International Journal of Medical Informatics 108 (2017) 1–8

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Prediction of lung cancer patient survival via supervised machine learning classification techniques

Chip M. Lynch[a], Behnaz Abdollahi[b], Joshua D. Fuqua[c], Alexandra R. de Carlo[c], James A. Bartholomai[c], Rayeanne N. Balgemann[c], Victor H. van Berkel[d], Hermann B. Frieboes[c,e,*]

[a] Department of Computer Engineering and Computer Science, University of Louisville, KY, USA
[b] Department of Electrical and Computer Engineering, University of Louisville, KY, USA
[c] Department of Bioengineering, University of Louisville, KY, USA
[d] Department of Cardiovascular and Thoracic Surgery, University of Louisville, KY, USA
[e] James Graham Brown Cancer Center, University of Louisville, KY, USA

ARTICLE INFO

Keywords:
Lung cancer
SEER database
Machine learning
Data classification
Supervised classification
Biomedical big data

ABSTRACT

Outcomes for cancer patients have been previously estimated by applying various machine learning techniques to large datasets such as the Surveillance, Epidemiology, and End Results (SEER) program database. In particular 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 underperformed with an RMSE value of 15.82, statistical analysis singles the SVM as the only model that generated a distinctive output. The results of the models are consistent with a classical Cox proportional hazards model used as a reference technique. 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 dataset may be on par with that of classical methods.

1. Introduction

1.1. Background and study description

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, such as advertisement, insurance, finance, social media, and fraud detection, accessing various forms of data collected in real-time and across multiple sources. Using these techniques to evaluate disease outcomes can be challenging, however, since patient data is generally unavailable for public analysis. One exception is the Surveillance, Epidemiology, and End Results (SEER) program [1,2] from the National Cancer Institute (NCI) at the National Institutes of Health (NIH). As the largest publicly available cancer dataset [3], this database provides de-identified information on cancer statistics of the United States population, thus facilitating large-scale outcome analysis.

We apply machine learning techniques to this dataset to analyze data specific to lung cancer, with the goal to evaluate the predictive power of these techniques. Lung cancer was chosen as it ranks as a leading cause of cancer-related death, with dismal 5-year survival rates

* Corresponding author at: Department of Bioengineering, Lutz Hall 419, Louisville, KY 40208, USA.
E-mail address: hbfrie01@louisville.edu (H.B. Frieboes).

http://dx.doi.org/10.1016/j.ijmedinf.2017.09.013
Received 7 April 2016; Received in revised form 29 August 2017; Accepted 23 September 2017

[4]. The disease is typically classified as either Small Cell Lung Cancer (SCLC) or Non-Small Cell Lung Cancer (NSCLC) [5], with the diagnosis dependent on cellular physical appearance evaluated through histology [6]. The goal of identifying survivability given a specific medical diagnosis is of strong importance in improving care and providing information to patients and clinicians. Given a dataset of lung cancer patients with particular demographic (e.g., age), diagnostic (e.g., tumor size), and procedural information (e.g., Radiation and/or Surgery applied), the question is whether patient survival can be computationally predicted with any precision.

Although survival time analysis may be considered clinically important in order to evaluate patient prognosis, clinicians have struggled to estimate prognosis of lung cancer patients. In a recent study, physician consultants predicted a survival time median of 25.7 months, while physician registrars and residents predicted survival times of 21.4 and 21.5 months, respectively, for patients on average with 11.7 months actual survival [7]. The study also found that only ~60% of patients whose physicians estimated survival time > 3 months actually survived this long. Another study found that physicians correctly predicted survival time to the month 10% of the time, to 3 months 59% of the time, and to 4 months 71% of the time, and tended to overestimate short term survival times but underestimate long term survival times [8]. Applying a correlational approach via machine learning to predict survivability could help to improve such predictions.

In this study, patients diagnosed with lung cancer during the years 2004–2009 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 analyze 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.

1.2. Related work

Previously published work has analyzed the SEER database via statistical [9–16] as well as classification techniques [17–21]. In earlier work [22], the concept of agglomerative clustering [23,24] 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 [22].

Prediction models for breast cancer survivability using a large dataset were developed in [25] applying two popular data mining algorithms, artificial neural networks and Decision Trees, as well as a commonly used statistical method, logistic regression. Ten-fold cross-validation methods were employed to measure the unbiased estimate of the three prediction models for performance comparison purposes. The results showed that Decision Tree (C5) was the best predictor with 93.6% accuracy on the holdout sample, artificial neural networks were second best with 91.2% accuracy, and logistic regression models attained 89.2% accuracy. In [26] a study was performed to develop prediction models for prostate cancer survivability, employing support vector machines (SVM) in addition to the previously mentioned three techniques. In this case, the results singled out SVM as the most accurate predictor (92.85% accuracy), followed by artificial neural networks and Decision Trees [26]. Similarly, in [27] prostate cancer survivability was evaluated using artificial neural networks, Decision Trees, and logistic regression. Various techniques were compared in [28] using SEER colon cancer patient data to predict survival, finding that neural networks were the most accurate. In [29], ensemble voting of three best performing classifiers resulted in optimal prediction and area under the Receiver Operating Characteristic (ROC) curve for colon cancer survival.

A few studies have evaluated lung cancer patient survival by analyzing the SEER database with machine learning techniques, including ensemble clustering-based approaches [30], SVM and logistic regression [31], and unsupervised methods [32]. Data classification techniques were evaluated in [33] to determine the likelihood of patients with certain symptoms to develop lung cancer. In [34], the performance of C4.5 and Naïve-Bayes classifiers was compared applied to lung cancer data from the SEER database, achieving ~90% precision in predicting patient survival. In [19,35], ensemble voting of five Decision Tree based classifiers and meta-classifiers was determined to yield the best prediction of lung cancer survivability in terms of precision and area under the ROC curve.

Association rule mining techniques have been employed to determine interesting association or correlation relationships among a large set of items; different techniques to extract the rules and standard criteria have been proposed, suggesting how to choose the best rules and select optimizations based on a given dataset [36]. In [17], an automated technique to create a tree of rules for lung cancer was implemented, some of which were redundant and were manually removed based on domain knowledge. Three factors were considered: the maximum branching factors, adding a new branch, and the factor to be used when adding a new branch. The authors proposed a tree-based algorithm using the entire dataset from the very beginning, and descending into the data in a depth-first fashion using a greedy approach. Each node of the tree represented a segment and hence an association rule. The attributes included: age, birth place, cancer grade, diagnostic confirmation, farthest extension of tumor, lymph node involvement, type of surgery performed, reason for no surgery, order of surgery and radiation, scope of regional lymph node surgery, cancer stage, number of malignant tumor, and total regional lymph nodes examined.

Measuring the efficacy of treatments and surgery is a desired result from analyzing the SEER dataset, even though the dataset lacks information regarding chemotherapy. The effectiveness of treatment was considered in [37]. The study explored the question whether lung cancer patients survive longer with surgery or radiation, or both. A Propensity Score was used, representing a conditional probability that a unit will receive a treatment given a set of observed covariates. Two methods were applied for estimating the score, namely, logistic regression and classification tree. Since patients can receive surgery or radiation separately or together, the score was calculated for each group and then the attributes were ranked. Statistical information related to the combination of survival time and radiation was extracted, and a classification tree was generated for each group. The results showed that patients who did not receive radiation with or without surgery had the longest survival time [37].

2. Methods

For this study, we chose linear regression, Decision Trees, ensemble learning algorithms Random Forest and Generalized Boosting Machines as logic-based methods, Support Vector Machine (SVM) using a polynomial kernel function as a non-probabilistic method, and a custom ensemble method that used a weighting function to sum the predictions of each of the five individual models into a final prediction. Although a plethora of supervised techniques exist, these particular models were chosen because they represent a set of modern, commonly used methods.

R was used for implementation, as it is an open source statistical language with access to machine learning algorithms. For testing and comparison of the models, Root Mean Square Error (RMSE) values from