EMORY UNIVERSITY

Machine Learning Final Project Proposal

Sub-Topic: Model Interpretation

Teammates: Chen, Lai, Wang, Yang, Zuo *in alphabetical order*

***Abstract---* In this proposal, using two tumor datasets, we tried to conduct survival analysis of patients and their genome features. In order to achieve our goal, we propose the use and comparisons of different methods for feature extraction, including traditional statistical method like PCA, LDA and genetic algorithm, simulated annealing, greedy forward selection, greedy backward elimination. Moreover, we plan to use neural network in feature extraction step, such as Autoencoder, Restricted Boltzmann Machine (RBM), convolutional neural network (CNN). For training phase, we plan to use Cox proportional hazard model. In the last, we will use cross-validation to verify our model.**

1. Introduction

Nowadays the development of biotechnology such as DNA microarrays permits the study of the relationship between genome and different diseases[1,2,3]. It is a great opportunity to human beings, while it also poses a huge challenge, as genome data usually provides massive information. How to choose and extract correct features for prediction is usually hard. In addition, how do we build a model that establishes relationship of correct features and diseases is another problem. Furthermore, the methods to utilize the data which contains both censored patients and uncensored patients also needs to be taken into considerations. Simply discarding the censored data is obviously not the best way as censored data also contains useful information even though they can not be used in the same way as uncensored data. In this project, we try to address the above three problems. We will explain especially which features are important in our model and how those features being selected.

2. Feature Selection & Extraction

In this project, the number of features, f (17,568), is much greater than the number of sample size, n (1,137). Since $f \gg n$, over-fitting is usually a potential problem. Because of the limit training set size, special mechanisms such as regularization, feature selection, and feature extraction are needed to prevent overfitting.

Feature selection is the process of selecting predominant features for model training. In this project, we plan to conduct several feature subset selection algorithms by using wrapper approach. Unlike feature selection, feature extraction transforms the original feature space to yield a subspace. We plan to implement several unsupervised learning methods to generate a new representation of the genomic data. A comparison between feature selection, feature extraction, and selection plus extraction will be performed under Cox model.

3. Model

Cox is a popular model used in survival analysis that can be used to assess the importance of various covariates in the survival times of individuals. We plan to use Cox proportional hazard model in the training phase. The hazard for a sample with feature x is given by the hazard function:

$$\frac{h(t)}{h(t_0)} = e^{\beta X}$$

where β is the regression parameter vector in Cox proportional hazards model. In the given dataset, not all features are associated with the survival time, mathematically, those unrelated features receive zero β_{x_i}, only those nonzero β_{x_i} need to be determined. To find a β that is able to accurately predict future data as well as receive a low test error rate, we plan to apply the wrapper method to find the best feature set to ensure the max prediction ability of Cox model.

4. Validation

We divide the samples into two parts, 80% of the data for training and the remaining 20% for testing. In validation, we plan to use 10-fold cross-validation to train for different feature sets. We divide our training data into 10 folds, take 9 folds for training and 1 fold for validation, each round we come up with one CI score (concordance index), which will be used as a standard to test how well the model performs. After the 10-fold validation, we have one CI score averaging from 10 CI scores for each selected feature set. After the comparison, our final model will take a subset of features with the best average performance. We retrain the model on the whole training set, then report the error on the testing dataset. We will also try other validation methods like leave one out validation and bootstrap.

5. MODEL FLOWCHART

Figure.1 shows the flowchart of our training process. Our purpose focuses on model selection. Before we start to perform feature selection/extraction, features which only contain NaN and negative value should be discarded in the pre-processing stage. Next, in feature extraction, we use sparse-denoising auto-encoder, RBM, CNN to compare the efficiency between each other. If possible, we will add some extra layer, such as max-pooling, to improve its behavior. In feature selection, to generate an appropriate subset, a heuristic algorithm such as genetic algorithm, forward-backward or greedy algorithm is needed.

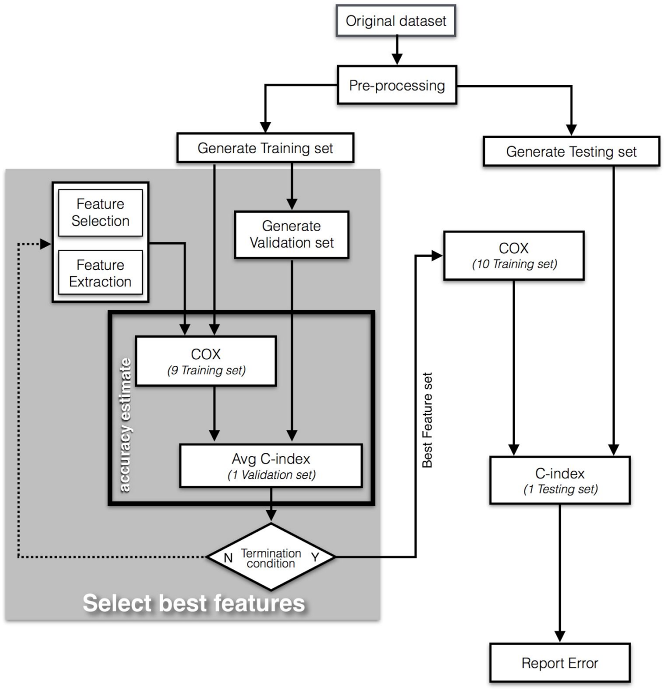
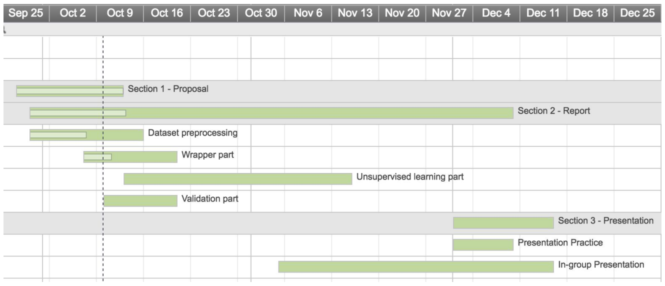
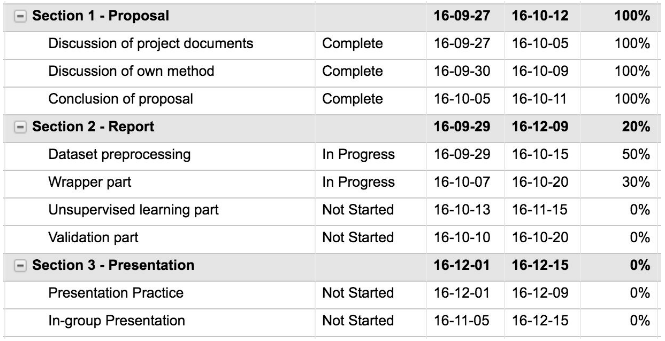


Figure 1

6. PROJECT TIMELINE





⊟ **Section 1 - Proposal**		**16-09-27**	**16-10-12**	**100%**
Discussion of project documents	Complete	16-09-27	16-10-05	100%
Discussion of own method	Complete	16-09-30	16-10-09	100%
Conclusion of proposal	Complete	16-10-05	16-10-11	100%
⊟ **Section 2 - Report**		**16-09-29**	**16-12-09**	**20%**
Dataset preprocessing	In Progress	16-09-29	16-10-15	50%
Wrapper part	In Progress	16-10-07	16-10-20	30%
Unsupervised learning part	Not Started	16-10-13	16-11-15	0%
Validation part	Not Started	16-10-10	16-10-20	0%
⊟ **Section 3 - Presentation**		**16-12-01**	**16-12-15**	**0%**
Presentation Practice	Not Started	16-12-01	16-12-09	0%
In-group Presentation	Not Started	16-11-05	16-12-15	0%

7. TOOLS

- Programming Language: Python, Matlab
- Source code version control: Github

8. REFERENCES

[1] A. Alizadeh et al., Nature 403, 503 (2000).
[2] T. R. Golub et al., Science 286, 531 (1999).
[3] M. E. Garber et al., Proc Natl Acad Sci USA 98, 13784 (2001).
[4] K, Ron, et al., Wrappers for feature subset selection (1995).