IT399 Minor Project
Report

Prediction of Lung Cancer patient survival via supervised machine learning classification techniques

Submitted in partial fulfillment of
the requirements for the award of the degree of

Bachelor of Technology
in
Information Technology

Submitted by

Roll No	Names of Students
15IT238	Shreyansh Sancheti
15IT125	Manohar Madanu
15IT104	Aditya Sharma

Under the guidance of
Ms. Nagamma Patil

Department of Information Technology
National Institute of Technology Karnataka
Mangalore, Karnataka, India – 575025

Monsoon Semester 2018

Department of Information Technology

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

Certificate

This is to certify that this is a bonafide record of the project presented by the students whose names are given below during Monsoon and Year 2018 in partial fulfillment of the requirements of the degree of Bachelor of Technology in Information Technology.

Roll No	Names of Students
15IT238	Shreyansh Sancheti
15IT125	Manohar Madanu
15IT104	Aditya Sharma

Ms. Nagamma Patil
(Project Guide)

Date: 02-May-2018

Declaration

We hereby declare that the project entitled PREDICTION OF LUNG CANCER PATIENT SURVIVAL VIA SUPERVISED MACHINE LEARNING CLASSIFICATION TECHNIQUES submitted by us in partial fulfillment for the completion of the course Information Retrieval is a record of bonafide project work carried out by us under the guidance of Ms. Nagamma Patil We further declare that the work done in this project will not be submitted either in full or part for the reward for degree or diploma in this institute or any other institute.

Shreyansh Sancheti
15IT238

Manohar Madanu
15IT125

Aditya Sharma
15IT104

Date: 02-May-2018

Abstract

Lung cancer patient survival have been previously estimated by applying various machine learning techniques to large data sets such as SEER program database. For lung cancer, it is not well understood which types of techniques would yield more predictive information, and which data attributes should be used in order to determine this information. In this study, a number of supervised learning techniques is applied to the SEER database to classify lung cancer patients in terms of survival, including linear regression, decision trees, gradient boosting machines (GBM), support vector machines (SVM), and a custom ensemble. Key data attributes in applying these methods include tumor grade, tumor size, gender, age, stage, and number of primaries, with the goal to enable comparison of predictive power between the various methods The prediction is treated like a continuous target, rather than a classification into categories, as a first step towards improving survival prediction. The results show that the predicted values agree with actual values for low to moderate survival times, which constitute the majority of the data. The best performing technique was the custom ensemble with a Root Mean Square Error (RMSE) value of 15.05. The most influential model within the custom ensemble was GBM, while Decision Trees may be inapplicable as it had too few discrete outputs. The results further show that among the five individual models generated, the most accurate was GBM with an RMSE value of 15.32. Although SVM under performed with an RMSE value of 15.82, statistical analysis singles the SVM as the only model that generated a distinctive output. We conclude that application of these supervised learning techniques to lung cancer data in the SEER database may be of use to estimate patient survival time with the ultimate goal to inform patient care decisions, and that the performance of these techniques with this particular data set may be on par with that of classical methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning uses mathematical algorithms implemented as computer programs to identify patterns in large datasets, and to iteratively improve in performing this identification with additional data. The algorithms are commonly used in different domains and diverse applications. Using these techniques to evaluate disease outcomes can be challenging. Since patient data is generally unavailable for public analysis. One exception is the SEER Program [1, 2] from the National Cancer Institute (NCI) at the National Institute of Health (NIH). This dataset provides information on cancer statistics of the United States population. Machine learning techniques are applied to this dataset to analyse data specific to lung cancer, with the goal to evaluate the predictive power of this techniques. Lung cancer was chosen as it ranks as a leading cause of cancer-related death, with dismal 5-year survival rates. Given a dataset of lung cancer patients with particular demographic (e.g., age), diagnostic (e.g., tumour size), and procedural information (e.g., Radiation and/or Surgery applied), the question is whether patient survival can be computationally predicted with any precision. In this study, patients diagnosed with lung cancer during the years 20042009 were selected in order to be able to predict their survival time. A number of supervised learning methods was employed to classify patients based on survival time as function of key attributes and, thus, help illustrate the predictive value of the various methods. The techniques chosen include linear regression, Decision Trees, Gradient Boosting Machines, Support Vector Machines, and a custom ensemble. This exercise also enables comparing the predictive value of the methods when applied with the chosen attributes to analyse the lung cancer patient data. The dataset in this study focuses on measurements available at or near the time of diagnosis, which represents a more proactive set of survival predictors.

1.1 Motivation

Respiratory (lung) cancer is the second most common cancer, and the leading cause of cancer-related deaths among men and women in the USA. Survival rate for lung cancer is estimated to be 15% after 5 years of diagnosis. Applying data mining techniques to cancer data is useful to rank and link cancer attributes to the survival outcome. Further, accurate outcome prediction can be extremely useful for doctors and patients to not only estimate survivability, but also aid in decision making to determine the best course of treatment for a patient, based on patient-specific attributes, rather than relying on personal experiences, anecdotes, or population-wide risk assessments.

Chapter 2

Literature Review

2.1 Related Work

Previously published work has analyzed the SEER database via statistical [3, 4] as well as classification techniques [5, 6]. In earlier work [7], the concept of agglomerative clustering [8, 9] was applied to generate groups of cancer patients. The algorithm of clustering of cancer data (ACCD) was proposed to predict outcomes, with any number of factors as input and with the goal of grouping patients uniformly in terms of survival. The approach was applied to a large breast cancer dataset from the SEER database using information concerning tumor size, tumor extension and lymph node status. The results showed the approach to be more effective than the traditional TNM (tumor-node-metastasis) cancer staging system [7]. A few studies have evaluated lung cancer patient survival by analyzing the SEER database with machine learning techniques, including ensemble clustering-based approaches [10], SVM and logistic regression [11], and unsupervised methods [12]. Data classification techniques were evaluated in [13] to determine the likelihood of patients with certain symptoms to develop lung cancer. In [14], the performance of C4.5 and Nave-Bayes classifiers was compared applied to lung cancer data from the SEER database, achieving 90Patient survival.

2.2 Outcome of Literature review

Applying a cluster analysis requires hypothesis that the groups exist. But this assumption may be false or weak. Clustering results should not be generalized. Cases in the same cluster are similar only with respect to the information cluster analysis was based on i.e., dimensions/variables inducing the dissimilarities. Need to develop model for conditional survival prediction (e.g. 5-year prediction, given that the patient has already survived for 1 year), and use of under-sampling/oversampling to deal with unbalanced data.

2.3 Problem Statement

To predict the survival time of the patient suffering from lung cancer using various machine leaning techniques

2.4 Objectives

- To do pre-processing on the raw data and choosing best features.
- To apply the regression models and predict and survival time of the patients.
- To compare the results of the various models using statistical measures.
- Advanced machine learning techniques to evaluate the model.

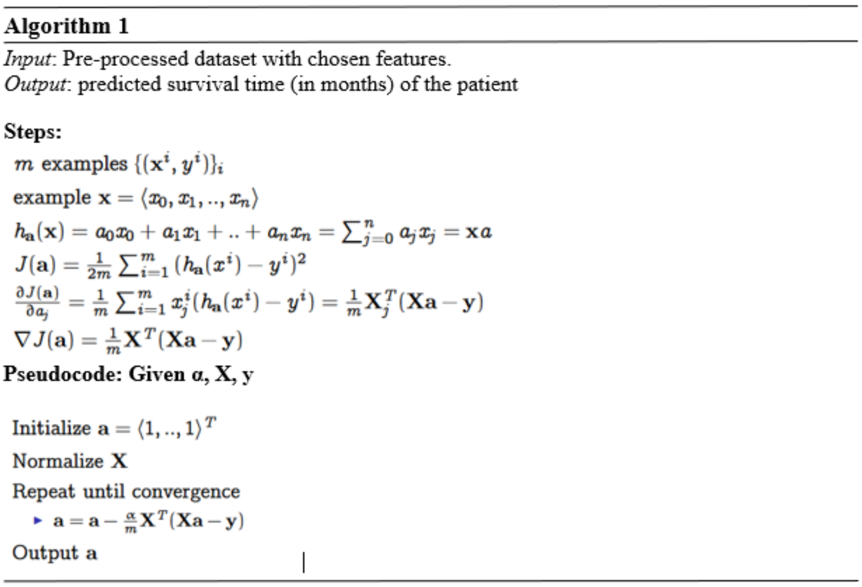
Chapter 3

Methodology

We present different methods to quantify the selection bias using the collected randomized data.

3.1 Algorithms

3.1.1 Linear Regression



Algorithm 1

Input: Pre-processed dataset with chosen features.
Output: predicted survival time (in months) of the patient

Steps:

```
$m$ examples $\{(\mathbf{x}^i, y^i)\}_i$
example $\mathbf{x} = \langle x_0, x_1, .., x_n \rangle$
$h_{\mathbf{a}}(\mathbf{x}) = a_0 x_0 + a_1 x_1 + .. + a_n x_n = \sum_{j=0}^{n} a_j x_j = \mathbf{x} a$
$J(\mathbf{a}) = \frac{1}{2m} \sum_{i=1}^{m} (h_{\mathbf{a}}(x^i) - y^i)^2$
$\frac{\partial J(\mathbf{a})}{\partial a_j} = \frac{1}{m} \sum_{i=1}^{m} x_j^i (h_{\mathbf{a}}(x^i) - y^i) = \frac{1}{m} \mathbf{X}_j^T (\mathbf{X}\mathbf{a} - \mathbf{y})$
$\nabla J(\mathbf{a}) = \frac{1}{m} \mathbf{X}^T (\mathbf{X}\mathbf{a} - \mathbf{y})$
```

Pseudocode: Given α, X, y

```
Initialize $\mathbf{a} = \langle 1, .., 1 \rangle^T$
Normalize $\mathbf{X}$
Repeat until convergence
    ▸ $\mathbf{a} = \mathbf{a} - \frac{\alpha}{m} \mathbf{X}^T (\mathbf{X}\mathbf{a} - \mathbf{y})$
Output $\mathbf{a}$
```

Figure 3.1: Pseudo-code

The simplest method implemented is linear regression, one of the oldest and most widely used correlational techniques. The goal of the method is to fit a straight line to a set of data points using a series of coefficients multiplied to each input, like a weighting function, and an intercept. The weights are decided within the linear regression function in a way to minimize the mean error. These weight coefficients multiplied by the respective inputs, plus an intercept, give a general function for the outcome, patient survival time. The psuedo code for linear regression with multiple variables is shown above in figure 3.1.

3.1.2 Decision tree

Decision tree builds regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Leaf nodes are nothing but indication of survival time.For regression decision trees there are more classifications than in a typical classification decision tree that makes the outcomes near continuous.

3.1.3 Random Forests

The Random Forest technique generates a number of decision trees during training which are allowed to split randomly from a seed point. This results in a forest of randomly generated decision trees whose outcomes are ensemble by the Random Forest Algorithm to predict more accurately than a single tree does alone. It is a meta estimator that fits a number of classifying decision trees on various sub-samples of the data set and use averaging to improve the predictive accuracy and control over-fitting. The number of trees was set to 500.

3.1.4 Gradient Boosting Machine

Gradient Boosting for regression. GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function. Loss function to be optimized i.e least squares regression. Learning rate shrinks the contribution of each tree bylearning rate. There is a trade-off between learning rate and n-estimators The number of boosting stages to perform are 100. Gradient boosting is fairly robust to over-fitting so a large number usually results in better performance. The maximum depth of the individual regression estimators is 3. It limits the number of nodes in the tree.

3.1.5 Support Vector Machine

Epsilon-Support Vector Regression. Assuming that a set of training data has been labeled as belonging to one of two sets, the algorithm represents them in space and specifies a hyper-plane maximally distant from both to separate them. The plane is called the maximal margin hyper-plane. If a linear separation is not possible, the algorithm employs kernel methods to obtain a non-linear mapping to a feature space. Kernel method selected is rbf. A drawback of SVM is that the method can be subject to over-fitting when the data is noisy.

3.1.6 Ensemble

A custom ensemble method was used to bring all these models together for a more accurate prediction. The results were expected to be better with the custom ensemble than with any single approach, and the ensemble was simple to implement and easily adaptable to model adjustments.

Chapter 4

Work Done

4.1 Experimental Framework

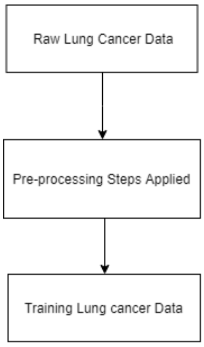


Figure 4.1: Pre-Processing steps applied to data

For this study we have raw lung cancer data taken from SEER database. Data was available in raw ascii format. Pre-processing steps were applied to convert raw data to readable .csv files. It had total 137 features. Out of 137 features 18 were selected to work. This were selected after taking expert opinion. So total 18 features with around 1 lakh tuple data was prepared and ready to feed to model. All pre-processing steps applied are shown in figure 4.1. To build models out of the data follow algorithms were proposed Linear Regression, Support Vector Machine, Decision Tree Classifier Gradient Boosting Machines and model was generated for each algorithms respectively. The model was trained by providing input test data of lung cancer patients (using 10 fold cross validation). Finally the model was tested on test data and predicted survival time for lung cancer patient was recorded. All steps applied are shown in figure 4.2.

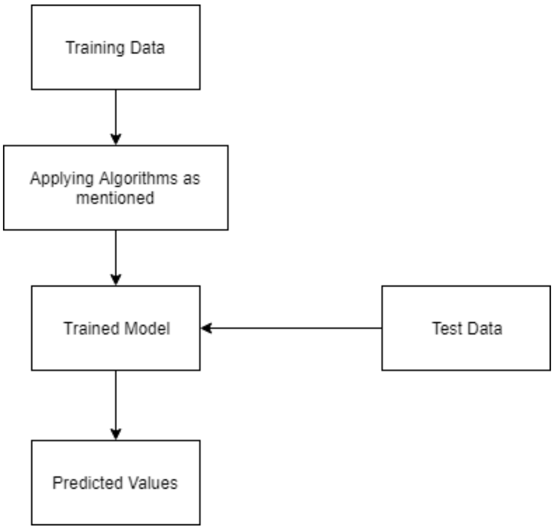


Figure 4.2: Experimental Framework for Training Model and Predicting Values

4.2 Results and Discussion

A set of key attributes was selected that consisted of both continuous (numeric) and categorical (discrete) variable types, for a total list of 19 attributes (including the target Survival Time, Table 1). The models were then used to correlate the remaining 18 attributes to the patient outcome, i.e., survival time. The data taken from the SEER database is from patients diagnosed from the years 2004 to 2009. Beginning in 2004, both stage and grade criteria were changed in definition, which makes 2004 the first available year for consistent five-year survival time evaluation. The techniques employed in this study are focused on creating a correlation between patient survival time and the attributes listed in table 4.2. Therefore, the range of survival time is from 0 to 72 months, and the output from the models is expected to be within this range.

The table 4.2 Selected SEER attributes and their respective descriptors

Number	Attribute	Type
1	Age	Discrete
2	Grade	Numeric
3	Radiation Sequence with surgery	Numeric
4	No. of Primaries	Discrete
5	T	Numeric
6	N	Numeric
7	M	Numeric
8	Radiation	Numeric
9	Stage	Numeric
10	Primary Site	Numeric
11	First Malignant Primary Indicator	Numeric
12	Sequence Number	Discrete
13	CS Lymphnodes	Numeric1
14	Histology Recode - Broad Groupings	Numeric
15	RXSumm-ScopeRegLNSur(2003+)	Numeric
16	RXSumm-SurgPrimSite(1998+)	Numeric
17	DerivedSS1977	Numeric
18	Tumor Size	Numeric
19	Survival Time	Discrete

The table 4.2 Comparing the standard metrics for different models

Model	RMSE	Standard Deviation	SD of residuals	Mean
Linear Regression	15.492	10.171	15.490	14.992
GBM	16.328	4.323	16.331	15.171
Decision Tree	15.252	10.786	15.264	15.253
Random Forest	15.197	10.453	15.199	15.075
SVM	18.073	4.988	16.867	8.697
Ensemble	15.811	5.967	5.967	14.893

After arranging attributes, the selected regression-based techniques were applied to the data standard deviation and mean of predictions, and standard deviation of the difference between the predicted and actual survival time are presented in Table .4.2 Across the entire training and validation sets, the mean survival time is 14.013 months and the standard deviation is 7.781 months

The Random Forest model was the most able model with an RMSE value of 15.197, as shown by the results in Table 4.2. This was followed closely by the Decision Tree model with an RMSE value of 15.252 and Linear regression with an RMSE value of 15.492. Trailing behind the other methods were GBM and SVM with RMSE values of 16.328 and 18.073, respectively. Once these models were all brought together and weighted using a linear regression, the resulting ensemble was slightly more adept, with an RMSE value of 15.811.

Chapter 5

Conclusion and Future work

The results from this study suggest that a correlational approach via supervised machine learning may be applicable to lung cancer patient survival prognosis, in the sense that meaningful predictions can be made with reasonable accuracy bands describable by the resulting statistics. The only model that may be non-applicable is Decision Trees, as it has too few discrete outputs. Despite the issues with the other models investigated, no model other than Decision Trees seems truly lacking, with the more advanced GBM model displaying stronger performance, and the SVM being worthy of independent attention as it predicts a distinct distribution but scores similarly to the others.

Future work could reevaluate the inputs for the selected models. While RMSE was chosen during our up-front design, other metrics may be warranted; the scores and standard deviations of linear regression and SVMs suggest that a deeper analysis may prove fruitful. A new effort to evaluate each individual criteria and how it relates to patient survival, especially in the case of longer-lived patients, could be key to more accurate and predictive correlational supervised machine learning algorithms.

References

[1] NCI SEER Training Lung Cancer Stats, Introduction to Lung Cancer. SEER Training Modules, National Cancer Institute, 2015 (Available from: http://training.seer.cancer.gov/lung/).

[2] NCI SEER Overview, Overview of the SEER Program. Surveillance Epidemiology and End Results, (2015) (Available from: http://seer.cancer.gov/about/).

[3] S. Ramalingam, K. Pawlish, S. Gadgeel, R. Demers, G. Kalemkerian, Lung cancer in young patients: analysis of a Surveillance, Epidemiology, and End Results database, J. Clin. Oncol. 16 (1998) 651657.

[4] T.K. Owonikoko, C.C. Ragin, C.P. Belani, A.B. Oton, W.E. Gooding, E. Taioli, et al., Lung cancer in elderly patients: an analysis of the surveillance, epidemiology, and end results database, J. Clin. Oncol. 25 (2007) 55705577.

[5] Identifying hotspots in lung cancer data using association rule mining, in: A. Agrawal, A. Choudhary (Eds.), 11th International Conference on Data Mining Workshops (ICDMW), IEEE, 2011.

[6] Finding survival groups in SEER lung cancer data. machine learning and applications (ICMLA), in: I. Skrypnyk (Ed.), 11th International Conference On; 2012, IEEE, 2012.

[7] A clustering-based approach to predict outcome in cancer patients, in: K. Xing, D. Chen, D. Henson, L. Sheng (Eds.), Sixth International Conference on Machine Learning and Applications (ICMLA), IEEE, 2007.

[8] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer- Verlag, New York, 2001.

[9] T.S. Madhulatha, An overview on clustering methods, IOSR J. Eng. 2 (2012) 719725.

[10] D. Chen, K. Xing, D. Henson, L. Sheng, A.M. Schwartz, X. Cheng, Developing prognostic systems of cancer patients by ensemble clustering, J. Biomed. Biotechnol. 2009 (2009) (632786).

[11] D. Fradkin, Machine Learning Methods in the Analysis of Lung Cancer Survival Data, (2006) (February).

[12] C.M. Lynch, V.H. van Berkel, H.B. Frieboes, Application of unsupervised analysis techniques to lung cancer patient data, PLoS One 12 (2017) e0184370.

[13] V. Krishnaiah, G. Narsimha, N.S. Chandra, Diagnosis of lung cancer prediction system using data mining classification techniques, Int. J. Comput. Sci. Inf. Technol. 4 (2013) 3945.

[14] G. Dimitoglu, J.A. Adams, C.M. Ji, Comparison of the C4: 5 and a naive bayes classifier for the prediction of lung cancer survivability, J. Comput. 4 (2012) 19.