Acute Hypotension Episode Prediction Using Information Divergence for Feature Selection, and Non-Parametric Methods for Classification

PA Fournier, JF Roy

Carré Technologies Inc., Montreal, Quebec, Canada

Abstract

Acute hypotension is a critical event that can lead to irreversible organ damage and death. When detected in time, an appropriate intervention can significantly lower the risks for the patient. The objective of this work is to describe an automated statistical method that produces an automated method to predict acute hypotension episodes, using the least data possible.

We first detailed the problem of having more features than samples in the PhysioNet/CinC Challenge 2009 training set. We constrained our analysis to the largest common subset of features available for all patients (arterial blood pressure measurements). We then used information divergence (or Kullback-Liebler divergence) between two distributions to identify the most discriminative features. We used these features in each training set to classify the samples in the test sets using a nearest neighbors (NN) algorithm. With this method, we obtained a score of 9/10 for event 1, and 32/40 for event 2 compared to a control method which gives us 10/10 for event 1, and 35/40 for event 2. Our preliminary results showed that our method leads to significantly better than random results, therefore it increases our information about the samples in the test sets.

1. Introduction

We describe in this paper an automated statistical method that produces an automated method to predict Acute Hypotensive Episode (AHE), using a minimal subset of the available data. We constrained our model in the number of parameters to satisfy Occam's razor, which tells us that when we have two models that make the same predictions, we should take the simplest one. Our interpretation is that if we can make a prediction without using a parameter, we should get rid of it.

Our automated statistical method uses information divergence to select relevant features, and the nearest neighbors algorithm as the non-parametric classifier. We compared this method's results with an ad hoc method to validate its effectiveness on the PhysioNet/CinC challenge 2009 dataset. Considering that prediction within a forecast window was required, temporal analysis was done, and a forecast window length was selected minimizing the training error while respecting the challenge's constraints.

2. Data

The training data available for the challenge consisted of various vital signs : ECG waveforms, pulmonary arterial pressure (PAP) and arterial blood pressure (ABP) measurements and statistics in mmHg, central venous pressure, heart rate, respiration rate, SpO2, cardiac output, and alarms annotations.

Low blood pressure is usually defined as blood pressure of less than 90/60 mmHg or 90/50 mmHg. An AHE, which we had to detect, is defined as any period of 30 minutes or more during which at least 90% of the Mean Arterial Pressure (MAP) measurements were at or below 60 mmHg. We used this definition when we designed our control method for AHE prediction.

2.1. Challenge requirements

We constrained our analysis to the largest common subset of features available for all patients in the training and test sets. The SpO2 measurements seemed discriminant in the training datasets but has been rejected because test datasets did not contain this information. Hence, the common subset contained only ABP measurements (ABP_{Dias}, ABP_{Mean}, and ABP_{Sys}). We had access to almost continuous 10 hours recordings of these signals for 60 labelled patients. For event 1, we had 10 unlabelled patients to classify. For event 2, we had 40 unlabelled patients to classify.

Event 1 consisted in identifying if a patient belonged to an AHE group or to a group with no documented AHE during their hospital stay (subgroups H1 and C1 respectively). Event 2 consisted in predicting which patient had experienced an AHE within the forecast window. The fact that a patient may have experienced an AHE before or after but had been classified as without AHE (subgroup C2) added complexity to the analysis because it caused data fea-

tures to overlap each other. Another subgroup with AHE, but with patient without pressor medications (subgroup H2), was also included for event 2 to minimize pressors data bias implication.

We could have chosen to extract many statistics from this common subset of information. However, there was a reason why we wanted to choose a small subset of features instead of using them all at once.

2.2. Sampling in high dimensions

A danger when modeling distributions from samples in high-dimension is to overfit a complex model using too few samples. For example, it is always possible to separate N samples in $N-1$ dimensions (features) with linear separators[1]. For event 1, it means that you are assured to always separate the 30 training samples in 2 classes using 29 statistics on the data with a linear model. Overfitting the parameters makes the estimated statistical model useless in predicting the labels of new patients.

When sampling in high dimensions, a greater number of samples points are close to the edges of the domain. When it happens, we cannot consider the domain as "local". We call this phenomenon "curse of dimensionality" [2].

When using local methods such as k-nearest neighbors, we suppose the neighbors of the sample of interest will be local to it. In high dimensions this assumption is no longer true.

Furthermore, increased model complexity makes the error on the test sample increase with respect to the error on the training sample[3].

For all these reasons, it is not reasonable to try to infer a complex rule from high dimensional data, hence the need for a proper feature selection method for the CinC challenge.

2.3. Information divergence and Feature Selection

Information divergence (or Kullback-Liebler divergence[4]) is a symmetric measure of the difference between two probability distributions P and Q.

$$D_{KL}(P,Q) = \sum_i P(i) log(\frac{P(i)}{Q(i)}) + \sum_i Q(i) log(\frac{Q(i)}{P(i)})$$

We used this measure of information to identify the most discriminative features. To discretize the domain, we used 20 equal-sized bins divided from the minimum and maximum value of each feature dimension.

We then found the features θ that made the distributions P and Q the most divergent.

$$\hat{\theta} = \arg\max_{\theta}(D_{KL}(P_\theta, Q_\theta))$$

2.4. Temporal features

There are many ways to see time series from a statistical perspective. In this work, we compared the relevance of taking more or less time before T_0 in the training sets, in a single window or many consecutive windows.

Since we used the information divergence factor as the decision factor, we required a single value for each feature. When we used many consecutive windows, we kept only the window with the minimum value for each feature. This is justified because we are trying to identify AHE, which is by definition a lower value of the ABP.

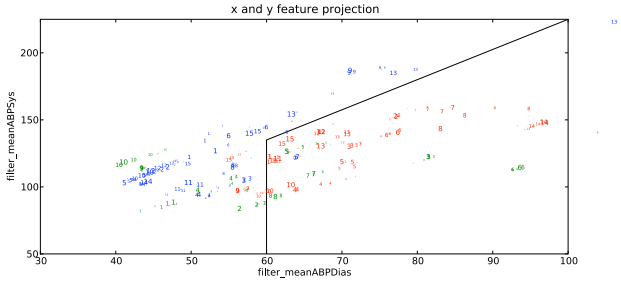

Figure 1. Mean ABP values for event 1, last 15 minutes before T_0, with 2.5 minutes windows.

Figure 1 shows the evolution of two features over time for different patients. Bigger numbers represent measurements closer to T_0.

3. Classification methods

In order to compare the behaviour of our completely automated method, we designed a control method based on human visual inspection of the data. This control method (threshold based on 2 features) is compared with the automated method (information divergence for feature selection and nearest neighbors for classification). We used the error on the training set to choose temporal features. We used the cross-validation approach to estimate the training error of the nearest neighbors algorithm.

3.1. Nearest neighbors

We used the two most discriminative features in each training set, defined by the information divergence, to classify the samples in the test sets using a simple nearest neighbors algorithm.

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

The nearest neighbors of a sample are the closest ones according to the Euclidian distance. The accuracy of this method depends on whether we have a labelled sampled

close to the new sample. In other words, the neighborhoods have to be "local". It does not perform well when outliers are present in the training set. In our analysis, $k = 1$.

3.2. Control method

We implemented the control method based on visual information of pairs of features. We manually selected the features we thought were good to separate the samples from the labeled distributions.

For event 1, we designed the control method knowing that 5 out of 10 had to be classified as AHE. The limit was set to 60 mmHg, according to the definition of the AHE, and a diagonal line was set to correctly classify two patients with high systolic blood pressure (see figure 2).

For event 2, we used the same kind of threshold as for the one of event 1, with the limit set to 50 mmHg because too many patients were classified with an AHE, and the challenge's rule was to have between 10 and 16 patients from that group.

3.3. Cross-validation

We estimated the classification error of the nearest neighbors algorithm on the training sets using each sample and comparing it to all $N - 1$ remaining samples (leave-one-out). This method is a special case of the K-fold cross-validation with $K = N$, and it is widely used when we have a small number of training samples. Leave-one-out is known to have low bias but possibly high variance [1].

4. Results

Our results showed that our automated statistical method leads to significantly better than random results, therefore it increases our information about the samples in the test sets. However, the control method gave us slightly better results for this challenge.

4.1. Procedure

Our final selection method for temporal windows follows these four steps :

1. Compute the error on the training samples for each window length and number.

2. Sort results according to the training error to select the window length with minimum error.

3. Apply analysis on the test samples.

4. Discard window length if the number of AHE detected does not respect the challenge's constraints (for event 2).

5. If two results are identical, choose the smallest analysis window.

For event 1, the procedure selected a single 2.5 minutes window for the control method, and a single 60 minutes window for the automated method.

Table 1. Training error on event 2 data using the control method with different window lengths and total duration (minutes).

len/dur	**2.5**	5	10	15	30	60
2.5	**16**	16	16	16	*15*	*15*
5	0	16	16	16	16	16
10	0	0	18	18	18	17
15	0	0	0	18	16	18
30	0	0	0	0	19	19
60	0	0	0	0	0	20

As depicted in table 1, the best results with the control method (2.5 minutes windows, 30 minutes duration) did not meet the challenge's requirement for event 2 (see step 4). The procedure selected a single window of 2.5 minutes for the control method, and 2.5 minutes windows on the last 30 minutes for the automated method.

4.2. Event 1 results

We achieved a 9/10 classification rate with the automated method and 10/10 using the control method. As shown on figure 2, the single error from the automated method was a false positive.

Table 2. Information divergence for event 1 features.

.	**0**	1	2	3	4	5
0	6.8	13.4	13.5	11.6	13.4	11.5
1	*13.4*	8.4	12.4	12.7	12.7	12.6
2	**13.5**	12.4	9.5	12.5	12.6	12.5
3	11.6	12.7	12.5	2.8	7.1	9.2
4	*13.4*	12.7	12.6	7.1	4.7	8.7
5	11.5	12.6	12.5	9.2	8.7	5.5

The information divergence matrix of 2-dimensional distributions using every pair of features is presented in table 2. The matrix is symmetric. The matrix shown here is for a single 2.5 minutes window before T_0. Features are mean (columns 0, 1, 2) and standard deviation (columns 3, 4, 5) values for ABP_{Sys}, ABP_{Dias}, and ABP_{Mean}. The features kept by the automated method in event 1 were the mean of ABP_{Sys} and the mean of ABP_{Mean}. We can see in the matrix that the pairs (mean ABP_{Sys}, mean ABP_{Dias}) (chosen in the control method) and (mean ABP_{Sys}, std ABP_{Dias}) have almost the same informational divergence as the one automatically chosen.

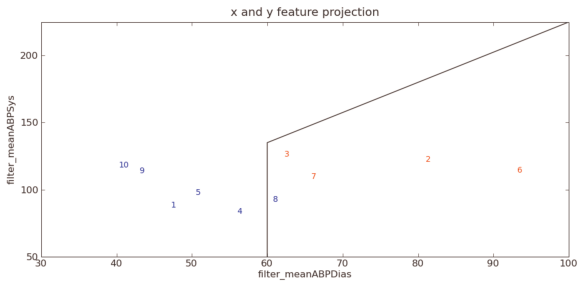

Figure 2. Comparison of automated (colors) and control (line) methods on test set for event 1.

Table 3. Performance of information-based and control methods on event 1.

	H	C	Accuracy
KL / 1-NN	6	4	90%
Control	5	5	100%
Required	5	5	

4.3. Event 2 results

We obtained a score of 32/40 with the automated statistical method, and 35/40 using the control method.

The features kept by the automated method in event 2 were the mean of ABP_{Mean} and the standard deviation of ABP_{Dias}. As shown in figure 3, there was 12 classification differences between the two methods.

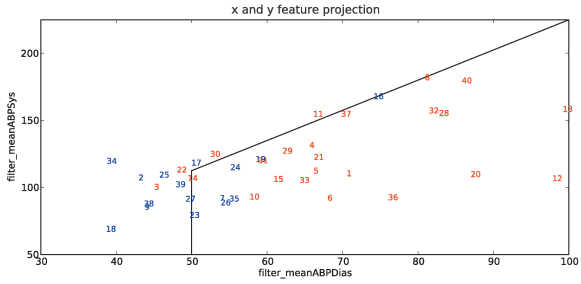

Figure 3. Comparison of automated (colors) and control (line) methods on test set for event 2.

Table 4. Performance of information-based and control methods on event 2.

	H	C	Accuracy
KL / 1-NN	16	24	80.0%
Control	15	25	87.5%
Required	10-16	24-30	

5. Conclusion

The perfect score obtained with the control method in event 1 can be explained by the simple fact that there were few samples to examine ; it was easy to clusterize them knowing that 50% were to be classified as with AHE. The completely automated analysis method based on information divergence proved to be effective with an almost perfect score of 90%. We could have achieved a perfect score simply by adding the AHE definition of MAP below 60 mmHg as a hard threshold but we wanted our automatic method to be automatically generated.

The 80% score obtained with the automatically generated method confirmed that purely statistical knowledge about the data can give results close to an ad hoc method based on human a priori on this dataset (87.5%).

A greater number of training samples would have considerably improved the information divergence automated method since it would define more clearly the boundary between groups. It would have allowed us to select a greater number of features instead of just 2, and would have enabled us to extend the 1-NN to a K-NN, which would have been more robust to noise and outliers. Cleaning training datasets of their outliers would also have improved the performance : these outliers were adding noise to the real distribution we wanted to estimate. Since the designed ad hoc method was built using visual inspection of the data, it may not have been possible to implement it with a large number of samples who would have required selecting more features in high-dimension. The latter case may benefit more from the information divergence method that the dataset used in this study.

References

[1] Hastie I, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY : Springer, 2001.

[2] Bellman R. Adaptive Control Processes. Princeton University Press, 1961.

[3] Vapnik V. The Nature of Statistical Learning Theory. New York, NY : Springer-Verlag, 1996.

[4] Kullback S, Leibler R. On information and sufficiency. Ann Math Statist 1951 ;22 :79–86.

Address for correspondence:

Pierre-Alexandre Fournier, Jean-François Roy
Carré Technologies Inc.
324, rue Villeray / Montreal (Quebec) H2R 1G7 / Canada
tel : +1 (514) 717-5226
fournier@carretechnologies.com