Glioblastoma detection using Machine Learning *Draft*

Patric Fulop & Alex Agachi*

The University of Edinburgh, Empiric Capital

February 23, 2018

1 Introduction

In 2018, brain tumors remain one of the most intractable types of cancers. As difficult to treat surgically as to reach with radio or chemotherapy, survival expectancy is bleak. For glioblastoma, survival expectancy from diagnosis is 3 months without treatment and 14 months with (WHO 2014, Gallego 2015, Schapira 2007).

Statistics and medicine have a joint history spanning centuries. Models started being successfully applied to clinical datasets as far back as the late 1960s. And machine learning methods have been applied to clinical datasets from the 1990s onwards, with a pick up in pace mirroring that in other fields over the past couple of years. This has been a factor of the progress in modeling and computational ability in the machine learning field, and of the explosion in medical data, clinical and unstructured, as genetic and structured, in oncology.

However, the vast majority of studies applying machine learning to medical datasets focus on the most common cancers, particularly breast cancer. In this study we set out to apply several families of methods from the field of statistical learning to one of the largest, and the most unique, brain tumor database in the world. Our goals were twofold: survival prediction and patient clustering. With a bias towards supervised learning and inference, we also approached unsupervised learning for prediction accuracy optimization and clustering. In addition to exploring the application of different ML methods to a mix of clinical and genetic data for brain cancers, we had a clear aim of developing insights and tools that can be of use to clinicians. We are not aware of any comparable study having been undertaken to date. Through this study, we also aim to set a standard method of exploring a dataset of clinical and genetic data using ML tools.

2 Context

Statistical learning is the subfield of statistics focused on modelling and prediction. Simple methods include regression, classification, and clustering. The field can be further split into supervised methods and unsupervised methods.

2.1 Supervised Learning

In supervised approaches, the goal is to build a model for a dataset D comprising of the training set $\mathbf{X}$ and the response variable $\mathbf{Y}$ such that the model can correctly associate a new datapoint x^* with y^*. We refer to the two most common scenarios encountered, *classification* and *regression*.

In **classification**, the goal is to learn the parameters of your model that best associate your data with a discrete class response variable in order to make predictions on new datapoints, i.e. *"Does my point belong*

*Equal contribution by both authors

to class 0 or 1?". Of course, choosing a model is not straightforward and requires domain knowledge.

In the case of **regression**, the goal is similar, but the response variable is continuous instead of discrete. From a high level view, we want to find the parameters θ of a model $\mathbf{M}$ that maximizes the likelihood of the data observations under the parameters of that model.

$$\hat{y} = \hat{\mathbf{M}}(x) = argmax_{\theta} p(y|\mathbf{M}(x), \theta) \tag{1}$$

2.2 Unsupervised Learning

If we consider learning under supervision to be a problem of finding a response variable $\mathbf{Y}$ by modelling previously known observable relationships $(\mathbf{X}, \mathbf{Y})$, we can define unsupervised learning as the *much more difficult problem* of finding patterns in data without prior knowledge.

One simple example is clustering high-dimensional data-points in some input space by using some intrinsic distance metric within that space. The right number of clusters and the metric is what the "learning" problem consists of. Another example would be reducing the dimensionality of a dataset by projecting onto a smaller space that takes into account the correlations between the data-points, in that way pointless interactions can be neglected. Formally, we can formulate the problem as finding the parameters θ for a model $\mathbf{M}$ that accurately captures the relationships within the input space.

$$\hat{\mathbf{M}}(x) = argmax_{\theta} p(\mathbf{M}(x), \theta) \tag{2}$$

There is more to be said about the intricacies of how to choose a good model for your problem, what training approach you deploy, maximum likelihood estimation (MLE) or incorporating priors and performing maximum a posteriori (MAP) estimation and whether you add flexibility by allowing your parameter space to change with the number of observations, i.e. parametric (fixed) or non-parametric models (varies with data). We refer the reader to Murphy's/Barber's excellent introduction [**?**] for a more detailed treatment.

3 ML applied to survival prediction and clustering

While medical research is steeped into careful data analysis and statistical analysis, it missed the nonlinear revolution of the past couple of decades. Most medical studies are still done on static datasets using the same linear analysis tools that were already in use a century ago.

There have been several meta studies of applications of machine learning to oncological datasets, notably: Cruz and Wishart (2006), Kourou et al. (2015), and focused on breast cancer studies specifically, Abreu et al (2016). Cruz and Wishart found more than 1500 papers on the topic of machine learning and cancer. However, the vast majority of them were focused on cancer diagnosis, with a small majority in the field of ML and cancer prediction/prognosis and even smaller combining clinical with molecular data. The authors reached several important conclusions:

1. Poor quality and poor validation of many studies, the lack of sufficient external validation or too few observations
2. Machine learning can substantially improve accuracy of predicting cancer susceptibility, recurrence and mortality (15-20%)

Kourou et al. performed a meta study of applications of machine learning to cancer, including diagnosis, and prediction/prognosis, which they further divided among susceptibility, recurrence and survival. They limited themselves to recent works, defined as those published in the five years prior to their publication dated 2014. While they note that one of the most common limitatations of the studies reviewed was the small size of the dataset, they highlighted a tendency towards increasing quality in such studies, with the use of multiple models, evaluation techniques and validation sets. While they found a very small amount of studies focusing on cancer survival prediction, and an even smaller one mixing clinical and molecular data, and each of them with very small datasets, these suggested that very high prediction accuracies could be reached by the models tested. These were as high as 98% for oral cancer (Rosado P et al).

While the meta study by Abreu et al (2016) focused on the prediction of breast cancer recurrence specifically, we decided to include it here due to its very high quality. We believe it is a model for meta studies of the cancer and ML literature. They follow a modelling logic in dividing their meta studies, including the relevant steps of building a model from feature selection to data cleaning to sampling to evaluation. For each, they provide context and summarize the approaches taken by the studies reviewed along with helpful commentary on the positives and less positive aspects.

There has been significant progress in recent years on applications of machine learning methods to the diagnosis of brain cancer, mostly focused on medical imagery (Biros et al. BRaTS?s 17; Erickson et al NVIDIA?s 2017 Global Impact Award; Panca and Rustam 2017). However, work on applying machine learning to brain cancer survival prediction and clustering is almost non existent. We came across one such study (Ma et al. 2017), which mixed molecular and clinical data for GBM cases from the Cancer Genome Atlas database, in order to predict survival. The study achieved AUC scores of 0.82 when mixing clinical and somatic copy-number alterations data, and of 0.98 when mixing microRNA data with clinical data.

Our review of the relevant literature highlighted several relevant problems.

1. Most application of machine learning tools to cancer survival prediction to date focused on the most common types of cancer, in particular breast cancer. In fact in the two general meta studies listed above, only one study focusing on brain cancer was referenced, Wei et al (2004, in Cruz and Wishart). Brain cancer survival prediction is very much absent from the literature at this point.
2. Many studies are plagued by quality problems, in particular the tiny datasets used
3. Each study uses different ML methods and variants. And indeed, trying different models and optimisations is a standard part of an ML workflow. And the science and art behind this is what makes machine learning such a challenging field still. Much more problematic is that even among the high quality studies, each study uses its own assessment metrics as a subjective choice among the set of possible ones. This makes assessment and comparison extremely difficult and it?s an easily avoidable problem. We propose that all studies, for whichever models they apply, stipulate clearly all standard assessment metrics for their models: sensitivity, specificity, ROC, and accuracy.

We could further add that almost no study makes the code used available ? another poor choice that detracts from the entire field in our opinion.

4 Dataset

Later on

4.1 Preprocessing steps and incorporating domain knowledge

We add the details later on, as seen in ML Cancer draft section 2. It will contain all steps from before the freeze.

Some questions remain:

- Should we add 2014-2015 for derniere nouvelles?
- Codeletion 1p19q mostly null?

4.2 Missing data

The dataset, in particular the genetic part, was very sparse. Several genetic indicators lacked data for approximately 50% of observations and one of them, Gene Mgmt, lacked data for 83% of observations. Complete case analysis was impossible even if considered, as no observation had data across all variables. As such, a statistically robust method for handling the missing data was not only desired but absolutely required. We settled on a type of fully conditional specification (FCS), multiple imputation by chained equations (MICE). For details on both the theory and implementation in R of this technique see Van Buuren (1999, 2007, 2009, 2011, 2012).

5 Modelling

Goals of the study

1. **Exploration:** Which predictors are associated with the target variable? How strongly are they related? Is the relationship between predictors and target linear? probably not How much of the variability in the response does this dataset explain? Does survival expectancy naturally group itself into several categories so that we can use QDA or a KNN classifier?

2. **Prediction:** Predict survival time in months for patients based on input data. We will use a regression model as well as a classification model. Clustering patients might prove beneficial when the target variable will be interval of survival, i.e. patients with survival between 3 months and 6 months. Performing categorical PCA (Multiple correspondence analysis) will perhaps prove beneficial. Naive Bayes too, if that's not too strong of a an assumption.

5.1 Decision Trees and Random Forests

Boosting, bagging, AdaBoost DT give you interpretability and more. Can apply an ensemble model of GNB and Random Forests?