Patient Specific Predictions in the Intensive Care Unit Using a Bayesian Ensemble

Alistair EW Johnson[1], Nic Dunkley[1], Louis Mayaud[1], Athanasios Tsanas[1], Andrew A Kramer[2], Gari D Clifford[1]

[1] University of Oxford, Oxford, United Kingdom
[2] Cerner Corporation, Vienna, VA, United States

Abstract

Introduction: An intensive care unit mortality prediction model for the PhysioNet/Computing in Cardiology Challenge 2012 using a novel Bayesian ensemble learning algorithm is described.

Methods: Data pre-processing was automatically performed based upon domain knowledge to remove artefacts and erroneous recordings, e.g. physiologically invalid entries and unit conversion errors. A range of diverse features was extracted from the original time series signals including standard statistical descriptors such as the minimum, maximum, median, first, last, and the number of values. A new Bayesian ensemble scheme comprising 500 weak learners was then developed to classify the data samples. Each weak learner was a decision tree of depth two, which randomly assigned an intercept and gradient to a randomly selected single feature. The parameters of the ensemble learner were determined using a custom Markov chain Monte Carlo sampler.

Results: The model was trained using 4000 observations from the training set, and was evaluated by the organisers of the competition on two new datasets with 4000 observations each (set b and set c). The outcomes of the datasets were unavailable to the competitors. The competition was judged on two events by two scores. Score 1 was the minimum of the positive predictive value and sensitivity for binary model predictions, and the model achieved 0.5310 and 0.5353 on the unseen datasets. Score 2, a range-normalized Hosmer-Lemeshow C statistic, evaluated to 26.44 and 29.86. The model was re-developed using the updated data sets from phase 2 after the competition, and achieved a score 1 of 0.5374 and a score 2 of 18.20 on set c.

Conclusion: The proposed prediction model performs favourably on both the provided and hidden data sets (set A and set B), and has the potential to be used effectively for patient-specific predictions.

1. Introduction

The intensive care unit (ICU) admits only the most severely ill patients who require life-sustaining treatments or extensive monitoring. Each patient receives a level of clinical care greater than that on a general medical-surgical floor. As such, the acquisition and storage of data collected from ICU patients could provide the research community with a rich data source for developing predictive models of patient outcomes. However, data emanating from ICUs are not always in an easily accessible format, and the fusion of such data into a form amenable to analysis is a primary task in the development of many clinical prediction systems.

The patient outcome that has been the focus of much predictive model development is mortality before hospital discharge. Early models include the Acute Physiology, Age, and Chronic Health Evaluation system [1], the Simplified Acute Physiology Score [2], and the Mortality Prediction Model [3]. These models have proven useful for comparing observed vs. predicted outcomes across ICUs and thus were used for benchmarking purposes. However, none of these models contain sufficient precision to be used on an individual patient level. More recent versions of these predictive models [4–6] address the tendency of predictive models to erode in calibration over time [7], but have not reached sufficient accuracy to be used in determining the appropriate clinical care for a patient. The goal of the Physionet 2012 competition was to encourage development of models whose main purpose was patient-specific mortality prediction.

2. Materials and methods

2.1. Dataset

This study used two datasets provided for the Physionet 2012 challenge. These datasets were originally extracted by the Physionet 2012 challenge coordinators from the open access Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II database, which was developed

to aid intelligent patient monitoring research in the critical care environment [8]. Three datasets were extracted from MIMIC II by the competition organizers, and are referred to as set A, set B, and set C. All datasets comprised of 4000 patient stays in the ICU lasting at least 2 days. The data were formatted as time-stamped measurements for 37 distinct variables. Furthermore, measurements for 5 static variables which were collected once at the beginning of the patient's ICU stay are also present in the datasets. Set A was made available for model development, with corresponding hospital mortality outcomes provided for each patient. A positive outcome indicates that the patient died in the hospital. Set B was also made available for model testing, and as such no outcomes were provided. The third dataset, Set C, was not available but used to evaluate the final models during the Computing in Cardiology 2012 conference.

2.2. Data preprocessing

Prior to using the dataset, each subjects' measurements were assessed using domain knowledge and distributional assumptions. That is, we impute physiologically implausible values for certain variables, either assigning the entry as unknown or substituting it with a physiologically valid entry. An example substitution would be converting height from an implausible value (e.g. 65, presumably an erroneous recording using inches) to a plausible value (e.g. 165 centimetres). This variable specific pre-processing was performed for age, height, diastolic blood pressure, heart rate, partial pressure of carbon dioxide, partial pressure of oxygen, hydrogen ion concentration, temperature, troponin I, white blood cell count, and weight. Missing data was present in the data and handled by the model.

2.3. Variable extraction

We converted each patient's time-stamped temporal variables into scalar features. For the static variables; age, initial weight, height, and gender, this was done by directly treating the value present as a feature. For the temporal variables, this was done by extracting the minimum, maximum, median, first, last, and number of values for each time stamped variable. Thus, for each temporal variable, six features were extracted.

2.4. Evaluation metrics

The agreement between predicted binary outcomes (after thresholding a predicted probability) and observed outcomes was assessed by the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). True indicates that the prediction and the observed outcome agreed, positive indicates an outcome of one, and negative indicates an outcome of zero.

For the Physionet 2012 competition, two official metrics were used to assess the performance of the submitted models. Score 1 (s_1) evaluates model *discrimination*, and score 2 (s_2) the model *calibration*. Further to these two official scores, additional metrics were used, including the normalized log-likelihood (**NLL**) and the area under the receiver operator characteristic curve (AUROC).

s_1 is the maximum of the sensitivity (Se) and the positive predictivity (PPV) at a given operating point. s_2 is a modified version of the Hosmer-Lemeshow $\hat{C}$ statistic ($HL_{\hat{C}}$) [9]. The $HL_{\hat{C}}$ involves grouping predictions into deciles of predicted risk, and calculating the mean error within each decile. s_2 is derived from the $HL_{\hat{C}}$ by further dividing the statistic by the difference between the mean prediction in the highest decile and the mean prediction in the lowest decile and a constant of 0.001. The AUROC is the probability that a patient with a positive outcome is given a higher probability of mortality than a patient with a negative outcome. Mathematically, this is interpretable as the $Pr(X = 1) > Pr(Y = 1)$, where X is the set of patients with observed positive outcomes, and Y is the set of patients with observed negative outcomes. The **NLL** is a metric based on information theory ranging between 0-1, with lower values indicating better model performance. The formulas for calculating these statistics are shown in Table 1.

Statistic	Equation
Sensitivity (Se)	$\frac{TP}{TP+FN}$
Positive Predictive Value (PPV)	$\frac{TP}{TP+FP}$
AUROC	$\sum_{i=1}^{N}[\sum_{j=1}^{M}\mathbf{1}(\mathbf{X_i}>\mathbf{Y_j})]$
NLL	$\frac{\sum_{i=1}^{N}(y_i ln(p_i)-(1-y_i)ln(1-p_i))}{N\times M}$
Hosmer Lemeshow $\hat{C}$	$\sum_{j=1}^{D}\frac{O_j-E_j}{n_j p_j(1-p_j)+0.001}$
Score 1 (s_1)	$min(Se, PPV)$
Score 2 (s_2)	$\frac{HL_{\hat{C}}}{p_D-p_1}$

Table 1. List of the various performance evaluation metrics used in this study and the formulae for calculating them.

2.5. Model description

A tree based classifier was developed using a Bayesian framework. The algorithm has many advantages, including

high overall performance and automatic handling of missing data, outliers, and normalization. Each tree selects a subset of observations via two regression splits. These observations are then given a contribution equal to a random constant times the observation's value for a chosen feature plus a random intercept. Furthermore, the tree also assigns a contribution to missing values for this chosen feature based upon a scaled surrogate. The contributions across all trees are summed to provide the contribution for a single "forest", where a "forest" refers to a group of trees plus an intercept term. The predicted probability output by the forest is the inverse logit of the sum of each tree's contribution plus the intercept term. The intercept term is set to the logit of the mean observed outcome.

The core of the new model is the custom Markov chain Monte Carlo sampler which iteratively optimizes the forest. This sampling process has a user defined number of iterations and a user defined number of resets (each reset involves reinitializing the forest and restarting the iterative process). After mapping the training data onto the quantiles of a normal distribution, the forest is initialized to a null model, with no contributions assigned for any observations.

At each iteration, the algorithm selects two trees in the forest and randomizes their structure. That is, it randomly re-selects first two features which the tree uses for splitting, the value at which the tree splits those features, the third feature used for contribution calculation, and the multiplicative and additive constants applied to the third feature. The total forest contribution is then recalculated and a Metropolis-Hastings acceptance step is used to determine if the update is accepted. If the update is accepted, the two trees are kept in the forest, otherwise they are discarded and the forest remains unchanged. After a set fraction of the total number of iterations to allow the forest to learn the target distribution (20%), the algorithm begins storing forests at a fixed interval, i.e. once every set number of iterations. Once the number of user-defined iterations are reached, the forest is re-initialized as before, and the iterative process restarts. Again after the set burn-in period, the forests begin to be saved at a fixed interval. The final result of this algorithm is a set of forests, each of which will contribute to the final model prediction.

In order to provide a binary prediction of survival, it was necessary to threshold the risk value (a probability between 0 and 1). The threshold was that which maximized the *estimated* value of s_1 on the test set. The calculation of the estimated s_1 for each risk threshold is identical to the calculation of s_1, except the true positive and false positive values are estimated from the predicted risks. The estimated true positives and estimated false positives at each risk value were calculated by sorting the risks across all patients and cumulatively summing the risks and the complement of the risks.

2.6. Model development and assessment

In order to estimate the performance of the model, jackknifing was performed. Each jackknife iteration involved redeveloping the model for randomly subsampled sets of Set A, followed by evaluating the predictive performance on the out of sample subset. These subsets are referred to as the training and validation sets, respectively. The training sets included 3000 patients while the validation sets included 1000 patients, repeated 32 times to assess the variability of the evaluation metrics.

The final model submitted to the competition utilized all 4000 patients in set A in an earlier version of the dataset. The threshold for survival was set based upon the predictions on the 4000 patients in set B. Though the dataset was modified during the competition, the entry submitted was trained using the earlier dataset. Additional results are provided for a model developed after the competition close, using all 4000 patients in the updated version of set A. The threshold was re-calculated for this model.

3. Results

The evaluation metrics on the out of sample data for set A are shown in Table 2 for the proposed model and the sample model provided by Physionet (SAPS). For set A, the data shown are the mean and standard deviation from 32 jacknknife repetitions of model development and evaluation. The evaluation metrics on all the data for set B and set C are shown in Table 3. The final entry achieved a s_1 of 0.5310 for set B and a s_1 of 0.5353 for set C. Furthermore, the entry achieved a s_2 of 26.44 for set B and a s_2 of 29.86 for set C. The threshold for the final entry was 0.3380 as chosen using the estimated s_1 on set B.

Table 2. Evaluation metrics of SAPS and the developed model on set A (1000 out of sample observations). None of the observations evaluated by the metrics were used in the model development.
*Set A statistics are presented as the mean and standard deviation from 32 jackknife repetitions.

Metric	SAPS Set A	Set A*	Set A 95% CI
AUROC	0.6668	0.8602 ($\pm$ 0.014)	5.13×10^{-3}
NLL	0.4023	0.2891 ($\pm$ 0.017)	5.95×10^{-3}
s_1	0.2957	0.4846 ($\pm$ 0.032)	1.14×10^{-2}
s_2	69.001	16.825 ($\pm$ 9.72)	3.505

The model developed after competition close using the same methods but with the newer data set achieved a s_1 of 0.5353 and an s_2 of 13.67 on set B and a s_1 of 0.5374 and an s_2 of 18.20 on set C.

Table 3. Evaluation metrics on set B (4000 observations) and set C (4000 observations) of the competition for the model (data not used for training).

Metric	Set B	Set C
s_1	0.5310	0.5353
s_2	26.44	29.86

4. Discussion and conclusions

Jackknifing is a common method used to assess both the inter-observation variability inherent in the data and the projected accuracy on an unseen set of observations. Since jackknifing requires a sub-sample of the training set, it is not unexpected for a model developed using the full training set to generalize better, as seen by the improvement of score 1 when developing using all observations.

It is worth noting the variability present in the competition metrics as revealed by the jackknife assessment. The variability of s_2 is particularly striking, and the metric is very sensitive to subtle variations in the evaluated data. This is likely due to the binning procedure of the observations. The increased variability of s_1, though less severe as s_2, may be caused by the added variability of the selection of a threshold, and the loss of numerical accuracy by rounding the final predictions. This increase in variability adds difficulty in distinguishing a model's superiority due to its methods or due to simple random chance. It is thus unsurprising to see the drastic changes in s_2 on set B and set C, as this may just be due to random chance and the large variability inherent to the metric. Conversely, the AUROC and the normalized-log likelihood both have the lowest variability of all the metrics. The AUROC is advantageous due to being a widely recognized and understood metric. The normalized log-likelihood, though not as ubiquitous, has the added advantage of also assessing the model calibration. For example, a model may generate predictions between 0-0.2 and still have an excellent AUROC as it distinguishes positive outcomes from negative outcomes well. The same does not hold true of the model's normalized log-likelihood, which effectively evaluates both a model's discrimination and a model's calibration. Nonetheless, the determination of a standard metric for evaluating mortality prediction systems is still an unsolved problem.

The lower than expected s_2 of the model is explainable by it being developed on an earlier version of the data set and utilizing information which was later removed. While the evaluation data sets were updated, the model still expected to use the missing information in calculating predictions. After the competition, the model was re-developed and the s_1 improved by 0.0021 while the s_2 improved by 11.65.

The proposed model had extremely good overall performance, with a median AUROC of 0.860, achieving much better performance than the SAPS sample model (AUROC of 0.667). The model has many advantages, such as automatic handling of missing data, and utilizing easily extractable features from regularly collected clinical parameters. The model's predictions discriminate patient mortality extremely well, and calibrate well. The model provides a promising new method of patient specific mortality prediction and decision support at the bed side.

Acknowledgements

AJ and LM acknowledge the support of the RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation). AT was funded by EPSRC and Intel Corporation. ND was funded by Cerner Corporation.

References

[1] Knaus W, Zimmerman J, Wagner D, Draper E. APACHE II: a severity of disease classification system. Critical Care Medicine 1985;13:818–829.

[2] Le Gall JR Lemeshow S SF. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270:2957–2963.

[3] Lemeshow S, Teres D, Klar J, et al. Mortality probability model (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993;270:2478–2486.

[4] Zimmerman JE KA. Outcome prediction in critical care: the acute physiology and chronic health evaluation models. Current Opinion in Critical Care 2008;14:491–497.

[5] Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, Kramer AA. Assessing contemporary intensive care unit outcome: an updated mortality probability admission model (MPM0-III). Critical Care Medicine 2007;35:827–835.

[6] Moreno R, Metnitz P, Almeida E, others, on behalf of the SAPS 3 Investigators. SAPS 3: from evaluation of the patient to evaluation of the ICU. part 2. development of a prognostic model for hospital mortality at ICU admission. Intensive Care Medicine 2005;31:1345–1355.

[7] Kramer A ZJ. Assessing the calibration of mortality benchmarks in critical care: The hosmer-lemeshow test revisited. Critical Care Medicine 2007;35(9):2052–2056.

[8] Saeed M, Villarroel M, Reisner A, Clifford G, Lehman L, Moody G, Heldt T, Kyaw T, Moody B, RG M. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): A public-access intensive care unit database. Critical Care Medicine 2011;39(5):952–960.

[9] Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics Theory and Methods 1980;9(10):1043–1069.

Address for correspondence: alistair.johnson@eng.ox.ac.uk