CS776 Final Project
May 12th 2018
Project Report

Data Integration for Predicting Survival in Cancer Patients

Trang Vu

Abstract

In this project, multiple classifiers were used to predict survival in cancer patients at 5 years and 10 years using data sets that incorporated various clinical and genomic information. In contrast to the intuition such that the more informative the model, the better predictive power the model, the result of this work showed that predictive accuracy of the classifier models did not always increase with the increased number of features and even the most complex model with the most integrated training data did not outperform the model whose predictions solely rely on patient's clinical information.

1 Introduction

The project is motivated by the "Cancer Data Integration Challenge", one of the 2018 CAMDA contest challenges (http://camda2018.bioinf.jku.at/doku.php/contest_dataset#cancer_data_integration_challenge) which asks participants to develop an approach to efficiently integrate various information such as gene expression profiles, and clinical data to predict survival (i.e. survival time upon diagnosis and treatments) in breast cancer and neuroblastoma patients. The most popular cancer survival models is Cox probabilistic model which models time-to-event scenarios (eg. survival time from a specific time point to remission) often relies on careful selection of subset of features from data sets with large number of features (Yousefi 2017). A number of supervised machine learning based approaches (such as decision tree, artificial neural network) have also been developed to model survival prediction as a regression and classification tasks (Wei 2004, Chen 2009, Bashiri 2016). All models used either clinical data, or genomic data or both to make the prediction and the extent to which how much data is used varies across models.

While these models have good predictive performance, it was often assumed that the models that incorporated more information will perform better than models that incorporate less information. There has been no study to the best of my knowledge that compares the predictive power of models on various levels of integrated data sets. In this project, the goal is to examine whether the survival prediction increases as different information are integrated into machine learning models, and to identify the classifier that outperform other classifiers. Briefly, from clinical and genomic data sets, I derived various data sets with different levels of integrated information (defined by more or less informative attributes). These data sets were used to train several classifiers and their predictive accuracies are compared across the data sets as well as the classifiers. I expected that the combination of most integrated data sets and neural network model will yield the best predictive accuracy. Surprisingly, there is no significant difference among selected classifiers' predictive accuracy upon cross-validation using different data sets regardless of how much information was incorporated. When predicting test instances using these classifiers, however, the most informative data sets did yield the highest accuracy of 60% with neural network model. A subset of these test instances were then used to compare the predicted outputs with those produced by one of the web-based models called PREDICT (http://www.predict.nhs.uk/predict_v1.2.html) to gain further insight into how far off the model's predictions from one of the standard models. The agreement between PREDICT outputs and neural network model's outputs was 74% while that between PREDICT and observations was 55%. It should be noted that PREDICT uses the popular Cox time-to-event model to make prediction so the predicted outputs (C-index) need to be converted to categorical values by imposing some predefined thresholds.

2 Methods

All the data preparation and analysis were written as Python scripts and executed on Linux server of Amazon Web Service EC2 instances (https://aws.amazon.com/ec2/). These scripts are included in supplementary files.

2.1 Data Acquisition, Preparation and Transformation

All data used in this project was downloaded from the CAMDA website (http://camda2018.bioinf.jku.at/doku.php/data_download) by creating an account to log in and accepting the download agreement. The downloaded zipped folder contains data and metadata of patients, breast tumor samples, gene expression levels, and copy number alteration data sets. The sizes of these data sets are summarized in Table 1 (see section Figures and Tables). Briefly, these data sets can be thought as clinical or genomic data sets in this project. The patient data set includes attributes such as age at diagnosis, the breast tumor type, whether patient receives chemo treatment while breast tumor samples data set includes information about the tumor such as size, stage, grade, etc. The information from these data sets are considered to be clinical information. The genomic information comes from the gene expression data set, which includes gene expression level (log of intensity) for 20k+ genes in different patients, and the copy variant alteration data set, which reports the categories for which copy number changes occurred.

The data integration step can be described informally as follow. First the patient data and breast sample data were merged into one data set called set A using patients' ID number. Set A was later merged with either gene expression data set or copy number alteration data to produce set B

and set C respectively. Set B was then merged with copy number alteration data to produce set D (note that set D can also be obtained by merging set C with gene expression data). Only a subset of genomic information was integrated with clinical data because otherwise, the number of attributes will be too large, which can make training classifiers impractically long. This subset includes 173 targeted known cancer genes.

After each merging step, the data sets were profiled to identify instances with missing values and remove them from the sets. Moreover, each attribute in each data sets was also classified to either numerical, categorical or binary attributes so that its values can be properly processed. If an attribute is numerical, its values are kept as floats. An exception to this rule is applied to the class attribute, which is the overall survival in months. This attribute is numerical in the original patient data set, however because the goal is to build a classifier that predict survival, the class values were converted to three categorical labels of "not survive 5 years" (Class1), "survive 5 years" (Class2), and "survive 10 years" (Class3). If an attribute is binary, its values are converted to 0 or 1. If an attribute is categorical and has *N* different categories, the attribute and its corresponding values are replaced by N new binary attributes, each represents one category of the original attribute (this is essentially one-hot-encoding). As a result of this data transformation, the numbers of attributes in the integrated data sets significantly increased compared to original data sets as shown in Table 1.

2.2 Classifier Training and Evaluation

The classifiers selected in this project are decision tree, random forest, logistic regression classifiers from scikit-learn (http://scikit-learn.org/stable/supervised_learning.html) and artificial neural network models from Keras (https://keras.io/). The scikit-learn classifiers were selected because based on my experience they perform reasonably well, quite fast, and relatively easy to debug. The neural network classifier was selected because neural network with hidden layers generally works well for non-linear decision boundary classification task, and also when there is little prior knowledge about how the attributes contribute to final output (Chen 2009, Yousefi 2017). In this project, the neural network model has an input layers with the number of nodes equals to the number of features a hidden layer with the number of hidden neurons equals to the number of input nodes in the input layer, and an output layer with 3 nodes representing the class labels described above. Neural network model is trained in 300 epochs at the learning rate at 0.01, with activation functions for the hidden layer and output layer being 'sigmoid' and 'softmax' respectively. The loss function is 'categorical_crossentropy', and the network training is optimized with stochastic gradient descent.

Each integrated data set was split to a train and a test sets at ratio 4:1. The train data set was used to train the classifiers and the performance of each classifiers were evaluated via 5-fold cross-validation by computing the average accuracy. Each trained classifier was then validated using the test sets by computing the predictive accuracy. A subset of 100 test instances was extracted from the test data sets upon which highest accuracy was achieved, and this subset was used to compare the predictions of the best trained classifier against the predictions of an existing model named PREDICT (http://www.predict.nhs.uk/predict_v1.2.html) that would predict the overall survival of breast cancer patients given patient's clinical information such as 'age at diagnosis', 'mode of detection', 'tumor size in mm', 'tumor grade', 'number of positive nodes', 'ER status', 'HER2 status', 'KI67 status', 'gen chemo regimen'. Based these information, PREDICT predicts survival as the percentage of patients alive at 5 years and 10 years respectively. In other words, PREDICT model does not produced deterministic categorical outputs but rather survival possibilities. To make the comparison straightforward, thresholds were used to convert each output probability to one of the 3 class labels. Specifically, if the survival probability at 10 years is greater than 50%, it will be labeled as Class3. If the survival probability at 10 years is less than 60% and the survival probability at 5 years is greater than 80%, it will be converted to Class2. Consequently, if the survival probability is less than 80% at 5 years, it will be labeled as Class1. The thresholds 80% and 60% were the average probabilities of the predicted survival probabilities at 5 years and 10 years of the 100 test instances. The agreement between predictions and observations or between predictions made by two predicted models is defined as the percentage of instances whose labels match.

3 Results

3.1 Integrated Data Sets

The data sets resulted from integration process are summarized in Table 1. The size of each data set represents the number of instances (patients) and the attributes (features) of each instance. Because copy number alteration data contains all patients in the clinical data, the data set 'Clinical' and 'Clinical + CNA' have the same number of rows. The number of instances in 'Clinical + GE' and 'Clinical + GE + CNA' are smaller than that in the other two data sets because not all patients in 'Clinical' data set present in the gene expression data sets. Missing values are not allowed in these integrated data sets because they will cause trouble for the classifiers. As mentioned earlier in the method section, all categorical attributes are converted to binary attributes using one-hot encoding, hence the number of attributes in the integrated sets are significantly larger than the original data sets. Consequently, the data in many attributes are very sparse. The distribution of class labels suggests that the data sets are not well-balanced with Class3 being the dominant class.

3.2 Classifier Evaluation

3.2.1 Cross Validation using Training Sets

Because the classification task is a multi-class classification problem, accuracy instead of precision and recall was used as metrics for classification evaluation. By performing cross validation on training sets, the average accuracy of each classifier can be assessed. Figure 1 shows the comparison of average accuracy computed from 5-fold cross validation of 4 different classifiers on 4 different training data sets. In most cases, for the same data set, the accuracy is not significantly different among the classifiers. Similarly, there is no significant difference in accuracy across the data set for a given classifier. This result did not confirm the hypothesis that the classifier is expected to perform better when more data is more available. An exception to this observation is the accuracy of neural network on 'Clinical + GE' data set. Neural network performed particularly poorly on this data set. It should be pointed out that the standard deviation in accuracy of neural network model varies greater than the other classifiers, suggesting that for some data, the neural network can perform slightly better than others.

3.2.2 Evaluation of Classifiers using Testing Sets and PREDICT model

The classifiers were trained and applied on held-out test sets of each integrated data sets. The accuracy of the predictions was computed as the percentage of instances whose labels match actual outputs. These values

were summarized in Table 2. Surprisingly, while neural network perform quite poorly on data sets with gene expression data ('Clinical + GE', 'Clinical + GE + CNA') during cross validation, the predicted accuracy was higher for the test data sets. This suggests that the trained neural network model probably adapted to some unknown peculiarities in the training sets that also present in the test set.

When comparing neural network's predictions on the 'Clinical + GE + CNA' data sets with the Cox-like model PREDICT's outputs, the agreement between PREDICT and neural network predictions was 74% (See supplementary files). In contrast, the PREDICT outputs and actual observations only agree at 55%. This suggests that at the imposed conversion thresholds specified in the method section, neural network's predictions using the highest level of integrated data set does not outperform the performance of a predictive model that only use patients' clinical information.

4 Discussion

Even though neural network classifier did perform the best on the most integrated data set, it was not the general trend that the more data the better predictive accuracy. I have hoped that the gene expression data and copy number alteration data can inform the classifiers hidden knowledge about cancer patients that could explain why patients that appear to have similar clinical profile can have different survival outcomes. One of the reasons why classifiers did not perform better when more features are available can be attributed to that not all features are informative and useful in helping classifier making correct predictions. It should also be pointed out that only a subset of gene expression and copy number alteration data was used to integrate with clinical data. This decision is to reduce the feature sets but it could miss some genes that are more informative. Perhaps feature selection in increment can be done in the data preprocessing step to select features that would be most informative to the classifiers. While neural network is expected to work well when no prior knowledge regarding the features are available, training and tuning hyper-parameters of neural network model is not trivial. For future works, perhaps simultaneous tuning network parameters can be done to find the best parameters to train the neural network. In this project, I have attempted to simultaneously tune epochs, learning rate, activation function and number of hidden neuron but the process took too long and often takes up too much memory space especially for network models with small learning rate and large number of hidden neurons. Perhaps parallel processing can be employed to speed up the process.

The use of one-hot-encoding to convert categorical attributes to binary attributes can result in the loss of context. For example, in copy number variant, the positive and negative values can indicate a gain or loss in certain phenotypes respectively but when these values were converted to binary, this context is loss, and the attribute may not be as useful as it could have been. Incorporating context information required domain knowledge and should be treated more carefully.

One key difference in the evaluation method used in this project compared with methods used in related works is that the prediction was modeled as a classification problem which outputs deterministic categorical outcomes. In most survival analyses, the concordance index (C-index) instead of accuracy is the popular choice in evaluating survival prediction (Laas 2014) and perhaps that is a more practical metric as physicians would like to know the survival probabilities of their patients in 5 – 10 years so that they can better design treatment plans. The method of evaluating the neural networks' prediction against existing model's prediction is not a direct comparison as two models produce two different type of outcomes. Therefore, this method of evaluation is subjected to biases stemmed from the threshold values used to convert survival probabilities to categorical values. Due to time constraints, only 100 test instances were used for the evaluation as PREDICT is a web-based model, which requires one to manually put in information case by case to get the predicted output. If time is not a constraint, implementing and testing PREDICT in batch maybe a better solution. Furthermore, it should be noted that PREDICT was trained on entirely different sets of patients and that could be one of the reasons why its agreement with the actual overall survival is low.

5 Tables and Figures

Table 1. Integrated Data Sets

Description	Shape	Class 1	Class 2	Class3
Patients	1981x20	NA	NA	NA
Samples	1981x8	NA	NA	NA
GE	24369x1906	NA	NA	NA
CNA	22545x2175	NA	NA	NA
Clinical	1171 x 68	228	243	407
Clinical + GE	1130 x 234	208	251	338
Clinical + CNA	1171 x 874	228	243	407
Clinical + GE + CNA	1130 x 1040	208	243	338

Summary of original and integrated data sets. Shape refers to the number of rows (instances) and the number of columns (attributes) in each data set. Class 1, Class 2, Class3 are the number of instances whose class labels are "not survive in 5 years", "survive 5 years", and "survive 10 years" respectively. These class labels are only applicable to integrated data. GE and CNA are gene expression and copy number alteration respectively.

Table 2. Performance of Classifiers

Data sets	DT	RF	LogReg	NN
Clinical	0.45	0.47	0.55	0.50
Clinical + GE	0.46	0.49	0.52	0.57
Clinical + CNA	0.43	0.44	0.48	0.32
Clinical + GE + CNA	0.39	0.46	0.46	0.60

Summary of the predicted accuracy of various classifiers on different integrated data test sets. DT = decision tree, RF = random forest, LogReg = Logistic Regression, NN = Neural Network. GE and CNA are gene expression and copy number alteration respectively.

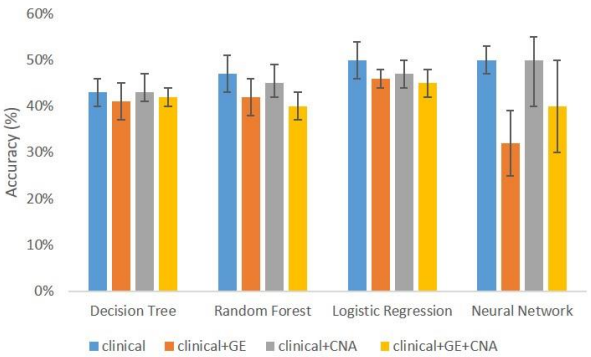

Figure 1: Average and standard deviation of accuracy from cross validation of classifiers on various integrated training data sets

6 References

Yousefi, S. *et al.* (2017) Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports*, **7**, doi:10.1038/s41598-017-11817-6.

Wei, J.S. *et al.* (2004) Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma. *Cancer Res*. **64**, 6883-6891.

Chen, Y. *et al* (2009) Artificial neural network prediction for cancer survival time by gene expression data. *3rd International Conference on Bioinformatics and Biomedical Engineering, IEEE*.

Bashiri, A. *et al.* (2016) Improving the prediction of survival in cancer patients by using machine learning techniques: experience of gene expression data: a narrative review. *Iran J. Public Healh*, **46**, 165-172.

Laas, E. *et al.* (2014) Are we able to predict survival in ER-positive HER2-negative breast cancer? A comparison of web-based models. *British J. Cancer*, **112**, 912-917.