Feature Representation for ICU Mortality

Harini Suresh

December 17, 2015

1 Problem Statement

Prognostic models to predict the outcome of patients in Intensive Care Units (ICU) are valuable for risk stratification, clinical decision-making, and evaluating ICU performance (1). When building these models, machine learning practitioners typically choose numerical feature representations based on the type and scale of the features in the data. Namely, the most prevalent representations are binary (yes/no) (7) or linear (counts of feature occurrence) (6). However, these representations are not usually adequate. Binary representations may discard information about the severity of a patient’s condition (i.e. a patient receiving multiple doses of drugs promoting urination may be having more trouble with water retention in heart failure) while linear representations may grant this information too much importance (i.e. a patient receiving three types of blood pressure lowering medications may not be at 3 times the risk of cardiovascular events relative to a patient receiving one).

Despite the weakness in these representations, these choices are rarely evaluated quantitatively or compared with other possibilities. Other possibilities include a new approach that will be introduced as the “hill” representation, which is centered around a feature’s median value. Different representations are also likely better suited to different features.

This study aims to both quantitatively evaluate the strengths and weaknesses of many different feature representations in this prediction task, with the long-term goal of introducing a method to choose optimal representations for each feature, from among a wide library of representations. If from a given set of selected features the best representations could be successfully determined, over/under-representation problems could be minimized to give the most accurate model of ICU mortality. This in turn has the potential to better treatment of ICU patients as well improve accuracy of population-level risk models.

2 Related Work

Many studies have attempted to build models using physiological values or counts (i.e. linear representation) or categorical variables (i.e. binary representation), and have explored using different subsets of features. Joon et al used both binary and linear representations to create patient similarity metrics and build a more personalizable model for a particular patient by using features extracted only from other similar patients (2). Pirracchio et al developed a “super learner” algorithm to predict ICU mortality that uses ensemble machine
learning to obtain better performance (3). They used features as represented in SAPS (Simplified Acute Physiology Score) II, which include values such as heart rate, systolic BP, and sodium and potassium levels, as well as the raw values of the features. Silva et al used general descriptors (height, weight, gender, etc) as well as outcome-related descriptors (SAPS-I score, SOFA score, length of stay, etc) to model patient mortality (4).

While these studies delve deep into model selection, the models being compared differ in feature sets or model hyperparameters rather than feature representations. In existing studies there is a lack of quantitative comparisons of the feature representations themselves, and I see an opportunity to more critically evaluate and develop those representations.

3 Methods

3.1 Features

I will use features extracted from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) Database (5). MIMIC II contains data collected between 2001 and 2008 from a variety of Intensive Care Units (ICUs) in a single tertiary teaching hospital. The database contain general patient information (ICD-9 codes, demographics, room tracking), physiological signals (vital metrics, SAPS), medications (IV meds, provider order entry data), lab tests (chemistry, imaging), fluid balance (intake, output), and notes (discharge summary, nursing progress reports). The MIMIC dataset is notable because it is publicly available for free use, encompasses a large and diverse set of patients, and contains numerous high resolution features for each patient.

The patients used this study were from the Medical Care Unit (MICU), Cardiac Care Unit (CCU), and Cardiac Surgery Recovery Unit (CSRU). The features extracted were from the following MIMIC tables: medevents, meddurations, additives, ioevents, iodurations, chartevents, chartdurations, procedureevents, and microbiologyevents. All clinical event data was represented as a count (so in the case of a procedure, the feature value is how many times that procedure was administered, rather than any values associated with it). This data comprised approximately 17,000 patients and 6,000 features (Figure 1 details the feature values for the extracted features).

3.2 Representations

Let the raw feature value for a patient $i$ and feature $j$ be $z_{ij}$, and its transformed representation in the feature matrix be $x_{ij}$ (see Figure 2 for graphical versions of each representation).

The first and simplest representation is the Binary representation, or the indicator function for a non-zero value of $z_{ij}$:

$$\text{binary}(x_{ij}) = I(z_{ij} > 0)$$  \hspace{1cm} (1)

An alternative representation that preserves the severity information contained in different non-zero raw feature values is a linear map of $z_{ij}$ to the range $[0,1]$:
The linear representation can also be truncated or untruncated, a distinction that comes about if the testing set has a different range than the training set. If the testing set has a value that is past the maximum of the training set, the testing set representation can either be cut off at 1, or scaled up proportionately (see Figure 3).

Finally, a novel representation similar to the Hill equation in biochemistry is as follows:

\[
\text{hill}\{x_{ij}\} = \frac{z_{ij}}{z_{ij} + m_j}
\]  

(3)

where \(m_j\) is the median value of non-zero \(z_{ij}\) across all patients feature \(j\). In this representation, values equal to the median are mapped to 0.5. Using the median allows for less sensitivity to outliers, and allows the representation to be independent of the feature’s minimum or maximum. It also provides greater resolution around the range of most common values.

3.3 Model

I used a logistic regression model with L2 regularization using sklearn’s linear_model.LogisticRegression method. The training and testing sets were split using a stratified 80/20 ratio. 3-fold stratified cross-validation was used on the training set to determine the optimum cost parameter for each model.

4 Results

After training each model on the training set, the AUCs on the testing set were measured (see Figure 4). The Hill representation performed the best, followed
Figure 2: A schematic of the different feature representations for the values 1 through 10 for a feature with a minimum of 0, maximum of 10, and median of 3.

Figure 3: A schematic of a truncated versus non truncated linear representation, using a dataset where the training maximum is 5, but values larger than 5 appear in the testing set.
Figure 4: The AUCs of each model on the testing set, for all care units together and each care unit on its own (the model was trained separately for each care unit, using training data from just that care unit).

by Binary, Truncated Linear, Raw Values, and Untruncated Linear. The top features for the model trained using all care unit are as follows:

| Feature | Description |
|---------|-------------|
| SAPS I Score | Endoscopic Retrograde Cholangiopancreatogram (Procedure) |
| | Aprotinin |
| | Incision Of Perirenal Or Periureteral Tissue (Procedure) |
| | Thoracic Operation (Procedure) |
| | Parital Hepatectomy (Procedure) |
| | Cava-pulmonary artery anastomosis (Procedure) |
| | zzO2AVI |
| | Bile Duct Repair (Procedure) |
| | Other percutaneous procedures on biliary tract (Procedure) |
| | Pericardial flush |
| | bupivacaine |
| | Hemispherectomy (Procedure) |

Bootstrap sampling was used to resample the test set 100 times, and the AUC for each model on all 100 test sets was measured (see Figure 5 for the AUC distributions for the models for all care units together). For each of the 100 resamplings, the order listed above was preserved.
Figure 5: caption
5 Discussion

With these preliminary experiments, the Hill representation performed better than linear, binary, or raw values on each of the datasets. This may be because the linear representation uses the maximum function, which is sensitive to outliers. Specifically, when some patients in the population have a high count of a feature, the bulk of the population feature values get compressed by the comparatively large value in the denominator of Equation 2. By contrast, the Hill representation spreads the density more evenly.

Another result to note is the distinction in AUC between the untruncated linear representation and raw values. Rather than the untruncated linear weights simply being a scaled version of the raw values, when learning is performed the regularization term pulls the weights to zero (shrinkage) with the loss function based on the data. This causes the learned weights to generally be not equal to the scaled weights.

Future work should include scaling up the data used to include data from MIMIC III (around 48,000 patients). It may also be interesting to include new types of representations, including a "discrete" representation where features are assigned a value based on certain quantiles of the feature distribution, or a linear representation that is scaled down by the values at certain quantiles rather than the actual minimum and maximum. While L2 regularization was used in the results presented in this study, it may also be beneficial to try different regularization methods such as elastic net, a balance between L1 and L2 regularization.

6 References

1. Power, G. Sarah, and David A. Harrison. Intensive Care National Audit & Research Centre (ICNARC), London, United Kingdom. “Why try to predict ICU outcomes?” Current Opinion in Critical Care. 20(5):544-549, October 2014.

2. Lee, Joon, David M. Maslove, and Joel A. Dubin. “Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric.” Ed. Frank Emmert-Streib. PLoS ONE 10.5 (2015): e0127428. PMC. 11 Oct. 2015.

3. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. “Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study.” Lancet Respir Med. 3(1):42-52. January 2015.

4. Silva, Ivanovitch, et al. ”Predicting in-hospital mortality of ICU patients: The physioNet/computing in cardiology challenge 2012.” Computing in Cardiology (CinC), 2012. IEEE, 2012.

5. Saeed, Mohammed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. “Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database.” 2011; 39:952-960. DOI: 10.1097/CCM.0b013e31820a92c6.
6. Neuvirth, H. M. Ozery-Flato, J. Hu, J. Laserson, M. S. Kohn, S. Ebadolali, and M. Rosen-Zvi. “Toward personalized care management of patients at risk: the diabetes case study.” *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, 395–403. ACM, 2011.

7. R. G. Krishnan, N. Razavian, Y. Choi, S. Nigam, S. Blecker, A. M. Schmidt, and D. Sontag. “Early detection of diabetes from health claims.” *Neural Information Processing Systems Workshop: Machine Learning for Clinical Data Analysis and Healthcare*, 2013.

8. Ghosh, D., M. Chinnaiyan. “Classification and selection of biomarkers in genomic data using LASSO.” *J. Biomed. Biotechnol*, 147-154. 2005.

9. Ng, A. “Feature selection, L1 vs. L2 regularization, and rotational invariance.” *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.