Combining Machine Learning and Clinical Rules to Build an Algorithm for Predicting ICU Mortality Risk

Michael Krajnak, Joel Xue, Willi Kaiser, William Balloni

GE Healthcare Systems, Wauwatosa, WI, USA

Abstract

In this study we aim to develop a decision support application for predicting ICU mortality risk that starts with a clinical analysis of the problem that also leverages machine learning to help create an algorithm with good performance characteristics. By starting from a clear basis in clinical practice we hope to improve algorithm development and the transparency of the resulting system.

We start with a general model structure for a fuzzy rule based system (FIS). The model can be specified by clinicians who identify the inputs and the rules. An optimizer based on a genetic algorithm generates the coefficients for the final solution. Using the 2012 PhysioNet/CinC Challenge data set we constructed a Phase 1 system using minimal clinical guidance. Our initial FIS's achieved scores of 0.39 for Event 1 and 94 for Event 2. In Phase 2 we updated the FIS based on clinician interviews. At the end of Phase 2 we achieved 0.40 for Event 1 and 60 for Event 2.

We hope to show that machine learning techniques that are modeled on the clinical understanding of a problem can be competitive with more abstract machine learning approaches but may be preferable because of their explainability and transparency.

1. Introduction

A fuzzy rule based system (or fuzzy inference system - FIS) can represent complex non-linear models as clinical rules. In contrast purely data driven techniques like neural networks or support vector machines can generate solutions that perform well, but are difficult to explain. Thus a key advantage of using a FIS is that it can be easily understood by clinicians. This allows clinicians to review the rules and provide feedback. Other regression approaches where the system behaves as a black box are harder to explain [4].

In previous work we had developed a FIS for identifying aesthetic overdose [1]. By applying the same approach to the 2012 PhysioNet Challenge [3] we hope to discover whether or not it is competitive with other decision support algorithms. If it performs reasonable well we believe the ability to reason about the the FIS with clinicians may make it preferable to other methods with similar performance.

2. Methods

2.1. Phase 1

We first created a Phase 1 FIS with the goals 1) Improve over the Challenge sample entry [3] given in the PhysioNet Challenge and 2) Compare our technique to a simple neural network (NN). If a simple neural network performed far better than a simple FIS we were ready to abandon our approach.

For our initial feature set we took the last value in each parameter and narrowed down the number of features using a NN based feature perturbation analysis. Our FIS took a naïve single feature per rule approach to constructing the rules.

Is our first pass comparison of the NN and FIS approaches we only measured the Event 1 score [3] and found that both approaches were comparable and performed slightly better than the Challenge example entry.

Our initial FIS entries for Event 1 and Event 2 had 15 features and 45 rules.

Table 1. Phase I Results.

Event	Sample Entry	NN	Phase 1
Event 1 Score	0.33	0.42	0.39
Event 2 Score	68	-	94

2.2. Clinical Analysis

We began Phase 2 with a literature review focusing on work done on the SAPS scoring system. Motivated by work done on the SAPS scoring system [2] we changed our initial feature definitions from the latest measurement of any parameter to the maximum or minimum of a parameter over an interval. We hoped this approach might also be more immune to certain kinds of noise in the data.

Then we began a series of clinician interviews. We recruited internally within our business and interviewed one respiratory therapist, two ICU nurses, and one emergency room doctor. They were given a list of the available parameters and asked to prioritize the utility of each parameter as both an absolute measure and as a trend. Then we asked a number of qualitative questions emphasizing time scales and relationships between the parameters. For example one such interaction that emerged during our discussions concerned the assessment of oxygenation based on inspired oxygen and O2 saturation. This lead to the rule, "If inspired oxygen is stable or increasing and O2 saturation is decreasing then the likelihood of mortality is high". The relationship between the two parameters provides more information than an assessment that only considers O2 saturation.

2.3. Data and feature extraction

We created a statistical survey of the Challenge data, including minimum, maximum, mean, and sigma of each parameter to establish preliminary upper and lower boundaries for allowable data. We cross checked it against the references provided by the challenge website to establish clinically significant upper and lower bounds. We then interviewed one clinician and established a final set of clinically significant bounds.

We noticed a number of other irregularities in the data including duplicate entries, other irregular entry timings, 0 valued entries, and noise. While we did clean up some minor irregularities, in the end due to time constraints we did not perform more aggressive artifact rejection. We assumed that derived features that were the minimum or maximum of a given time series would be somewhat immune to this kind of embedded noise that does is not associated with out of range values.

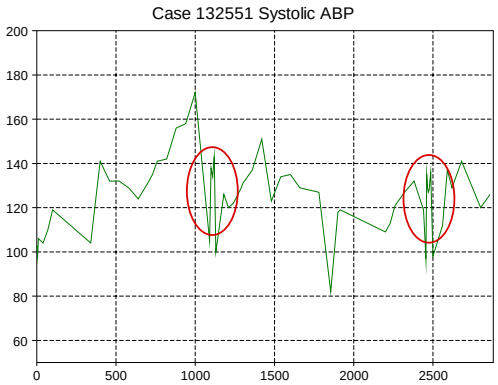

Figure 1. Systolic arterial blood pressure with artifacts.

Our clinicians also indicated a strong preference for trended values, in many cases over intervals as short as four hours. While it was simple to compute trends by taking the slope of a linear fit over some interval, we were concerned about the impact of noise and by the low data rates of some time series. In the end we only included the highest priority trended features as indicated by our clinicians, and only used eight hour intervals for time series with at least one data point per hour. We used twenty four hour intervals for the other trends.

2.4. FIS optimization

A FIS has rules and coefficients. The coefficients are used to convert features to fuzzy values that are processed by the rules. A fuzzy value is a category and a weight. One set of coefficients might map a heart rate of 90 to the fuzzy value "high" with a weight of 0.7. The optimizer is used to select the conversion coefficients that yield good results for the given rules.

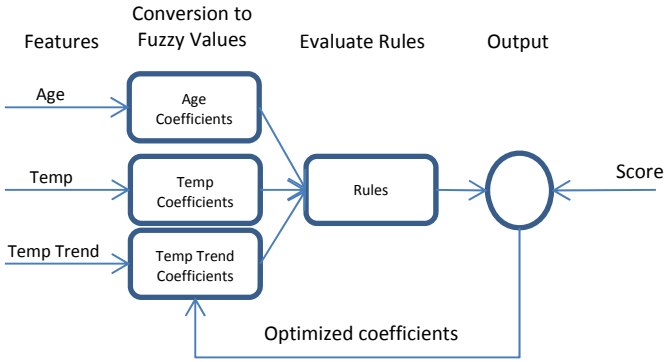

Figure 2. FIS optimizer structure.

Our FIS optimization system takes as settings: 1) A list of features, 2) For each feature the number of fuzzy categories to divide it into, 3) The clinically relevant maximum and minimum of each feature, 4) The rules, and 5) Event 1 and Event 2 scoring functions used to score an optimization result. We used the Event 1 scoring function to generate a FIS for Event 1 and the Event 2 scoring function to generate a separate FIS for Event 2.

The optimizer uses its settings to generate a fixed number of FIS's. Each FIS starts with random coefficients for converting parameters to fuzzy values and the same set of rules. Each FIS is scored and a genetic algorithm is used to create a new set of FIS's that have coefficients from set that was evaluated. The genetic algorithm uses each FIS's score to weight the probability that its coefficients will be reused so that the new set yields a better result[1]. Figure 2 shows the overall structure of the optimizer.

We then made several runs adding features to try to improve performance. With runs with more than 16 features it became very difficult to tell if adding features improved performance.

If we ran the optimizer for more generations we did see some small improvement in the score. But this resulted in over training the algorithm. Our Phase 1

algorithms had had similar performance for Challenge data set A and set B, but our first set of Phase 2 algorithms performed significantly worse on set B.

To analyse the over training problem we split the Challenge data set A into a training set with 70% of the cases and a validation set with 30% of the cases. We generated solutions using the reduced training set and over time plotted the improvement in the algorithm on the training set and on the validation set, see Figure 3.

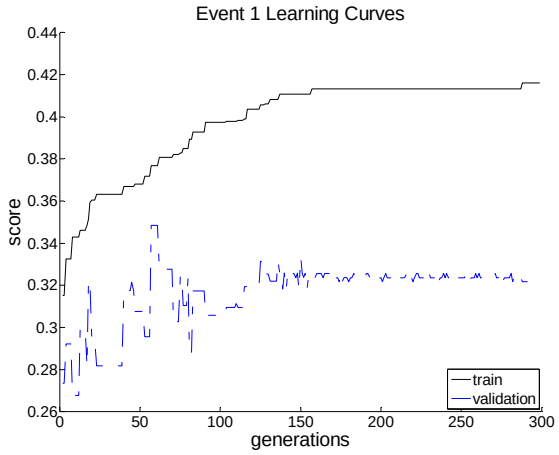

Figure 3. Learning curve for Event 1 scores showing a maximum for the validation score at generation 61.

Figure 3 shows the curves for Event 1 with increasing performance on the training and validation sets until generation 61 then performance increases for the training set but it falls on the validation set indicating over training. To address this we added additional features. Given that our previous evaluation showed that increasing performance with features was difficult, we restricted the new features to the parameters identified as most valuable by the clinicians. Our final FIS used 19 features.

Table 2. Final set of features used.

Feature
Age
Bilirubin. Max over 24 hours
BUN. Max over 24 hours
Creatinine. Max over 24 hours
Glasgow Coma Score. Min over 24 hours
HCO3. Max over 24 hours
Heart Rate. Max over 24 hours
PaO2. Min over 24 hours
PaO2. Trend over 24 hours.
pH. Min over 24 hours
Platelets. Min over 24 hours
Potassium. Min over 24 hours
Systolic ABP. Min over 24 hours
Systolic ABP. Trend over 8 hours
Temp. Max over 24 hours
Temp. Trend over 24 hours
Urine. Total over 24 hours
Urine. Trend over 8 hours
White Blood Cell count. Max over 24 hours

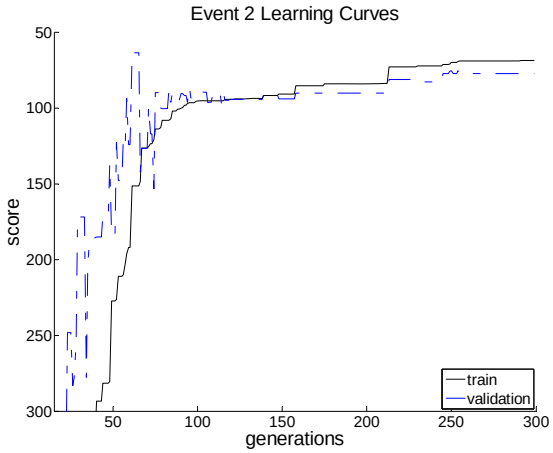

Figure 4. Learning curve for Event 2.

The learning curve for Event 2 shows the performance on both the training and validation sets improving and leveling off. To improve this we included the new features we added for our Event 1 FIS and we increased the size of the initial set of FIS's used by the genetic algorithm.

2.5. Error analysis

To look for rules that contributed to errors in the results we looked at a summary of the rule outputs over all of the cases in the validation set.

Each rule produced a fuzzy output for mortality risk as either "very low", "low", "medium", "high" or "very high" and an activation weight from 0 to 1. For each rule we captured its output for each case in the validation set. The outputs were binned by the Challenge set A outcome, by whether or not the patient survived.

Rules that output "high" or "very high" mortality risk more often when the risk was actually low were identified. Likewise for rules that indicated low when the risk was really high. These rules were deleted and the resulting rule set was resubmitted to the optimizer for further tuning.

Figure 5 shows the learning curves for the previous and the latest Event 1 FIS's. Three additional features did result in better validation scores. But adding additional features did not always increase the score. We theorize that issues with feature extraction and the number of rules and features restricted any additional benefit.

Figure 6 shows the same comparison as figure 5, but for Event 2. In addition to the updated FIS the size of the initial set of FIS's was doubled. Increasing the size of the

initial set of FIS's may have yielded a better result but may not have addressed issues with the adequacy of the FIS model. Because we combined several feature changes in the final FIS runs in order to meet the final challenge deadline it is not clear which activities impacted the performance the most.

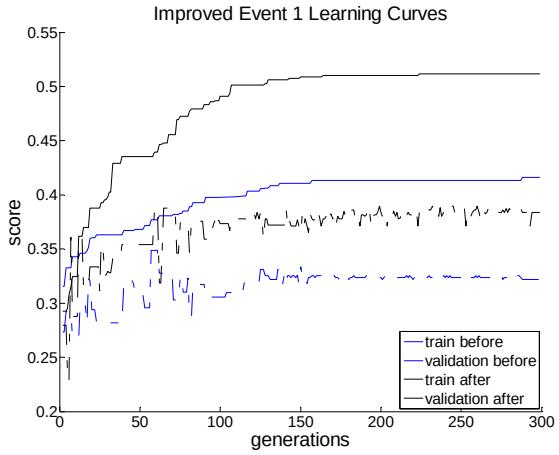

Figure 5. Learning curves for Event 1 before and after adding 3 features.

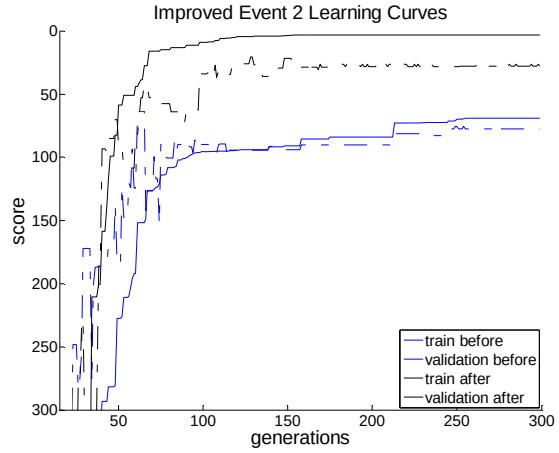

Figure 6. Learning curves for Event 2 before and after adding 3 features and doubling the initial size.

3. Results

Table 2 shows a comparison of the results of our Phase 1 and Phase 2 FIS's on set B, and the final results on set C. While we improved our overall performance between Phases 1 and 2, we did not place in the top 10 scores of the final ranking.

Table 2. Comparison of Phase 1, Phase 2, and final results.

Event	Phase I (set B)	Phase 2 (set B)	Final (set C)
Event 1	0.39	0.40	0.36
Event 2	94	60	67

4. Discussion

We felt that the ability to talk about the system with clinicians benefitted our development process in ways that are hard to quantify. However, our overall results were not especially encouraging.

One significant limitation was the amount of time it took to generate a solution. A medium size FIS took 8-10 hours to generate on a 3.2 GHz Xeon CPU. A few very large FIS's that we tried took up to 30 hours. Because of the statistical nature of the optimization we reran each configuration three times. The delays it took to generate and review each solution became a significant hindrance to the project.

It is not clear if the lower then desired final performance is due to our approach or due to the implementation details or due to not applying sufficient time to explore alternatives. Given more time we would have liked to include more features and rules and explored variations like including the type of ICU (a parameter that was added when Phase 2 started) [3].

Acknowledgements

John Pendergast RCP CRT, Faye Aebly RN, Traci Bartolomei RN BSN, and David Barash MD, for their patience and clinical insight.

References

[1] Krajnak M, Xue J. Optimizing fuzzy clinical decision support rules using genetic algorithms. Engineering in Medicine and Biology Conference 2006;5173-5176.

[2] Rui P, Moreno R, Metnitz P, Almeida E, Jordan B, Bauer P, Campos R, Iapichino G, Edbrooke D, Capuzzo M, Le Gall J. 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 Med 2005; 31:1345–1355.

[3] Silva I, Moody GB, Scott DJ, Celi LA, Mark RG. The PhysioNet/Computing in Cardiology Challenge 2012: predicting mortality of ICU patients. Computing in Cardiology 2012; 39 [in press].

[4] Xue Q, Taha B, Reddy S, Aufderheide T. An adaptive fuzzy model for ECG interpretation, presented at Marquette ECG Conference, 1998.

Address for correspondence.

Michael D. Krajnak
8200 West Tower Avenue
Milwaukee, WI 53223-3219