Artificial Intelligence in Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Bayesian averaging over Decision Tree models for trauma severity scoring

V. Schetinin [a,*], L. Jakaite [a], W. Krzanowski [b]

[a] University of Bedfordshire, United Kingdom
[b] University of Exeter, United Kingdom

ARTICLE INFO

Article history:
Received 28 February 2017
Received in revised form 4 December 2017
Accepted 13 December 2017

Keywords:
Bayesian method
Decision Tree
Predictive posterior distribution
Injury severity scoring

ABSTRACT

Health care practitioners analyse possible risks of misleading decisions and need to estimate and quantify uncertainty in predictions. We have examined the "gold" standard of screening a patient's conditions for predicting survival probability, based on logistic regression modelling, which is used in trauma care for clinical purposes and quality audit. This methodology is based on theoretical assumptions about data and uncertainties. Models induced within such an approach have exposed a number of problems, providing unexplained fluctuation of predicted survival and low accuracy of estimating uncertainty intervals within which predictions are made. Bayesian method, which in theory is capable of providing accurate predictions and uncertainty estimates, has been adopted in our study using Decision Tree models. Our approach has been tested on a large set of patients registered in the US National Trauma Data Bank and has outperformed the standard method in terms of prediction accuracy, thereby providing practitioners with accurate estimates of the predictive posterior densities of interest that are required for making risk-aware decisions.

1. Introduction

Health care systems based on Machine Learning (ML) technologies are increasingly demanded for the prevention of lifestyle-related and chronic diseases as well as for emergency care and life support, see e.g. [1,2]. This interest is explained by the need for efficient access to data related to patients' conditions in ambulance, hospital or home environments. To assist health care practitioners with decision making, these data can be analysed using various ML approaches.

Million people worldwide are injured and admitted to hospitals for emergency treatment. Just in 2014, around 40 million people were treated in the US and 192,945 of them obtained fatal injuries and died [3]. Reliable, accurate, and timely information about a patient's condition is therefore of critical importance for improving trauma care outcomes.

For evaluation of injury severity and prediction of survival, practitioners exploit a logistic regression model known as the Trauma and Injury Severity Score (TRISS). The TRISS model predicts the probability of survival for a patient on arrival at a hospital, see e.g. [4–7]. The prediction model combines screening information and physiological parameters recorded by a paramedic at an accident scene. For some patients, data that are obtained with medical devices, such as blood pressure and heart rate, can be missing at the moment of the examination.

Uncertainties that exist in data as well as in the prediction model will affect the results and might lead to fatal errors or inadequate treatment. For this reason, practitioners have raised a concern about the ability of TRISS to provide reliable and accurate predictions and estimates of uncertainty [8,5].

The accuracy of predictions is compared against actual survival during model calibration. A relationship between the predicted and actual probabilities can be visualised as a calibration curve [9]. In this sense, the TRISS calibration curve has drifted away from the ideal curve, see e.g. [10,6,8].

In [5], it has been found that the accuracy of TRISS predictions is acceptable when the types and severities of patient's injuries are typical. However, for patients with four or more injuries as well as those with atypical combinations of injuries, the accuracy has to be improved. In practice, it is critically important to accurately estimate the uncertainty in a predicted survival probability. The uncertainty estimates are required in order to minimise risks of fatal errors. Uncertainty can be represented by confidence intervals. These intervals are reliably estimated when the density of predicted probabilities is fully tractable, which is achievable only in trivial cases. Thus TRISS methodology that is based on theoretical

* Corresponding author.
E-mail address: v.schetinin@beds.ac.uk (V. Schetinin).

https://doi.org/10.1016/j.artmed.2017.12.003
0933-3657/

assumptions cannot realistically estimate the uncertainty [11,12]. To tackle the above problems, we employ the Bayesian approach to learn prediction models from data. This methodology in theory provides the most accurate predictions and uncertainty estimation, see e.g. [13–16]. This approach, however, requires intensive computations, see e.g. [17–19]. In our approach we use Bayesian averaging over Decision Tree (DT) models, also known as Classification and Regression Trees, which are well-known for their ability to select input variables that maximally improve the performance [20–22]. DT models split the given data along input variables recursively, which is relatively simple to compute. The strategy, however, cannot provide a global view on the entire data. At the same time, the partitions which are made along variables are transparent, and when the number of the partitions is reasonably small then DT models can assist users with new insights into the data, see e.g. [23].

We analyse existing approaches and describe our approach based on Bayesian averaging over DT models. Then we test and compare the proposed and TRISS methods on the main trauma data benchmark, the US National Trauma Data Bank (NTDB) [24]. The comparison of the methods is made in terms of the Area Under the receiver operating characteristic Curve (AUC), which is a summary measure of the accuracy of a quantitative diagnostic test, see e.g. [25]. Finally, we discuss a DT model that can be used for purposes of interpretation with the maximum predictive ability.

2. Logistic regression model for predicting survival probability

Logistic regression modelling is a way of calculating probabilities of survival for given predictors, see e.g. [4,9]. As such, the TRISS model includes both continuous and categorical tests. The former include: age, systolic blood pressure, and respiratory rate, while the latter include: severity scores of injuries that a patient can obtain, the Glasgow Coma Scale (GCS), and the type of injury. Screening tests are evaluated on the patient's arrival by a trained scorer, see e.g. [6].

The above screening tests form two aggregated predictors: Injury Severity Score (ISS), and Revised Trauma Score (RTS). However, practitioners have found that such an aggregation causes unexplained fluctuations of the ISS over observed probabilities of survival, which affects the prediction accuracy, see e.g. [5,26]. The calculation of survival probabilities has been made available online as a TRISS Calculator [27].

The current standard TRISS allows for up to three of the most severe injuries that a patient can obtain in six regions of the body: head, face, chest, abdomen, extremities, and external (skin, subcutaneous tissue and burns).

Within this methodology, a density of predicted values is assumed to be a Gaussian distribution, $N(\mu, \sigma^2)$, where μ and σ^2 are defined by the parameters b and by the regression error, respectively. As follows from [28], under such an assumption, the uncertainty interval for a prediction cannot be realistically estimated for a patient.

3. Methodology

In most practical cases, any given model is incapable of fully explaining the real-world data, which means that a single "true" model does not exist. The method of Bayesian averaging over models, adopted in our study, assumes that different models can be mixed together so that their average under certain conditions will approximate the true model of interest. The averaging strategy is often more efficient than model selection in real-world applications when the predictive ability (or fitness function) is not unimodal, see e.g. [28,16].

The use of DT models within Bayesian method gives us the following advantages [20,29]. In comparison with other Machine Learning methods, the DT technique is directly applied to the given data without time-consuming data preprocessing or careful tuning of the learning algorithm, and so the DT technique is often called "off-the-shelf". The following advantages could be also important when dealing with real-world problems. DT models are robust to outliers in the given data. When a domain problem is represented by a mix of numerical and categorical variables, the DT technique naturally captures the relationships between them.

DTs perform internal feature selection as an integral part of the learning procedure, and so the use of general data transformations, such as Principal Component Analysis, is not required, see e.g. [30].

Models learnt from given data are calibrated and their accuracy is statistically evaluated by goodness-of-fit tests. In the medical domain the calibration is usually assessed via the Hosmer-Lemeshow (HL) statistic [9,8]. HL statistics are typically calculated for 10 intervals of predicted values. Under certain conditions, the larger the HL statistic, the worse is the calibration. The HL-test, however, is statistically significant in 100% of models when the number of patients is 50,000 or more. So this test has to be analysed along with the overall number of patients, see e.g. [31], which has been taken into account in our experiments.

The HL-test of goodness-of-fit is typically used along with others metrics of medical decision-making models, such as sensitivity and specificity, True Positive (TP) and False Positive (FP) rates. We also compare the diagnostic potentials of the proposed and existing methods in terms of AUC as discussed in Section 1.

Table 1
Screening tests and ranges of NTDB.

#	Name	Min	Max
1	Age	0	100
2	Gender	0 female	1 male
3	Injury type	0 penetrating	1 blunt
4	Blood pressure	0	300
5	Respiration rate	0	200
6	GCS Eye	1	4
7	GCS Verbal	1	5
8	GCS Motor	1	6
9	Head severity	0	6
10	Face severity	0	4
11	Neck severity	0	6
12	Thorax severity	0	6
13	Abdomen severity	0	6
14	Spine severity	0	6
15	Upper extremity severity	0	6
16	Lower extremity severity	0	6
17	External severity	0	6

4. Data

For comparison of the proposed and standard TRISS methods, we use a set of patient records from the US NTDB, the major source of data about injured patients admitted to hospitals and emergency units [24]. The data include patient age, gender, type and regions of injuries along with some clinical and background information about patient state. The NTDB also includes the TRISS prediction and the outcome of care, alive or died, for each patient.

Table 1 shows the screening tests (or predictors) that are used by the TRISS method based on the NTDB. The variables *Age*, *Blood pressure*, and *Respiration rate* are continuous, and the remaining variables are categorical. The patient outcome is the *discharge status*, $y \in \{0, 1\}$, where 0 is alive, and 1 is died. The table also shows the minimal and maximal values of each test.

Please cite this article in press as: Schetinin V, et al. Bayesian averaging over Decision Tree models for trauma severity scoring. Artif Intell Med (2017), https://doi.org/10.1016/j.artmed.2017.12.003