Data preprocessing and mortality prediction: the Physionet/CinC 2012 challenge revisited

Alistair EW Johnson[1], Andrew A Kramer[2], Gari D Clifford[1,3]

[1]University of Oxford, Oxford, UK
[2]Cerner Corporation, Vienna, VA, USA
[3]Emory University, Atlanta, GA, USA

Abstract

The Physionet/CinC 2012 challenge focused on improving patient specific mortality predictions in the intensive care unit. While most of the focus in the challenge was on applying sophisticated machine learning algorithms, little attention was paid to the preprocessing performed on the data a priori. We compare four standard pre-processing methods with a novel Box-Cox outlier rejection technique and analyze their effect on machine learning classifiers for predicting the mortality of ICU patients. The best machine learning model utilized the proposed preprocessing method and achieved an AUROC of 0.848. In general, the AUROC of models using our novel preprocessing method increased, and this increase was as much as 0.02 in some cases. Furthermore, the use of preprocessing improved the performance of regression models to a higher level than that of non-linear techniques such as random forests. We demonstrate that proper preprocessing of the data prior to use in a prognostic model can significantly improve performance. This improvement can be even greater than that provided by more complex non-linear machine learning algorithms.

1. Introduction

With the burgeoning supply of medical data, prognostic systems are becoming increasingly complex and more accurate. This is especially so in the intensive care unit (ICU), which hosts both the most severely ill and the most heavily monitored patients. Prominent examples of ICU prognostic systems include the APACHE IV [1], SAPS III [2, 3], MPM III [4], and OASIS [5] models, which predict patient mortality given their physiology and demographics. These models are primarily used for risk-adjustment as they do not have sufficient calibration for patient specific predictions. However, there is increasing opportunity to leverage large datasets in order to provide patient specific predictions. The aim of the Physionet/CinC 2012 challenge was to spurn innovation in this field [6].

Much of the competition entries focused on training machine learning algorithms which were much more sophisticated than prior work. However, an often overlooked aspect are the artefacts present in the data. Due to the chaotic nature of the ICU, data sourced from monitors is often corrupted by noise. This noise can reduce the performance of a predictive model as the underlying machine learning methods will have optimized parameters based upon incorrect information. For these reasons data preprocessing to remove artefacts is of the utmost importance, yet it has received relatively little exposition in previous publications. Furthermore, there is no systematic analysis of performance of machine learning methods under various levels of data preprocessing. Here we compare three data preprocessing methods across four widely employed machine learning techniques. The data preprocessing methods include no data preprocessing, utilizing domain knowledge, and a novel Box-Cox based iterative outlier rejection method. The preprocessing methods are evaluated when used in conjunction with a Random Forest (RF), Support Vector Machine (SVM), Regularized Logistic Regression (RLR), and a Regularized Logistic Regression with additional square terms (RLR^2).

2. Methods

The data utilized was prepared by the challenge organizers and was originally sourced from the MIMIC II database, a freely available public access dataset hosted on Physionet containing data for ICU admissions to the Beth Israel tertiary care hospital [6, 7]. Data for 4,000 patients who stayed at least 48 hours in the ICU were extracted from the dataset. The outcome of interest was in-hospital mortality. Variables included patient age, gender, height, weight, ICU type, and 37 time-stamped physiological measurements (e.g. heart rate, systolic blood pressure). Further to this, data for an additional 4,000 patients were extracted and provided online, but the outcomes for this subset were kept hidden. This allowed for unbiased

evaluation of model performance by the challenge organizers. The data was first converted from timestamped measurements into features usable in a supervised classification setting. The overall development process involved: 1) (optionally) preprocessing the data by aggregating all time-stamped measurements across all patients, 2) extracting features for use in the machine learning method, 3) (optionally) further preprocessing the data (using the same techniques as before), and finally 4) training and validating the models.

2.1. Preprocessing

Three pre-processing methods were evaluated: no preprocessing, domain knowledge, and the proposed iterative Box-Cox outlier rejection technique (*BCOR*). Note that all these preprocessing methods focused on removing outliers. When a preprocessing method detected an outlier, its value was set to missing, and would later be replaced by mean imputed values.

No pre-processing involved using the features directly extracted from the data, but it is worth noting that these features were still univariately standardized to have zero mean and unit variance (with normalization coefficients calculated from only the training set). This is to allow for use of scale sensitive methods, such as an SVM.

Domain knowledge pre-processing involved first correcting human transcription errors (such as recording temperature in degrees Farenheit rather than Celsius), then removing values which were unphysiological (by applying upper and lower bounds). For features where limits were not obvious (e.g. heavy tailed distributions like urine output), no thresholding was applied.

BCOR proceeds iteratively and univariately. Each feature is Box-Cox transformed to increase its similarity to a normal distribution. Thresholds are then determined using a critical value at the 0.01 significance level ($\alpha = 0.01$) with application of the Bonferroni correction. Specifically, given a data vector $\mathbf{x}$, we determine from the data a value λ which maximizes the profile log-likelihood [8] of transformed data $\mathbf{x}' = \frac{\mathbf{x}^\lambda - \mathbf{1}}{\lambda}$ being sourced from a normal distribution. For $\lambda = 0$ the transformation takes the form $\mathbf{x}' = \mathbf{log}(\mathbf{x})$. Thresholds are generated from the transformed dataset $\mathbf{x}'$ as:

$$\pm \frac{(1 - \frac{\alpha}{2})}{N} \Phi^{-1}\left(\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right), \quad \Phi \sim \mathcal{N}(\mu(\mathbf{x}'), \mathbf{s}(\mathbf{x}')) \quad (1)$$

... where $\mu(\cdot)$ calculates the mean, $s(\cdot)$ calculates the standard deviation, $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function for a normal distribution, and N is the number of data points observed. Data not residing within the thresholds were replaced with a missing value indicator. This removes values which, given the number of samples, are extremely unlikely to have been drawn from the overall distribution. This process is repeated until no values are removed, and all transformation parameters and associated thresholds are saved for later application to the validation sets. Data after the BCOR process remains in the transformed space, and thus is more Gaussian than the original data. Binary and ordinal variables are not preprocessed in this manner.

2.2. Feature extraction

For the temporally evolving measurements, in general, the standard deviation, first, last, highest, lowest, and median of all the measured values were used as features. These measures are gross aggregates of each patient's vital sign trajectories, and have been shown to have surprising performance despite their simplicity [**?**]. Traditionally, only the highest and lowest values are used in severity scoring systems [1–4]. Variables which were not processed in this way include urine output and mechanical ventilation. For urine output, the cumulative sum over all values was used. For mechanical ventilation three flags were created: the first indicated whether the patient was ventilated within the first 4 hours of ICU admission, the second indicated whether the patient was ventilated between hours 44 and 48, and the final flag indicated whether the patient was ever mechanically ventilated. For static measurements (such as gender), the admission value was used. The 'ICU-type' variable was converted into four indicator variables for each ICU type. This resulted in a a design matrix with 4,000 observations and 198 features.

2.3. Futher processing

The design matrix was then handled in one of four ways. The first involved no additional processing. The second involved applying the domain knowledge step. The third involved applying the *BCOR* step to the design matrix. This allows for quantification of preprocessing both before and after synthesis of the data into summary features. Finally, the above three were repeated when missing value indicator features were added to the design matrix. Missing value indicators were created only if a feature had missing values, and contained a value of '1' if a feature value was missing (and 0 otherwise).

2.4. Model development

Predictive models were then developed using 4-fold cross-validation, and performance measures are calculated across the held-out folds. Relevant hyperparameters were learnt using a further internal cross-validation. All data was standardized to the training set to prevent scaling issues affecting model performance. Furthermore, missing

data was imputed using the mean value of its respective feature in the training set. The models used were: Regularized Logistic Regression (RLR), Regularized Logistic Regression with the addition of each covariate squared to the design matrix (RLR^2), Support Vector Machines (SVM), and Random Forests (RF).

The methods were evaluated using two metrics of performance. The first, the area under the receiver operator characteristic curve (AUROC), represents the probability of correctly ranking a positive outcome higher than a negative outcome, or mathematically $p(\hat{y}|y = 1 > \hat{y}|y = 0)$.

As the AUROC is invariant to a lack of model calibration the negative log-likelihood is also presented. If the model developed is treated as a function $f(\mathbf{x})$ which outputs a probability then the negative log likelihood is calculated as $-log(p(y|f(x)))$. Assuming a binomial likelihood function for the target of interest, y, the negative log likelihood can be calculated as:

$$\log(\mathrm{p(y|f(\mathbf{x}))}) = -\log\left(\mathrm{f(\mathbf{x})}\right)\times \mathrm{y} - \log\left(1-\mathrm{f(\mathbf{x})}\right)\times(1-\mathrm{y})$$

The closer this value is to 0, the better the calibration and discrimination of the model.

3. Results

The performance as measured by the AUROC is shown in Figure 1. The performance as measured by the negative log likelihood is shown in Figure 2. The ideal value for the AUROC is 1, and higher values indicate better discrimination. The ideal value of the negative log likelihood is 0, and lower values are better.

4. Discussion

All models have an AUROC over 0.80, which is considered to provide excellent clinical utility in the field of outcome prediction [4]. The proposed pre-processing method, BCOR, is an effective means of further improving this efficacy. For most models the BCOR resulted in large performance increases as compared to traditional domain knowledge techniques. It is interesting to note the relative insensitivity of RF to pre-processing techniques. This is likely because a RF offers a substantial amount of flexibility. Values which are outliers can be assigned a distinct contribution, and thus RFs are capable of naturally handling outliers. However, this flexibility comes at a cost, and RFs usually require a large dataset to learn from. The SVM has been shown to be effective for smaller datasets, as is apparent in this work. As SVMs are sensitive to scale, the removal of values which are more extreme than would reasonably occur in the class distributions improves the performance. Surprisingly, the addition of missing value flags

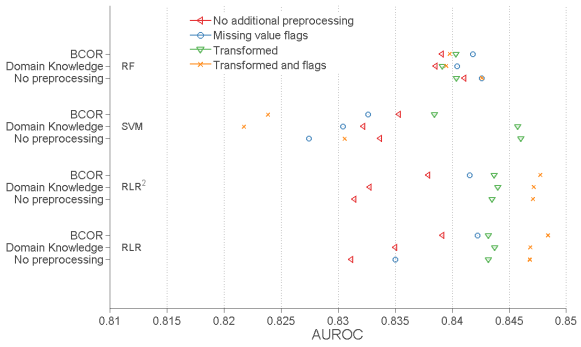

Figure 1. Discrimination of each method as measured by the area under the receiver operator characteristic curve (AUROC). The preprocessing methods applied prior to feature extraction are shown on the y-axis. The symbol and color combinations represent: no pre-processing (red diamond), addition of missing value flags (blue circle), BCOR which includes feature transformation (green triangle), and BCOR combined with missing value indicators (orange cross). These methods are compared when classifying ICU mortality with different machine learning techniques as specified on the figure.

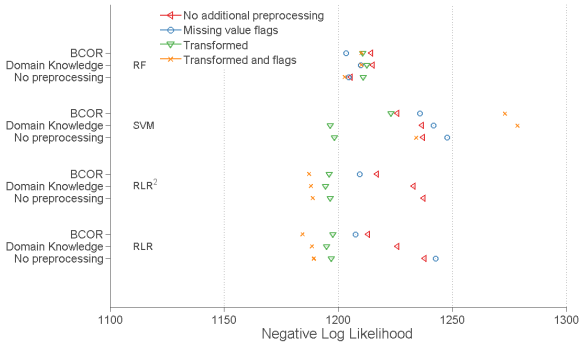

Figure 2. Calibration and discrimination of each method as measured by the negative log likelihood. Lower values are better. The preprocessing methods applied prior to feature extraction are shown on the y-axis. The symbol and color combinations represent: no pre-processing (red diamond), addition of missing value flags (blue circle), BCOR which includes feature transformation (green triangle), and BCOR combined with missing value indicators (orange cross). These methods are compared when classifying ICU mortality with different machine learning techniques as specified on the figure.

strongly hinders the performance of the SVM models. This is likely due to the high feature to data ratio, which is further increased by the addition of missing value flags (up to

385 features for 3000 observations in each fold).

The best performing combination of models and preprocessing methods was an RLR which utilized the BCOR method prior to feature extraction, the BCOR method after feature extraction (which contains a feature transformation step causing feature distributions to be more Gaussian), and the addition of missing value flags. This is not entirely unsurprising as the RLR method strongly benefits from L1-regularization, which prevents model overfitting and reduces the impact of high feature to observation ratios. The AUROC of the RLR model, one of the most common predictive models used in the medical literature, was increased from 0.83 to 0.85. This demonstrates the significant part that preprocessing data plays in model performance, even though it is an often neglected topic.

Interestingly, the RLR^2 model did not improve upon the RLR model performance appreciably, or at all when used in conjunction with BCOR. This indicates that the addition of a second degree of freedom for each feature was not useful. This may be explained by the presence of a high degree of colinearity caused by features extracted from the same variable. The addition of a term for median heart rate squared, for example, is not useful when the model can already utilize features representing the first and last heart rate as additional degrees of freedom.

Domain knowledge pre-processing was also effective at improving model performance. The removal of unphysiological values is an effective method of controlling for transcription errors and other artefacts. However, domain knowledge is not trivial to implement. First, it requires agreement of a range of allowable values, which is often arbitrarily selected. For example, a common upper threshold on heart rate is 300, but one can imagine a heart rate of 299 is also unreasonable. In contrast, the BCOR method chooses thresholds in a data-driven way, and also selects thresholds in a tractable fashion. Additionally, it is not always possible to select meaningful thresholds by hand, especially for long-tailed distributions such as white blood cell count.

5. Conclusion

Three methods of preprocessing medical data were compared using four machine learning methods. The methods were shown to consistently improve performance across all models. Using indicator variables for missing values, and thus allowing algorithms to learn the weighting for these missing values, generally improves performance except in the case of SVMs. The proposed preprocessing method BCOR is completely automatic, reducing the burden of quality assurance which is prevalent in healthcare. Furthermore, the method substantially improves the performance of the most common regression model. While data preprocessing is rarely studied in the for mortality prediction, these approaches can provide dramatic improvements.

Acknowledgements

AEWJ acknowledges the support of the RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation).

References

[1] Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Critical Care Medicine May 2006;34(5):1297–1310. ISSN 0090-3493.

[2] Metnitz PGH, Moreno RP, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR. SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. Intensive Care Medicine October 2005; 31(10):1336–44. ISSN 0342-4642.

[3] Moreno RP, Metnitz PGH, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR. SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Medicine October 2005;31(10):1345–55. ISSN 0342-4642.

[4] 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 March 2007;35(3):827–35. ISSN 0090-3493.

[5] Johnson AE, Kramer AA, Clifford GD. A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. Critical Care Medicine 2013;41(7):1711–1718.

[6] Silva I, Moody G, Scott DJ, Celi LA, Mark RG. Predicting in-hospital mortality of ICU patients: The physioNet/computing in cardiology challenge 2012. Computing in Cardiology 2012;39:245–248.

[7] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[8] Box GE, Cox DR. An analysis of transformations. Journal of the Royal Statistical Society Series B 1964;26(2):211–252.

Address for correspondence:

Alistair E. W. Johnson
Institute of Biomedical Engineering, Old Road Campus Research Building
University of Oxford
alistair.johnson at eng.ok.ac.uk