The Multiple Myeloma DREAM Challenge

Khalil Ouardini
CentraleSupélec
ENS Paris-Saclay
ouardini.k@gmail.com

12 January 2022

1. Introduction

Multiple Myeloma (MM) is a cancer of the plasma cells in the bone marrow, and its clinical course depends on a complex interplay of clinical traits and molecular characteristics of the plasma cells. Since risk-adapted therapy is becoming standard of care, there is an urgent need for a precise risk stratification model to assist in therapeutic decision-making. While progress has been made, there remains a significant opportunity to improve patient stratification to optimize treatment and to develop new therapies for high-risk patients. This challenge aims to accelerate the development and evaluation of such risk models in MM.

The purpose of this study is to develop a model for prediction of clinical progression by 18 months using RNA-seq and clinical based features.

In this work we explore several methods for feature engineering of RNA-seq data, leveraging statistical analysis (differential expression) and a priori knowledge from the literature. We also explore several predictive models, from the classic ML and the survival analysis literature. Finally we experiment with different ways of including right censored patients to the training data.

2. Data

Training data: extracted from a single cohort study of The Multiple Myeloma Research Foundation (MMRF) dataset. It includes RNA-Seq (TPM normalized and raw counts) and microarray information for a total of 582 patient samples. The data was collected from bone marrow tumor cells from patients with newly diagnosed active MM. It also includes clinical feature such as ISS score (International Staging System), age and gender.

Validation Data: We only know that we have access to the TPM normalized RNA-seq data and clinical features. Patients could potentially come from several cohort studies. We also know the patients on the validation set are **not censored**.

In the following section we describe the feature engineering process.

2.1. Clinical features

Feature engineering of the clinical features:

- **Gender**: We encode the gender in a numeric binary variable
- **Age**: We bin the age variable in 6 intervals.
- **ISS**: Missing values are imputed with the mean (i.e $ISS = 2$)

2.2. Gene expression data

We start by verifying that all the patients have a a record in the gene expression matrix. After this sanity check we:

1. **Select the overlapped genes**: Since patients in the train and validation datasets come from different cohort studies, we can't guarantee that their transciptomic data is mapped to the same set of genes. Fortunately, the challenge orgzanizers gave access the list of overlapped genes between all the cohorts included in the study. Therefore, we make sure we only select the overlapped genes before the feature selection. The file is accessible here https://www.synapse.org/!Synapse:syn10792130
2. **Raw counts and TPM overlap**: We can also notice that the raw counts gene expression matrix (https://www.synapse.org/!Synapse:syn9744874) has 57997 genes compared to 24128 in the TPM normalized matrix. Only considering the overlapped genes from the start will also make sure the genes we select with the raw count gene expression matrix can be found in the TPM normalized matrix
3. **Gene-level QC**: filter genes with zero counts in all samples, genes with an extreme count outlier, genes with low mean normalized counts and genes with low variance.

4. **Differential Analysis**: Select differentially expressed genes with DESseq2 [3]. The hypothesis is that these genes can offer biological insight into the processes affected by the condition of interest. The DESeq2 software expects the raw gene expression matrix as input.The DESeq2 method fits a Negative Binomial Generalized Linear Model into the observed counts. A gene specific dispersion parameter that defines the relationship between the variance of the observed count and its mean value is also learned through maximum likelihood estimation. Using TPM data for DE is not advised because it significantly reduces the variability between genes (which is what we want to detect). DESEq2 returns a p-value for each gene representing the probability of rejecting the null hypothesis (H_0: the gene does not differentially express (on average)) It also returns the Log2FoldChange, i.e the effect size estimate that tells us how much the gene's expression seems to have changed due to the observed condition. We experiment with different ways of selecting the genes. In practice, selecting genes according to Log2fold change has shown the best results on predictive performance. The volcano plot 1 shows that many genes pass the significance test (blue points), but only a few show a a Log2FoldChange ≥ 2 (red points). Considering the Log2FoldChange is more straightforward to select up/down regulated genes.

5. **Gene expression signatures** We also try to include gene expression signatures for high-risk multiple myelomas that have been documented in the literature [2], [4]. (Empirically, we see that including this set of genes does not improve performance metrics so we do not include them)

   We end up with a list of 40 genes selected according to log fold change (i.e the most up-regulated or down-regulated genes). We select these genes from the TPM normalized data.

2.3. Microarray data

Although I did not include it in my study, we could run a differential expression analysis on the microarray gene expression data, and look at genes that overlap with the genes detected by DESeq2 on the RNA-seq data.

3. Model Assessment

3.1. Cross-validation

An assumption in the challenge is that the test data does not include any censored patients. When running cross-validation, we need to make sure our validation set reflects the distribution of the test set. To do so, we split the data in censored (191 patients) and non-censored patients (392 patients). We cross-validate on the non-censored patients, to make sure our validation set only contains non-censored patients for each split. Then we optionally augment each train split with the censored patients. With this method we can also assess the influence of censored data on the training.

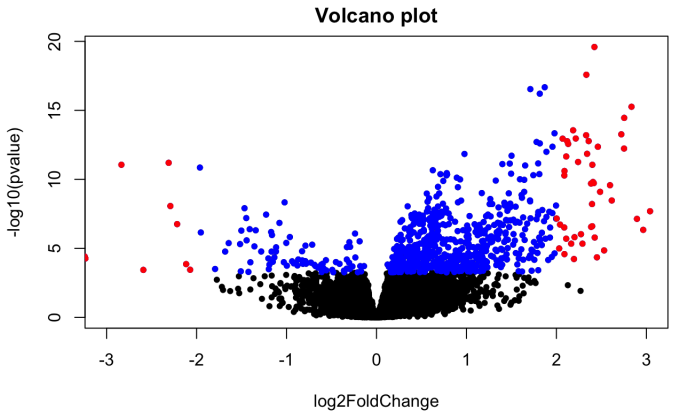

Figure 1. **Volcano Plot of the DESeq2 results**. The black dots and blue respectively represent the genes that fail and pass the significance test. Red dots genes are the genes that pass the significance test, and that show —Log2FoldChange— ≥ 3

We evaluate all the models with 5-fold cross-validation. We repeat the process over 10 different runs (with different seeds) and report the mean and the standard deviation of the reported metrics.

3.2. Metrics

Besides Accuracy, we report classification metrics that take into account class imbalance such as AUC, precision and recall.

4. Models

Given the size of the dataset, we prioritize the use classic supervised algorithms in this study.

We use a Logistic Regression model and Random Forest to include a non linear model in the comparative study. Finally, we also explored models designed to include censored data from the survival analysis literature. This part is covered in 4.3.

4.1. Results

As shown in Table 1, including gene expression improves the model on almost all metrics. The *Clinical Baseline* is logistic regression trained on the the clinical features only. We reach our best performance (AUC=65) on cross-validation with a Random Forest model (RF in Table 1) classifier trained on the clinical and transcriptomic data of non-censored patients, which is an improvement of 5 AUC

	Accuracy	AUC	Precision	Recall
Clinical Baseline	65.9 ± 4	60.9 ± 9	21.2 ± 4.5	52.3 ± 19
LogReg	69.3 ± 3.6	63.6 ± 7.3	34.3 ± 9.7	54.4 ± 13.9
RF	69.6 ± 2.4	65 ± 4.5	29.9 ± 7.8	57.9 ± 9.7
RF-ensemble	68.1 ± 9.9	63.6 ± 4.8	36.8 ± 11	54.7 ± 21
RF-ensemble-censored	52.7 ± 4.9	56.1 ± 5.0	65.3 ± 16.1	38.1 ± 16.7

Table 1. Average and standard deviation (in percentage) of Accuracy, AUC, Precision and Recall scores over 10 runs. The **Clinical Baseline** is a Logistic Regression trained on the clinical features only. All the other models are trained on clinical and gene expression features.**LogReg** refers to Logistic Regression. **RF** refers to a Random Forest classifier. **RF-ensemble** refers to the Ensemble of classification (HR FLAG) and regression (OS and PFS) models. **RF-ensemble-censored** refers to the same model but including the censored data in the training.

	Random Forest
N estimators	200
# Max depth	60
Max features	auto

Table 2. Hyperparameters of the *Random Forest* model. The parameters are selected with Random Search.

points over the clinical baseline.

The hyperparameters used for the experiments are summarized in Table 2.

We can notice that on average the recall and precision are rather low (and their standard deviation is high!) for all methods, which testifies of the difficulty to characterize the high risk patients.

4.2. Effects of Ensembling and censored data

We also have access to the OS (Overall survival time) and PFS (Progression Free Survival) labels in the training dataset. A high risk patient is defined as OS or PFS ≤ 18, therefore we could train two regressors to predict the PFS and OS labels, and use this condition to make a prediciton for the High Risk Flag (HR FLAG). The regressors can be combined with the HR FLAG classifier to make one final prediction for each patient of the test set. We implement this procedure and train the ensemble (RF ensemble in Table 1) on the non-censored patients. We notice more variance in the predictions (higher standard deviation) that could come from combining the variance of 3 different models. However it does not show an improvement compared to the Random Forest classifier on the AUC (RF)

By using the OS and PFS flag, we can include the censored data into our training splits. The train splits of the regressors are augmented with the censored data, but the test splits remain the same (i.e they only contain non-censored patients). Empirically, we see that this training scheme negatively impact the performance as AUC drops to 56.1 ((RF ensemble-censored in Table 1). Since we don't have additional information, we can only hypothesize that the clinical condition and the biology of those patients does not reflect the Multiple Myeloma group. Including this data during training could make it harder for a model to discriminate the HR patients by shifting the distribution of the training set.

4.3. Survival analysis

Quite unsatisfied with the performance of the classic ML models, I tried using some methods from the survival analysis literature, that are designed to take into account time and censoring.

Survival analysis methods should improve predictive accuracy of the model (compared with classification) because survival models "use all the information" by incorporating the time to death (or another event) in the development of the model and, more importantly, by accounting for subjects with unknown event times (censoring).

To compare survival methods to other classifiers, we need to produce a prediction. Typically, survival models use a random variable $T \geq 0$ representing the survival (or event) time. The survival function is the probability that an individual survives beyond time t:

$$S(t) = P(T > t), 0 \leq t \leq \infty \tag{1}$$

In practice we use **Cox proportional hazards model** [1] Once our regression model is fitted, we can estimate $S(t)$ for a new patient in the test set. To make a prediction we need to compute the probability of survival after 18 months (540 days) defined as $S(t = 540)$. If this probability is lower than 0.5, the patient is classified as "High Risk".

This reframing also allows us to use standard classification metrics, such as the area under the curve (AUC) to assess model performance.

Although this model was implemented in the codebase, I did not include the results in the main table as I'm still not sure how to produce a reliable prediction in the cross-validation setup (So far only a few patients ($\leq 5\%$) are predicted as High Risk which biases the classification metrics towards the majority class).

4.4. Multi-task learning

One idea I would have experimented with if had more time, would be to train a Multi Layer Preceptron (MLP) on

a Multi-task learning objective. We can define the loss function as a weighted sum of binary cross-entropy and mean squared error losses, to simultaneously train a neural network on classification and regression. As neural networks are prone to overfitting on small datasets, multi-task learning can be used as a regularizer, as previously discussed in the literature.

With such a model, we could also include censored data and simply set the classification loss to 0 for those data points.

5. References

References

[1] DR .Cox. Regression models and life-tables. *Journal of the Royal Statistical Society.*, 1972.

[2] Broyl A. de Knegt Y. van Vliet M. H. van Beers E. H. van der Holt B. el Jarari L. Mulligan G. Gregory W. Morgan G. Goldschmidt H. Lokhorst H. M. van Duin M. Sonneveld P Kuiper, R. A gene expression signature for high-risk multiple myeloma. *Leukemia, 26(11), 2406–2413.*, 2012.

[3] Huber W. Anders S. Love, M.I. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol 15, 550*, 2014.

[4] Jr Zhan F. Burington B. E. Huang Y. Colla S. Hanamura I. Stewart J. P. Kordsmeier B. Randolph C. Williams D. R. Xiao Y. Xu H. Epstein J. Anaissie E. Krishna S. G. Cottler-Fox M. Hollmig K. Mohiuddin A. Pineda-Roman M. Tricot G. ... Barlogie B. Shaughnessy, J. D. A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood, 109(6), 2276–2284.*, 2007.