Data-Driven Mortality Prediction for Trauma Patients

Yuanyang Zhang
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106
yuanyang@cs.ucsb.edu

Bernie J Daigle Jr
Institute for Collaborative Biotechnologies
University of California, Santa Barbara
Santa Barbara, CA 93106
bdaigle@gmail.com

Lisa Ferrigno
Santa Barbara Cottage Health System
Santa Barbara, CA 93106
lferrigno@sbch.org

Mitchell Cohen
Department of Surgery
University of California, San Francisco
San Francisco, CA 94143
MCohen@sfghsurg.ucsf.edu

Linda Petzold
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106
petzold@cs.ucsb.edu

Abstract

Trauma is the leading cause of death between the ages of 1 to 44. A large number of these deaths occur within days of the arrival of the patient at the hospital. Accurate prediction of the outcomes of trauma patients and the identification of a few key predictors would be highly valuable. In this paper we focus on (1) the prediction of mortality within any given time frame after arrival, and (2) the selection of key predictors. We consider that patients have both static and temporal data, and that a large number of missing values is inevitable in this type of dataset.

We propose a novel mortality prediction model, which extends the logistic regression with elastic net to accommodate large numbers of missing values as well as highly-correlated time-course data. Specifically, we formulate the prediction of mortality as a logistic regression with elastic net regularization by employing static data and all temporal measurements up to the current time. We include an impact function in the weights of each group of time measurements to reduce the correlation across different time points. The impact function is a function of time, which can be adjusted to emphasize earlier or later measurements. We also include a scaled missing value indicator function, which allows us to accommodate variable numbers of missing values. Our method preserves the sparsity property of the elastic net while extending its applicability to time course data with numerous missing values. We compare our method to unmodified logistic regression with elastic net run on either the first or the last time point measurements in the dataset.

1 Introduction

Trauma is the leading cause of death between the ages of 1 to 44. More than 180,000 deaths from injury occur each year in the United States [1]. For treatment of trauma patients, time is of the

essence. An accurate prediction of patient outcome, with a need to measure few key predictors, would be highly valuable.

Our dataset is from the UCSF/San Francisco General Hospital and Trauma Center. It consists of both static data including demographic information, background information and injury information, and temporal data, including basic measurements, blood factor concentration, and other clinical features. Due to the intense environment in the treatment of trauma patients, missing values are common in this type of dataset.

In this paper we propose a temporal importance-adjustment model, a mortality prediction algorithm which extends the applicability of logistic regression with elastic net to time course data with large number of missing values. Specifically, we formulate the prediction of mortality as a logistic regression with elastic net regularization, using both patients' static and time course measurements to make predictions. In order to reduce the correlation of the same feature across different time points, we introduce an impact factor which is a function of time. For each patient, we scale the impact factors of each variable among available time points. With our formulation, we can easily accommodate variable numbers of missing values from time course variables. Furthermore, using an elastic net regularizer allows us to decrease the number of features needed for predicting mortality.

2 Approach

Suppose we have data from n patients $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$. For each patient, we have both static data $\mathbf{x}_{i,s}$ and temporal data $\mathbf{x}_{i,t_1}, \mathbf{x}_{i,t_2}, \cdots, \mathbf{x}_{i,t_N}$ which is measured at N time points. We seek to predict the mortality $y_1, y_2, \cdots, y_n$ of each patient at some pre-specified final time, where $y_i = 1$ if a patient dies, and $y_i = -1$ otherwise.

We model the conditional probability that patient i dies by

$$p(y_i = 1|\mathbf{w}, \mathbf{x}_i) = \phi(\mathbf{x}_i\mathbf{w}), \tag{1}$$

where $\mathbf{w}$ is the weight for features. Using a logistic link function, we obtain a logistic regression model [3]:

$$\phi(\cdot) = \frac{\exp(\cdot)}{1 + \exp(\cdot)}. \tag{2}$$

For all patients, the negative log likelihood function is given by

$$l(\mathbf{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{x}_i\mathbf{w}))). \tag{3}$$

Using an elastic net [5] to regularize the weights, we obtain

$$l(\mathbf{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{x}_i\mathbf{w}))) + \lambda_1||\mathbf{w}|| + \lambda_2||\mathbf{w}||_F^2, \tag{4}$$

where $||\cdot||$ and $||\cdot||_F$ are the L^1 and L^2 norm respectively, and λ_1 and λ_2 are hyper-parameters that penalize the L^1 and L^2 norms respectively.

Considering that for each patient, $\mathbf{x}_i$ contains both static data and temporal measurements, we propose to replace model (1) with

$$p(y_i = 1|\mathbf{w}, \mathbf{x}_i) = \phi(\mathbf{x}_{i,s}\mathbf{w}_s + g(t_1, \delta)\mathbf{x}_{i,t_1}\mathbf{w}_{t_1} + g(t_2, \delta)\mathbf{x}_{i,t_2}\mathbf{w}_{t_2} + \cdots + g(t_N, \delta)\mathbf{x}_{i,t_N}\mathbf{w}_{t_N}), \tag{5}$$

where $g(t_i, \delta)$ is an impact function in terms of time and δ is the impact parameter. The impact function allows us to adjust the importance of measurements at different times.

The conditional probability (5) can be expressed as

$$p(y_i = 1|\mathbf{w}, \mathbf{x}_i) = \phi((\mathbf{G}_i \odot \mathbf{x}_i)\mathbf{w}), \tag{6}$$

where $\mathbf{G}_{i,j}$ is the impact factor for patient i at feature j, and $\odot$ is the element-wise product.

Missing values are very common in clinical measurements, especially in a trauma situation. We use $\mathbf{S}$ to indicate missing values, where $\mathbf{S}_{i,j} = 0$ if the target value of patient i in feature j is missing, and $\mathbf{S}_{i,j} = 1$ otherwise. The impact factor for patient i becomes $\mathbf{S}_i \odot \mathbf{G}_i$. Normalizing the impact factor, we obtain

$$p(y_i = 1|\mathbf{w}, \mathbf{x}_i) = \phi\left(\left(\frac{\mathbf{S}_i \odot \mathbf{G}_i}{\|\mathbf{S}_i \odot \mathbf{G}_i\|} \odot \mathbf{x}_i\right)\mathbf{w}\right). \tag{7}$$

Letting $\hat{\mathbf{G}} = \frac{\mathbf{S}_i \odot \mathbf{G}_i}{\|\mathbf{S}_i \odot \mathbf{G}_i\|}$ we can write (7) simply as

$$p(y_i = 1|\mathbf{w}, \mathbf{x}_i) = \phi((\hat{\mathbf{G}} \odot \mathbf{x}_i)\mathbf{w}). \tag{8}$$

The modified negative log likelihood function is given by

$$l(\mathbf{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i((\hat{\mathbf{G}} \odot \mathbf{x}_i)\mathbf{w}))) + \lambda_1||\mathbf{w}|| + \lambda_2||\mathbf{w}||_F^2. \tag{9}$$

We use a coordinate descent method with one-dimensional Newton direction [4] to minimize the negative log likelihood (9).

3 Experiments

3.1 Dataset and Preprocessing

Our dataset, from the UCSF/San Francisco General Hospital and Trauma Center, includes comprehensive demographic, injury, physiologic and biomarker data on consecutive patients admitted to an urban Level I Trauma Center who required ICU admission. We have a total of 859 ICU patients and 1532 measurements in the dataset. 124 of the features are static data, including demographic information, background information and injury information. 118 attributes have been measured at hours 0, 2, 3, 4, 6, 12, 24, 48, 72, 96 and 120, including basic measurements, blood factor data and lab test data.

Patients are excluded if they have only one time point with measurements. Features are excluded if they have $\geq 30\%$ missing values across all included measurements. After preprocessing, we have 233 patients with 96 static features and 22 temporal features at each time point. Data from 67% of the time points is still missing. We randomly split the dataset into 80% training data and 20% testing data for evaluation.

3.2 Evaluation

A patient's mortality status at a given time after arrival at the hospital is determined based on known information about the patient at that time. In other words, if a patient dies before the given time, we treat the outcome as dead. If the patient has not yet died or has been discharged from the hospital, we treat the outcome as survived. We consider mortality predictions at 5 days, 30 days and 1 year.

We compared our proposed method to that of unmodified logistic regression with elastic net regularization (using the R package `glmnet` [2]). Because this method cannot handle large numbers of missing values in the time course data, we apply it to: (1) static data and initial time point measurements, (2) static data and final time point measurements.

Since the outcomes are largely unbalanced in all 3 prediction models (5.5%, 11.2% and 12.4% deaths respectively), we randomly sub-sample the survived patients in the training dataset to keep a $30\%/70\%$ ratio between dying and surviving patients. We select all the hyper-parameters in the three prediction models using 5-fold cross-validation in the training dataset with AUC (area under the ROC curve) as an objective. The learned hyper-parameters are then used on the training dataset, and predictions are made on the testing dataset. We point out that the impact function itself is also a hyper-parameter, which we either select from linear, square root or exponential function. In our experiments, we find that a linear impact function has the highest AUC in cross-validation. The results for all methods and models are compared in Table 1.

Table 1: AUC of temporal importance-adjust model compared with two other models: logistic regression with elastic net run on initial time measurements only, and the same method run on final time measurements only. We predict mortality at three final time points: 5 days, 1 month and 1 year. We also list the true positive rate and true negative rate, where the total number of samples is 47.

METHOD	TIME	AUC	True Positive Rate	True Negative Rate
Temporal Importance-Adjustment		**0.67**	**0.66**	**0.89**
Elastic net + Initial	5 days	0.59	0	1
Elastic net + Final		0.65	0	1
Temporal Importance-Adjustment		**0.79**	**0.5**	**0.73**
Elastic net + Initial	1 month	0.67	0	1
Elastic net + Final		0.78	0	1
Temporal Importance-Adjustment		**0.81**	**0.71**	**0.7**
Elastic net + Initial	1 year	0.66	0	1
Elastic net + Final		0.66	0	1

Table 2: Key terms identified with highest weights for 5 days, 1 month and 1 year mortality.

Key Features	5 days	1 month	1 year
Static	Abbreviated injury scale-head, Congestive heart failure, Valproic, Keppra, Pedestrian-vehicle accident.	Abbreviated injury scale-head, Congestive heart failure, Keppra, Chronic obstructive pulmonary disease, Valproic .	Abbreviated injury scale-head, Congestive heart failure, Glasgow coma scale, Injury severity score, Keppra.
Hour 0- 12	Heart rate, Temperature, Respiratory rate, Partial pressure of oxygen, ATIII .	pH, Temperature, Factor VII, Prothrombin time, Protein C .	Temperature, Heart rate, pH value, Factor VIII, Bicarbonate.
Hour 24- 48	Factor VII, Heart rate, Prothrombin time, Partial pressure of oxygen.	Respiratory rate, Factor IX, Base deficit/excess, Factor VII, D-dimmer.	Respiratory rate, Factor VII, Prothrombin time, Heart rate, Factor VIII.
Hour 72- 120	Factor IX, Factor V, Bicarbonate, Respiratory rate, Base deficit/excess.	Factor IX, Heart rate, Respiratory rate, Factor II, Factor V .	Factor IX, Bicarbonate, Blood pressure, Protein C, Factor VIII.

We can see that our method always obtains the highest AUC. Also, we note that even though logistic regression with elastic net on one time point measurement can achieve similar AUCs in the first two final time cases, this method does so by predicting that all patients survive, which leads to a zero true positive rate.

Lists of the 5 features obtained by the temporal importance-adjustment model are shown in Table 2.

3.3 Discussion

In this work, we propose a temporal importance-adjustment model for predicting mortality, which incorporates both time-dependent and missing value indicator functions in a logistic regression model with elastic net regularization. By doing so, we can accommodate time-course data with large numbers of missing values, as well as enjoy the simplicity provided by logistic regression and the sparsity provided by regularization.

Acknowledgements

We gratefully acknowledge financial support from the U.S. Army Research Office (Coagulopathy grant W911NF-10-2-0114).

References

[1] N. C. for Injury Prevention and C. u. W. . Control. 10 leading causes of death by age group, united states @ONLINE, 2010.

[2] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[3] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[4] G.-X. Yuan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *The Journal of Machine Learning Research*, 11:3183–3234, 2010.

[5] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.