Multi-task machine learning for joint diagnosis and prognosis of human cancers

Pang Wing Kwan
Bachelor of Science in Actuarial Science

NAME: PANG WING KWAN
UNIVERSITY NO.: 3035572105
STUDENT'S MAJOR: ACTUARIAL SCIENCE
SUPERVISOR: DR. YU LEQUAN

RESEARCH COLLOQUIUM FOR SCIENCE UG STUDENTS 2022-23

Abstract

- **Cancer diagnosis**: classification of the cancer stage
- **Cancer prognosis**: prediction of the survival time of cancer patients
- **Objective**: Compare the model performance of the single-task models (penalized logistic regression for diagnosis and penalized COX regression for prognosis) and multi-task model (double-head neural network jointly doing the two tasks)
- **Predictors**: mRNA sequence and demographic information (age and gender) **Results**: multi-task model outperformed the single-task models.

Introduction

- Cancer is one of the deadliest diseases in the world while breast cancer is the most prevalent type of cancer developed in women → set the study scope to be breast cancer
- Correlation between the stage and the survival time of the cancer patients → possibility of boosting model performance by combining the two tasks (jointly predict the survival time and the stage of the patients)
- Compare the performance of the single-task models to that of the multi-task model to evaluate the effectiveness of multi-task models in boosting the performance

Materials and Methods

- **Data**: from Cancer Genome Atlas (TCGA) and only included Breast Invasive Carcinoma (BRCA) patients
- **Predictors**: Eigengene modules obtained from mRNA sequence and demographic information (age and gender) of patients
- **Testing and training set**: 854 patients for model-fitting; 214 patients for evaluation of model performance

Conversion of mRNA data to eigengenes

- **Reason**: To prevent the problem of the "curse of dimensionality" from happening
- **Main idea**: Genes were converted into co-expression modules (eigengene modules) through mining co-expression networks → reduced dimensions
- **Procedure**: filter out 50% of the genes with the lowest mean and a further 50% with the lowest variance (reduce the robustness of the correlational computations + reducing the impact of noises) → Group remaining genes into gene co-expression modules using a weighted network mining algorithm, local maximal QuasiClique Merger (lmQCM)
- **Results**: 20,531 genes → 29 eigengene modules

Model performance evaluations

Accuracy for cancer diagnosis

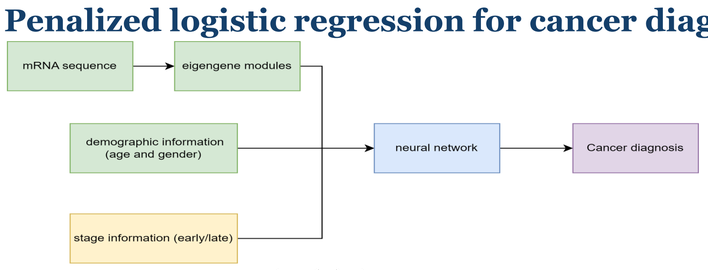
- **Measurement**: Accuracy of the classification task on the testing dataset
- **Evaluation**: Higher accuracy means a better performance

C-index for cancer prognosis

- **Measurement**: concordance between the actual ranking and the predicted ranking of the survival times, i.e. the proportions of patients correctly ranked
- **Formula**: $C-index = \frac{1}{n}\sum_{i\in[1,\dots,N|\delta_i=1]}\sum_{t_j>t_i} I(x_i\beta > x_j\beta)$
- **Evaluation**: A higher C-index indicates a better performance

Single-task models

Penalized logistic regression for cancer diagnosis



- Binary classification: predict the stage of the patient (early-stage or late-stage)
- Elastic net penalty
- Outcome Y: $log\left(\frac{p(X)}{1-p(X)}\right)$ where $p(X) = Pr(Y^s = 1|X)$
- Model: $Y^s = log\left(\frac{p(X)}{1-P(X)}\right) = \beta_0^s + \beta_1^s X_1 + \beta_2^s X_2 + \cdots + \beta_p^s X_p$
- Objective:

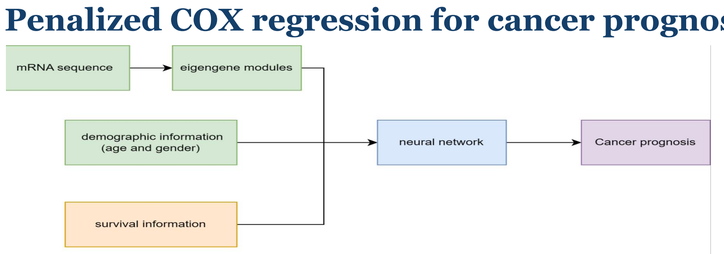
$$\min_{\beta} \Sigma_{i=1}^{854}(-y_i^S\ log(\hat{p}(x_i)) - (1-y_i^S)\ log(1-\hat{p}(x_i))) + \alpha\ [\rho\|\beta\|_1 + \frac{1-\rho}{2}\beta^T\beta]$$

COX regression

- Hazard function: $h(t|X)$
- Survival function: $S(t) = 1 - F(t) = Pr(T \geq t) = exp(-\int_0^t h(t|X)dt)$
- Baseline of hazard function: $h_0(t)$, hazard rate obtained when the values of all the predictor variables are set to be 0
- Hazard ratio (HR): $\frac{h(t|X)}{h_0(t)}$
- δ_i: survival status (0 when the patient is uncensored, 1 otherwise)
- t_i = survival time when δ_i = 0; t_i = observation period if δ_i = 1

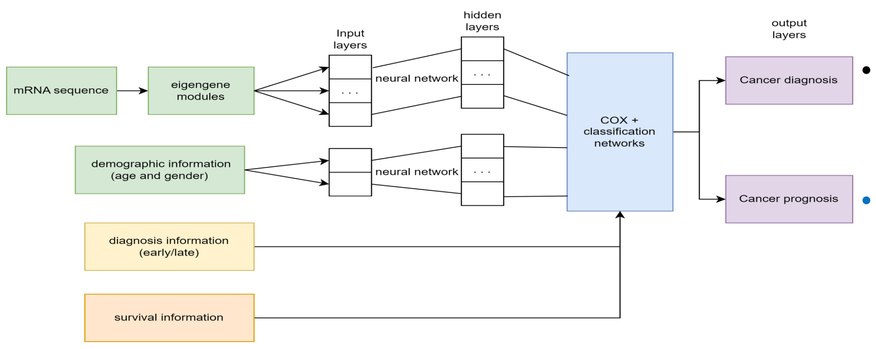
Materials and Methods

Penalized COX regression for cancer prognosis



- COX regression: Predict the hazard function
- Elastic net penalty
- Outcome Y: the hazard ratio, $log(\frac{h(t|X)}{h_0(t)})$
- Model: $log\ (h(t|X)) = \beta_1^T X_1 + \beta_2^T X_2 + \ldots + \beta_p^T X_p$
- Objective : $\min_{\beta} \sum_{i=1}^{854} -\delta_i[\beta^T x_i - log(\sum_{j\in R(t_i)} exp(\beta^T x_j))] + \alpha\ [\rho\|\beta\|_1 + \frac{1-\rho}{2}\beta^T\beta]$

Multi-task model (Multi-head neural network)

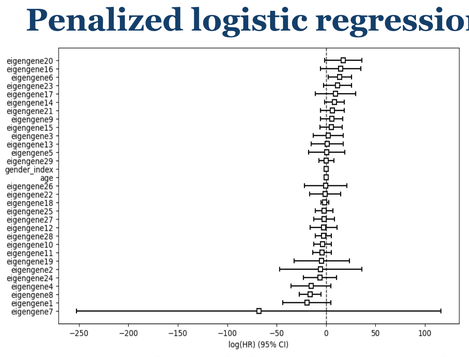


- Predict the stage of patient and hazard function jointly
- Has "COX regression" head and "classification" head
- 5-fold cross-validation (80% of the data used in training and 20% for purposes)
- Adaptive moment estimation (Adam) optimization algorithm for optimization
- Number of epoch: 100 (trained the network with all training data for 100 times)
- Batch size: 256 (training data will be divided into batches with size 256)
- Learning rate and dimension of hidden layers chosen with the minimum loss
- Loss function: $\min_{\theta}\ \sum_{i=1}^{n} -\delta_i\left(\beta^T x_i - \log\left(\sum_{j\in R(t_i)} \exp(\beta^T x_j)\right)\right) - \omega_i[y_i^S\ log(x_i) + (1-y_i^S)\ log(1-x_i))] + \alpha\ \|\beta\|_1$

Results

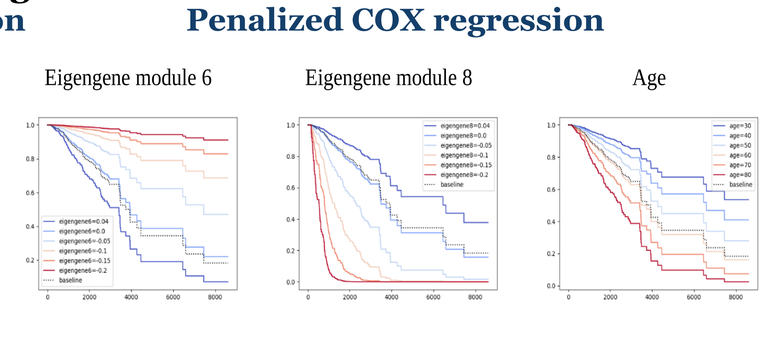
Single-task Models

Penalized logistic regression



- Optimal model: α = 0.01 and ρ = 1 (Accuracy = 0.668)
- Features chosen by the model: eigengene 2, 6, 8, 11, 20, 21, 23, 25, 26
- Ranked the importance of predictors using the value of coefficient

Penalized COX regression



- Optimal model: α = 0.0001 and ρ = 0.5 (C-index = 0.598)
- Features chosen by the model: eigengene 6 (+ve), 8 (-ve) and age (-ve)

Multi-tasks Model

- Optimal model: dimensions of hidden layers are set to be 6 (eigengene modules) and 1 (demographic data) (Accuracy = 0.761; C-index = 0.626)
- Better performance than single-task ones
- Feature selected: eigengene modules 8, 20, 26, 23 and 21 (also included in single-task models)

Evaluation metrics	Selected variables				
	Eigengene Module 8	Eigengene Module 20	Eigengene Module 26	Eigengene Module 23	Eigengene Module 21
Loss Value	+2.93	+1.56	+1.54	+1.37	+1.04
Accuracy	0	0	0	0	0
C-index	-0.010	+0.016	-0.045	+0.001	-0.037

Note: Only the features giving a positive impact on the loss value is included

- Features chosen by the neural network is closely related to the spreading of the cancer cells to neighboring tissues and other organs/their response to the various kinds of cancer treatments → able to select the right features

Conclusions

- Multi-task model performs better than the single-task models in both diagnosis and prognosis task
- Most important features: eigengenes 8, 20, 26, 23 and 21
- Limitations: lack in computational power for conducting more complex analysis
- Future studies: involve in other genomic information (miRNA sequence /histopathological images); testing the multi-head model on other cancers