Predicting the Likelihood of Survival from Thoracic Surgery Data

Samuel Jackson, Aberystwyth University

***Abstract*—Abstract goes here...**

I. Introduction

Thoracic surgery is a major invasive surgery involving operating on the lungs of a patient. The authors of ref. [1] collected several pieces of possibly relevant data on a number patients who went on to have thoracic surgery. The data also includes a record of whether a given patient survived for longer than one year after the surgery. This paper looks at using a reduced subset of the features and patients from the dataset in [1] to classify patients based on whether or not they will survive for one year after the surgery. This paper compares three different classifiers: random forests [2], extremely randomised trees [3], and gradient boosting [4].

The format of the result of this paper is structured as follows: section II outlines the preprocessing steps performed on the dataset and describes the classifiers used. Section III presents the performance of the classifiers on the dataset. Section IV discusses the results and presents possible justification for the performance based on the properties of the classifier and dataset. Finally, a summary and discussion of possible future directions are discussed in section V.

II. Methods

A. Dataset

The thoracic surgery dataset used consists of 16 predictors and 300 instances. Table I gives a description of each predictor derived from the original UCI dataset repository [5]. The dataset includes a mixture of both categorical (nominal and ordinal) and continuous data. The final (17^{th}) column of the dataset is the binary class label with value 0 if the patient survived and 1 if they died within one year of surgery.

TABLE I
Description of Columns in the Thoracic Surgery Dataset

Column	Type	Description
DGN	Nominal	Diagnosis: Specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1)
PRE4	Numeric	Forced vital capacity (FVC)
PRE5	Numeric	Volume that has been exhaled at the end of the first second of forced expiration (FEV1)
PRE6	Ordinal	Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)
PRE7	Nominal	Pain before surgery (T,F)
PRE8	Nominal	Haemoptysis before surgery (T,F)
PRE9	Nominal	Dyspnoea before surgery (T,F)
PRE10	Nominal	Cough before surgery (T,F)
PRE11	Nominal	Weakness before surgery (T,F)
PRE14	Ordinal	T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11, OC14, OC12, OC13)
PRE17	Nominal	Type 2 DM - diabetes mellitus (T,F)
PRE25	Nominal	Peripheral arterial diseases (PAD) (T,F)
PRE30	Nominal	Smoking (T,F)
PRE32	Nominal	Asthma (T,F)
AGE	Numeric	Age at surgery
Risk1Y	Nominal	1 year survival period - (T)rue value if died (T,F) (Class Label)

III. Results

IV. Discussion

V. Conclusions

Appendix A
IPython Notebook and Additional Python Files

Appendix B
Predictions for Testing Dataset

References

[1] M. Zięba, J. M. Tomczak, M. Lubicz, and J. Świątek, "Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients," *Applied soft computing*, vol. 14, pp. 99–108, 2014.

[2] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[4] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, 2013.

[5] "Thoracic Surgery Data Set," http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data, accessed: 2016-04-29.