McMaster University

Research Proposal

Using Machine Learning Methods to Predict the Risk of Death of Gastrointestinal Cancer Patients

Student:
Steve Cygu
400164479

Supervisor:
Prof. Jonathan Dushoff

ICES Scientist:
Dr. Hsien SEOW

May 15, 2019

1 Introduction

Machine learning (ML) techniques allows computers to 'learn' and discern patterns from large, noisy or complex datasets. This capability makes ML approaches best-suited to cancer prognosis and prediction. Application of various ML methods in development of cancer predictive models have resulted into improved understanding of cancer patients and better decision making. In addition, clinical management of cancer patients can be improved by early diagnosis and prognosis, especially, when correct information about the prognostic indicators are available. One of the problems which has led to emergence of applications of ML methods in cancer research is the need to classify cancer patients into low and high risk groups using robust techniques [4]. An example of such methods is machine machine learning models. These models can be used to identify important features contributing to tumor growth, diagnosis outcome and prediction [5].

Traditional models such Kaplan-Meier test and Cox-Proportional hazard models have been used to predict survival and identify prognostic indicators of cancer patients [2]. Compared to convectional survival models, ML methods have the capability of incorporating a wider range of covariates and identifying the most important ones. Traditional models do not predict the event but instead estimate the survival probability [1].

One of the challenges in accurate prediction of survival in cancer patients emerges from growing complexity of cancer, variant treatment options and heterogeneous patient populations. Therefore, accurate and reliable estimates could assist in designing personalised care and treatment, and improve institutional performance in cancer management. The use of ML methods in predicting risk of death of canter patients from clinical data is not a new one [3]. However, due to wide range of machine learning algorithms and improved validation methods there is a potential of more accurate predictions.

In this research, we intend to use various machine leaning methods, such as Neural Networks, Classification trees, Bagging, Boosting, Random Forest, $k-$Nearest Neighbours, Naive Bayes, Linear and Quadratic Discriminant Analysis to construct a classification model for Gastrointestinal Cancer Patients using *PROVIEW* datasets. The performance of the resulting models will be evaluated based on AUC and ROC, sensitivity, specificity and accuracy. The results will also be compared to Cox-Proportional hazard model estimates. We may also use clustering and/or Bayesian methods to jointly model the multiple outcomes.

2 Methods

2.1 Objectives

1. Use machine learning methods to predict: i) risk of death; ii) level of pain; iii) performance status and identify important factors for each of the 3 cases.

2. Jointly model multiple outcomes.

2.2 Analysis Plan

2.2.1 Predicting death

To incorporate censoring (presumably past 5 years) in ML approaches, we will follow the approach suggested by Zupan et al. [6]. The method involves assigning a distribution to the outcomes based on the Kaplan-Meier estimates. Suppose the censoring time, $T = 5$ years, the death outcome is split into three groups: the first consists of patients who died (outcome known); the second consists of patients who did not die but were followed upto the 5 years after diagnosis (not dying assumed); and the third consists of patients who did not die but were followed upto more than 5 years. For the last group, the probability of not dying at a particular followup time, T_f, is computed using Kaplan-Meier.

Patient-specific probability of not dying (the third group), P_s, is computed using the Kaplan-Meier estimates from all the three groups, i.e.,

$$\begin{aligned} P_s &= P(\text{not dying in 5 years}|\text{not dying in}(T_f)) \\ &= \frac{P(\text{not dying in 5 years})}{P(\text{not dying in}(T_f))}, \end{aligned}$$

and the probability of dying is thus $P_d = 1 - P_s$. The outcome for the particular patient is then a distribution (P_s, P_d). The entries for the patients in the third group are then modified by creating two copies: one labeled with the death outcome "yes", weighted by P_d; and the other labeled with the death outcome "no", weighted by P_s. This result will then be used for prediction using various machine learning methods.

2.2.2 Other outcomes

Various machine learning methods will be used to directly predict the outcomes.

2.2.3 Modeling multiple outcomes

Multiple outcomes are possibly correlated since they are related quantities collected from the same individual and may be useful when evaluating risk factors or characterizing treatment effectiveness, and as such jointly modeling might help in understanding the latent variations. In this research we will apply Bayesian Multivariate approaches to jointly model the multiple outcomes.

References

[1] Dursun Delen, Glenn Walker, and Amit Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127, 2005.

[2] Mogana Darshini Ganggayah, Nur Aishah Taib, Yip Cheng Har, Pietro Lio, and Sarinder Kaur Dhillon. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC medical informatics and decision making*, 19(1):48, 2019.

[3] Sunil Gupta, Truyen Tran, Wei Luo, Dinh Phung, Richard Lee Kennedy, Adam Broad, David Campbell, David Kipp, Madhu Singh, Mustafa Khasraw, et al. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ open*, 4(3):e004007, 2014.

[4] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[5] Mitra Montazeri, Mohadeseh Montazeri, Mahdieh Montazeri, and Amin Beigzadeh. Machine learning models in breast cancer survival prediction. *Technology and Health Care*, 24(1):31–42, 2016.

[6] Blaž Zupan, Janez DemšAr, Michael W Kattan, J Robert Beck, and Ivan Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine*, 20(1): 59–75, 2000.