Prepared for submission to JHEP

Machine Learning for Survival Analysis: A New Approach

D. Dooling P. Green A. Kim D. Scroggin L. Stevens J. Webster

Innovative Oncology Business Solutions,
4901 Lang Ave NE,
Albuquerque, NM 87109, USA

E-mail: ddooling@innovativeobs.com, pgreen@innovativeobs.com, akim@innovativeobs.com, dscroggin@innovativeobs.com, lstevens@innovativeobs.com, jwebster@innovativeobs.com

Abstract: We have applied a little-known data transformation on subsets of the Surveillance, Epidemiology, and End Results (SEER) publically available data of the National Cancer Institute (NCI) to make it suitable input to standard machine learning classifiers. This transformation properly treats the right-censored data in the SEER data and the resulting Random Forest and Multi-Layer Perceptron models predict full survival curves. Treating the 6, 12, and 60 months points of the resulting survival curves as 3 binary classifiers, the 18 resulting classifiers have AUC values ranging from .765 to .885. Further evidence that the models have generalized well from the training data is provided by the extremely high levels of agreement between the random forest and neural network models predictions on the 6, 12, and 60 month binary classifiers.

Contents

1 Introduction and Background

Extracting actionable information from data is changing the fabric of modern business. A class of techniques that transforms data into actionable information goes by the name of Machine Learning [13]. Machine Learning has recently become a popular method to answer questions and solve problems that are too complex to solve via traditional methods. The Surveillance, Epidemiolgy, and End Results (SEER) Program of the National Cancer Institute (NCI) has been collecting data because intuitively researchers feel confident that this data is capturing information that has buried within it useful information in the form of relationships between the types of data collected (demographic as well as staging information) and the survival outcomes. Though this relationship evades capture by traditional methods, it is possible to surface it with the two machine learning techniques known as **Random Forests** and **Neural Networks**. These two methods produce very similar results when applied to the SEER dataset, and are based on two almost diametrically opposed learning philosophies, which lends confidence in the validity of the results.

The Surveillance, Epidemiolgy, and End Results (SEER) Program of the National Cancer Institute (NCI) is the most recognized authoritative source of information on cancer incidence and survival in the United States. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 28 percent of the US population.

Quoting directly from the SEER website [11]:

> The SEER program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status. This program is the only comprehensive source of population-based information in the United States that includes stage of cancer at the time of diagnosis and patient survival data. The mortality data reported by SEER are provided by the National Center for Health Statistics. The population data used in calculating cancer rates is obtained periodically from the Census Bureau. Updated annually and provided as a public service in print and electronic formats, SEER data are used by thousands of researchers, clinicians, public health officials, legislators, policymakers, community groups, and the public.

One characterstic of the SEER data and that is shared by many datasets in the medical field goes by the name of "censored data." The SEER data contains the number of months each patient survived, as well as an indicator variable showing whether or not the patient is still alive at the end of the data collection period. Methods to deal effectively with this kind of "right-censored data" include Kaplan-Meier curves and Cox's Proportional Hazard models [4]. The Kaplan-Meier techniques only give estimates for cohorts of patients and are not applicable for predicting the surival curve for a single patient, and the Cox Proportional Hazard models require a fairly restrictive set ot assumptions to be satisifed in order to yield reliable results. In addition, the Cox Proportional Hazard models are not able to capture the nonlinear relationships between the given data fields that go into making predictions; they can only capture the first-order linear relationships.

To overcome these limitations of the traditional methods, IOBS has applied a little-known technique to transform the SEER data to make it amenable to more powerful machine learning methods. The essential idea is to recast the problem to an appropriate discrete classification problem instead of a regression problem (predicting survival months). Treating months after diagnosis as just another discrete feature, the SEER data (or any other right-censored data) can be transformed simply so as to make predictions for the hazard function, probability of dying in the next month, given that the patient has not yet died. The full survival function can then be derived from the hazard function. Details of this transformation can be found in this blog post [1].

2 Methodology

2.1 Data acquisition

We used the publically available 1973-2012 SEER incidence data files corresponding to colon, breast and lung cancer contained in the following list. SEER requires that researchers submit a request for the data, which includes an agreement form. Detailed documentation explaining the contents of both the incidence data files used in this study as well as a data dictionary for the 1973-2012 SEER incidence data files are available without the need to register or submit a data request [12].

- incidence\yr1973_2012.seer9\COLRECT.txt
- incidence\yr1973_2012.seer9\BREAST.txt
- incidence\yr1973_2012.seer9\RESPIR.txt
- incidence\yr1992_2012.sj_la_rg_ak\COLRECT.txt
- incidence\yr1992_2012.sj_la_rg_ak\BREAST.txt
- incidence\yr1992_2012.sj_la_rg_ak\RESPIR.txt
- incidence\yr2000_2012.ca_ky_lo_nj_ga\COLRECT.txt
- incidence\yr2000_2012.ca_ky_lo_nj_ga\BREAST.txt
- incidence\yr2000_2012.ca_ky_lo_nj_ga\RESPIR.txt
- incidence\yr2005.lo_2nd_half\COLRECT.txt
- incidence\yr2005.lo_2nd_half\BREAST.txt
- incidence\yr2005.lo_2nd_half\RESPIR.txt

2.2 Data preparation and preprocessing

A great deal of data munging is necessary before using these SEER incidence files as input into machine learning algorithms. A preprocessing step common to each of three cancer types studied involves the `STATE-COUNTY RECODE`. The `STATE-COUNTY RECODE` field is a state-county combination where the first two characters represent the state FIPS code and the last three digits represent the FIPS county code. This particular field illustrates an important feature of machine learning, that between *categorical features* and *numeric features*. All input into a machine learning algorithm must be numeric, but real numbers carry with them the usually extremely useful property known as the well-ordering property of the real numbers. But if one is tasked with encoding a categorical feature into suitable numeric format for machine learning, it is necessary to do so in a way that removes the well-ordering property [2].

As a simple example of how to correctly treat categorical variables in a mchine learning context, consider the SEER variable `SEX`. This variable is encoded with a numeric 1 for males and a numeric 2 for females as shown in Table 1. Values such as "Male" and "Female" encoded as numbers are dangerous because if not handled properly, they can generate bogus results [7]. The proper way to transform the SEER `SEX` variable is to create two additional variables: `sex_Male` and `sex_Female`, and then to eliminate the variable `SEX`. For example,

Code	Description
1	Male
2	Female

Table 1. Encoding of gender in the SEER incidence files. These types of categorical variables need to be transformed via one-hot-encoding.

$$\left[\begin{array}{c} \text{SEX} \\ \hline 1 \end{array}\right] \rightarrow \left[\begin{array}{c|c} \text{sex\_Male} & \text{sex\_Female} \\ \hline 1 & 0 \end{array}\right] \tag{2.1}$$

and

$$\left[\begin{array}{c} \text{SEX} \\ \hline 2 \end{array}\right] \rightarrow \left[\begin{array}{c|c} \text{sex\_Male} & \text{sex\_Female} \\ \hline 0 & 1 \end{array}\right] \tag{2.2}$$

The procedure outlined in Equations (2.1, 2.2) needs to be applied to all of the nominal categorical variables in the SEER data that we wish to include in our predictive models. In particular, in order to include the geophgraphical information contained in the SEER categorical variable `STATE-COUNTY RECODE`, it becomes necessary to create a new feature variable for each of the distinct (state,county) pairs in the data. In the United States, there are approximately 3,000 counties. Clearly, transforming the `STATE-COUNTY RECODE` data representation into distinct (state_county) columns will explode the data to become wider than is optimal for machine learning. Adding extra columns to your dataset, making it wider, requires more data rows (making it taller) in order for machine learning algorithms to effectively learn [2]. Because one-hot coding `STATE-COUNTY RECODE` would cause such drastic shape changes in our data, we wish to avoid doing so. Fortunately, this variable, though given as a categorical variable, is actually an ordinal variable. There is an ordering among the (state_county) columns, name longitude, latitude, and elevation. We can transform the data in `STATE-COUNTY RECODE` into three new numerical columns: `lat`, `lng`, and `elevation`.

For example, Table (2) shows how five entries of `STATE-COUNTY RECODE` corresponding to counties within New Mexico would can be represented by the `elevation`, `lat`, and `lng` features.

It is a simple exercise to construct the full lookup table from the SEER `STATE-COUNTY RECODE` variable to the corresponding three values `elevation`, `lat`, and `lng`. Using the publically available dafafile from the United States Census Bureau [3] to construct query strings like the values of the `address` field in Table (2), it is possible to then programmatically query the Google Maps Geocoding API for the latitude and longitude [6], and the Google Maps Elevation API for the corresponding elevation [5]. An added benefit of this shift from the single categorical variable `STATE-COUNTY RECODE` to the three continuous numerical variables `lat`, `lng`, and `elevation` is that input into the web applications described later are not restricted to the states and counties covereed in the SEER registries. The

STATE-COUNTY RECODE	address	elevation	lat	lng
35001	Bernalillo+county+NM	5207.579772	35.017785	-106.629130
35003	Catron+county+NM	8089.242628	34.151517	-108.427605
35005	Chaves+county+NM	3559.931671	33.475739	-104.472330
35006	Cibola+county+NM	6443.415570	35.094756	-107.858387
35007	Colfax+county+NM	6147.749089	36.579976	-104.472330

Table 2. Example of the transformation of `STATE-COUNTY RECODE` to `elevation`, `lat`, and `lng`.

full lookup table analogous to Table 2 is available from a GitHub repository containing supplemental information for this study [8].

2.3 Colon Cancer Data

In this section we describe the data processing steps that were specific to the colon cancer model development. the files

- incidence\yr1973_2012.seer9\COLRECT.txt
- incidence\yr1992_2012.sj_la_rg_ak\COLRECT.txt
- incidence\yr2000_2012.ca_ky_lo_nj_ga\COLRECT.txt
- incidence\yr2005.lo_2nd_half\COLRECT.txt

were imported into a pandas DataFrame object. The following filter was then applied:

- `SEQUENCE NUMBER-CENTRAL` $\neq$ `"Unspecified"`
- `AGE AT DIAGNOSIS` $\neq$ `"Unknown age"`
- `BIRTHDATE-YEAR` $\neq$ `"Unknown year of birth"`
- `YEAR OF DIAGNOSIS` ≥ 2004
- `CS TUMOR SIZE EXT/EVAL` $\neq$ `""`
- `CS TUMOR SIZE` $\neq 999$
- `SEER RECORD NUMBER` $= 1$
- `PRIMARY SITE` $=$ `"LARGE INTESTINE, (EXCL. APPENDIX)"`
- `SEQUENCE NUMBER-CENTRAL` $= 0$

Model	6 Months AUC	12 Months AUC	60 Months AUC
Breast RF	.846	.885	.844
Breast NN	.855	.867	.836
Colon RF	.804	.806	.828
Colon NN	.797	.804	.841
Lung RF	.772	.796	.874
Lung NN	.765	.796	.875

Table 3. AUC values for the Random Forest and Neural Networks model binary classifiers derived from the full survival curve predictions; see text for details.

Cancer Type	% agreement 6 months	% agreement 12 months	% agreement 60 months
Colon	.981	.971	.915
Breast	.994	.984	.938
Lung	.861	.883	.900

Table 4. Percentage agreement for the Random Forest and Neural Network classifiers for 6, 12, and 60 month survival predictions on the test data for each cancer type.

3 Prediction Models

3.1 Random Forests

3.2 MLP Neural Networks

4 Performance Metrics

4.1 Model Agreement

5 Web Applications

6 Further Directions

Discussion of causality. A certain Marital status is not a "cause" of a better prognosis; c.f. Simpson's Paradox. Implementation of Judea Pearl's Causality Calculus.

A Selected Features

In this Appendix we explicitly list the features chosen for each of the Colon, Breast and Lung cancer predictive models. For each cancer type, the features chosen for the random forest and neural network models were the same, so as to be best be able to compare the two models.

A.1 Colon Cancer Feature Selection

Reference the GitHub Repository containing the notebooks.

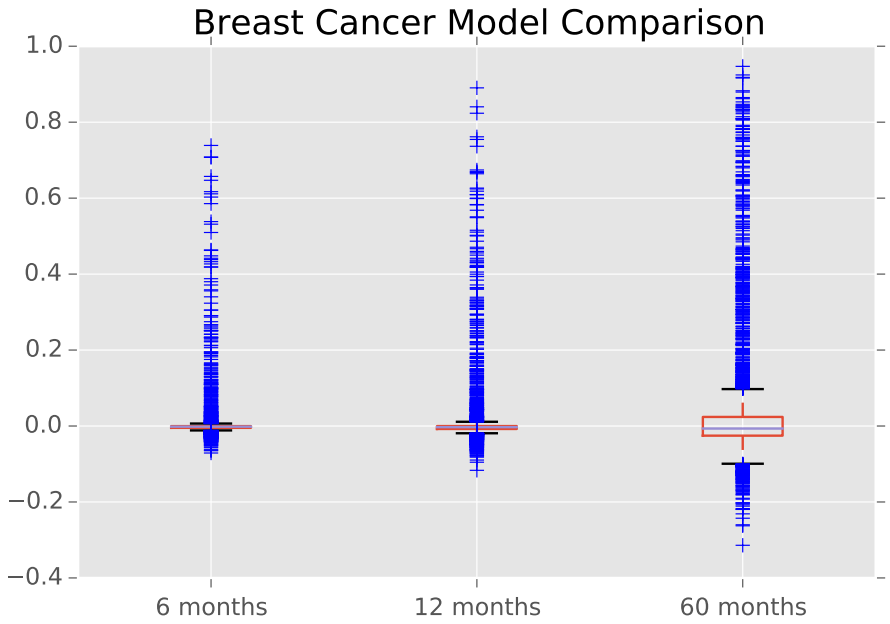

Figure 1. Box plots showing the distributions of the signed difference between the MLP model's prediction for the probability of surviving 6 months and the Random Forest model's prediction of the same quantity for breast cancer. The plot shows the same quantity for the 12 and 60 months classifiers. It is apparent from the figures that the outliers are due to the neural network models predicting higher survival probablitlies than the random forest for some few cases. These differences were evaluated for the 3300 test patients in the breast cancer data.

B Model Architecture and Python Code

C GitHub Repositories

Please always give a title also for appendices.

Acknowledgments

This is the most common positions for acknowledgments. A macro is available to maintain the same layout and spelling of the heading.

Note added. This is also a good position for notes added after the paper has been written.

References

[1] Ben Kuhn. Decision trees for survival analysis. http://www.benkuhn.net/survival-trees, year="2016 (accessed 14 Jan 2016)",.

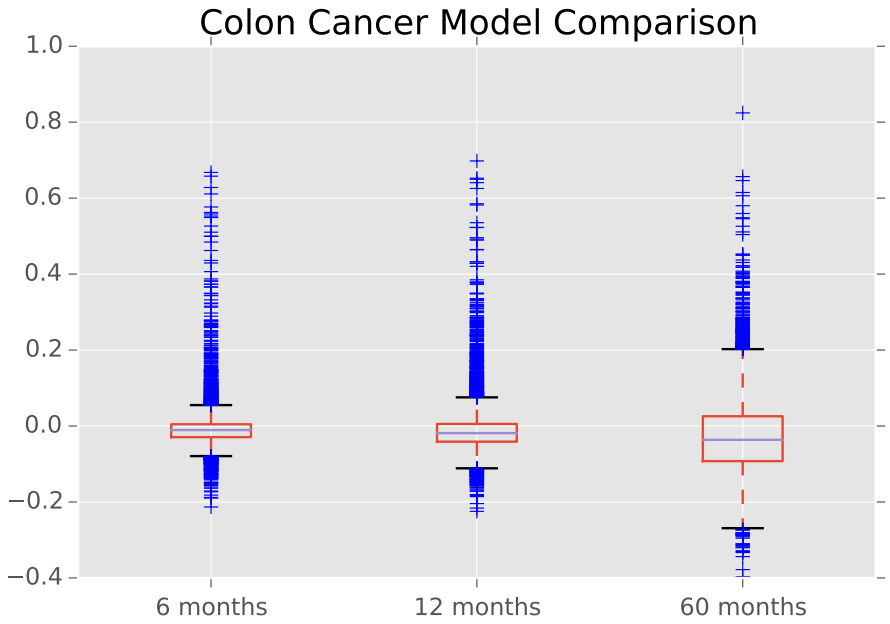

Figure 2. Box plots showing the distributions of the signed difference between the MLP model's prediction for the probability of surviving 6 months and the Random Forest model's prediction of the same quantity for colon cancer. The plot shows the same quantity for the 12 and 60 months classifiers. It is apparent from the figures that the outliers are due to the neural network models predicting higher survival probablitlies than the random forest for some few cases. These differences were evaluated for the 5654 test patients in the colon cancer data.

[2] Michael Bowles. *Machine Learning in Python: Essential Techniques for Predictive Analysis.* Wiley, 2015.

[3] United States Census Bureau. 2010 fips code files for counties - geography - u.s. census bureau. https://www.census.gov/geo/reference/codes/cou.html, 2016 (accessed 18 Jan 2016).

[4] Cam Davidson-Pilon. Quickstart – lifelines 0.8.0.1 documentation. http://lifelines.readthedocs.org/en/latest/Quickstart.html, 2016 (accessed 14 Jan 2016).

[5] Google Developers. The google maps elevation api | google maps elevation api | google developers. https://developers.google.com/maps/documentation/elevation/intro?hl=en, 2016 (accessed 18 Jan 2016).

[6] Google Developers. The google maps geocoding api | google maps geocoding api | google developers. https://developers.google.com/maps/documentation/geocoding/intro, 2016 (accessed 18 Jan 2016).

[7] Allen Downey. *Think Stats.* O'Reilly Media, 2014.

[8] IOBS. Supplemental material | paperdata.

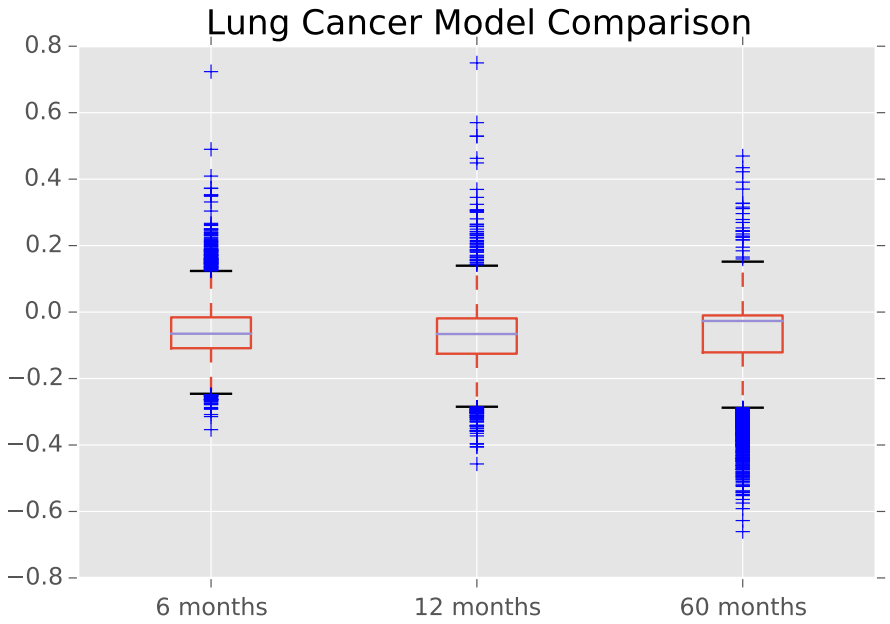

Figure 3. Box plots showing the distributions of the signed difference between the MLP model's prediction for the probability of surviving 6 months and the Random Forest model's prediction of the same quantity for lung cancer. The plot shows the same quantity for the 12 and 60 months classifiers. These differences were evaluated for the 5654 test patients in the colon cancer data. The Interquartile Ranges for lung cancer are visibly larger than those for breast cancer and colon cancer shown in fig 1 and fig 2.

https://github.com/doolingdavid/PAPERDATA.git, 2016 (accessed 18 Jan 2016.

[9] Rabbert Klein. Black holes and their relation to hiding eggs. *Theoretical Easter Physics*, 2010. (to appear).

[10] Johann A. Makowsky, Saharon Shelah, and Jonathan Stavi. Δ-logics and generalized quantifiers. *Annals of Mathematical Logic*, 10:155–192, 1976.

[11] National Cancer Institute, the Surveillance, Epidemiology, and End Results Program. About the SEER Program - SEER. http://seer.cancer.gov/about, 2016 (accessed 14 Jan 2016).

[12] National Cancer Institute, the Surveillance, Epidemiology, and End Results Program. Documentation for ASCII Text Data Files - SEER Datasets. http://seer.cancer.gov/data/documentation.html, 2016 (accessed 15 Jan 2016).

[13] Sebastian Raschka. *Python Machine Learning Essentials*. Packt Publishing, 2015.