Ensemble Approaches To Prostate Cancer Dream Challenge

The Data Wizard team

July, 2015

1 Introduction

Ensemble methods are popular method for increasing prediction accuracy. Ensembles tend to be more accurate than their component predictors. In traditional ensemble methods, the individual predictors vote, and a prediction is returned by the ensemble based on the collection of votes. In subchallenge q1a and q2, the predicted risk rank of all patients in the evaluation sets corresponds to the performance metrics, ROC AUC and precision-recall AUC, respectively. We therefore combine the predicted ranks from different component predictors.

2 Methods

2.1 Data preprocessing

For missing values, if the variables are related to blood information, we impute the missing data with normal blood values from [WebMD, 2015]. Continuous numeric attributes are "discretized" or binned into a small number of distinct ranges by entropy-based method. In subchallenge q1a, we train the traditional Cox models on the three training sets (ASCENT2, CELGENE, EFC6546) separately and select features according to their p-value in the Cox models. In subchallenge q1b and q2, we select the most relevant features for each subchallenge according to their information gain.

2.1.1 Subchallenge 1a

We derive different feature sets for the three training sets (ASCENT2, CELGENE, EFC6546), respectively and then train a specific model for each training set. Finally, we apply the Cox model with maximum penalised likelihood (CoxMPL model) on the selected features to train the three models. [Ma et al., 2014]. According to [Raykar et al., 2007], the survival problem can be cast as a *ranking problem* - where the task of survival analysis is to rank the data instances based on their survival times rather than to predict the actual survival times. Therefore we combine the ranking results from the different models. In this

challenge, we compare five ways for the ranking combination, such as average, minimum, maximum, median, and trimmed mean. Trimmed mean combiner eliminates the greatest and smallest values before calculating the standard mean of the given ranks.

2.1.2 Subchallenge 1b

To predict the exact time to event (death of a patient), we not only use uncensored data points but also apply the censored data points. We develop an adaboost-like regression algorithm called "adaboost-s" for survival problem, especially to predict time to event. The key technique under our algorithm is how we assign weights to the censored training data points. We think of the right censored data has a lower bound of his time to event. In the training phase, if the predicted time to event of a censored data point is smaller than the lower bound, the data point is considered incorrectly predicted and its weight is increased.

2.1.3 Subchallenge 2

We apply ensemble technique to combine the ranking results of several methods. The methods includes three probabilistic classifiers. The first classifier is a random forest classifier that trains on the whole training data set with the variable "DISCONT" as the class labels. The second classifier is also a random forest classifier that trains on the entire training data set but uses the data points which variable "ENDTRS_C" is "AE" or "possible_AE" as one class and the other points as another class. The final classifier is a gradient boosting classifier that trains on the same training set the second classifier trains on.

References

[Ma et al., 2014] Ma, J., Heritierc, S., and Lô, S. N. (2014). On the maximum penalized likelihood approach for proportional hazard models with right censored survival data. *Computational Statistics and Data Analysis*, 74:142–156.

[Raykar et al., 2007] Raykar, V. C., Steck, H., Krishnapuram, B., Dehing-Oberije, C., and Lambin, P. (2007). On ranking in survival analysis: Bounds on the concordance index. In *Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems (NIPS 2007)*, pages 1209–1216.

[WebMD, 2015] WebMD, L. (2015). http://www.webmd.com/.