Detecting artifacts in clinical alerts from vital signs

Madalina Fiterau[1], Artur Dubrawski[1], Karen Chen[1], Donghan Wang[1],
Gilles Clermont[2], Marilyn Hravnak[2], and Michael R. Pinsky[2]

[1]Auton Lab, Carnegie Mellon University
[2]University of Pittsburgh

Abstract

Ensemble methods have shown predictive utility in analyzing vital sign (VS) data collected from physiologically unstable monitored patients. Training classification models requires labeled ground truth data obtained via laborious annotation or events or manual chart reviews by expert clinicians. We present an approach that distinguishes correct alerts from artifacts in multivariate non-invasive vital signs data collected at the bedside of critical care patients. Our framework makes the decision process which aids the expert adjudication transparent and comprehensible. The expert intervention is reduced to simply validating the outcome produced by an automated system using a small part of the available data.

1 Clinical alerts in vital sign monitoring data

Clinical monitoring systems are designed to process multiple sources of information about the current health condition of a patient and issue an alert whenever a change of status, typically an onset of some form of instability, requires the attention of medical personnel. In practice, a substantial fraction of these alerts are not truly reflective of the important health events, but instead they are triggered by malfunctions or inaccuracies of the monitoring equipment. Accidentally disconnected ECG electrodes, poorly positioned blood oxygenation probe, and many other such problems unrelated to the patient's clinical condition may in practice yield instability alerts. Frequency of such false detections has been shown to cause the alert fatigue syndrome, pervasive among medical personnel, particularly in critical care environments. Alert fatigue has adverse effects on the quality of care and patient outcomes. To maintain and enhance effectiveness of care, it is important to reliably identify and explain these non-consequential artifacts.

Noninvasive monitoring data including ECG-derived heart rate (HR), respiratory rate (RR), systolic and diastolic blood pressure (BP), and pulse oximetry (SpO_2) were sampled at 1/20Hz and alerts were issued whenever VS exceed any of preset stability thresholds ($40<HR<140$/min, $8<RR<36$/min, $SpO_2>85\%$). Each alert is associated with a category indicating the type of the chronologically first VS signal that exceeds its stability limits. A total of 147 statistical features were extracted from each raw VS stream independently during the alert window. The data density or duty cycle is the normalized count of signal readings during the alert period. A low value of this metric indicates the temporal sparseness of the data, while a value of zero simply means there was no data captured in that period. We also record the minimum and maximum of the first order difference of VS value during alert window. Extreme values of these statistics typically indicate a sharp increase/decrease of the VS value. The difference of means of VS values for the 4-minute window before and after the alert is also used, as is the value of the slope which results from fitting linear regression to the VS values versus the time index.

2 Projection-assisted annotation of alerts

In order to ascertain the effectiveness of the informative projection models in assisting domain experts, we have performed a user study in which two expert clinicians were asked to adjudicate alerts based on the projection models and, separately, based on vital signals. An example of the visual representations shown to clinicians for adjudication is in Figure 1. In total, 80 samples were labeled, each sample being assigned four scores, two by each clinician, one based on the projection as well as one the vital sign time series for the alert. The scores range from -3, indicating high reviewer confidence that the alert constitutes an artifact, to 3, indicating high reviewer confidence that the alert is real. Based on the scores assigned by each reviewer, the alert falls into one of three confidence categories, represented in Table 1. If there is disagreement between reviewers, or a reviewer is uncertain, the sample is marked as ambiguous and no label can be assigned.

Table 1: Annotation scoring matrix. Category and label assignment based on reviewer scores. C1 (strong agreement), C2 (weak agreement), C3 (disagreement). A (artifact), R (real alerts), - (ambiguous sample, no label assigned).

		Reviewer 1 Confidence						
	Category	**3**	**2**	**1**	**0**	**-1**	**-2**	**-3**
Reviewer 2 Confidence	**3**	R	R	R	-	-	-	-
	2	R	R	R	-	-	-	-
	1	R	R	R	-	-	-	-
	0	-	-	-	-	-	-	-
	-1	-	-	-	-	A	A	A
	-2	-	-	-	-	A	A	A
	-3	-	-	-	-	A	A	A

By merging the reviewer scores as shown in Table 1., each of the samples is assigned a label and confidence category from the projection-assisted annotation, and a separate one from the adjudication based on vital signs. The latter is considered the ground truth. Table 2 shows the extent of overlap between the confidence categories. 36 samples are labeled with the same confidence (and label) irrespective of the manner of annotation, 3 samples that could not be annotated based on the trace were annotated by analyzing the projections, 10 of the samples were annotated with more confidence based on the VS, while the remaining 31 (38.75% of total) could not be annotated based on the projected representation, but could be adjudicated using VS. This experiment points out that projection-assisted annotation was useful in obtained labels for 35 samples.

Table 2: Categories of Projection-assisted labeling and VS-based labeling.

Number of samples		**Category of vital sign -based labeling**			***Total***
		C1	**C2**	**C3**	
Category of projection-assisted labeling	**C1**	19	0	1	*20*
	C2	10	3	2	*15*
	C3	24	7	14	*45*
Total		*53*	*10*	*17*	*80*

We have evaluated the success of the IP-assisted annotation by comparing the resulting labels with the ground-truth obtained through the VS analysis. The comparison, shown in Table 3, was performed separately for the sets of labels of the two experts and for the final labels obtained by combining their scores. 27 of the samples were correctly classified using the projections, 31 could not be classified due to either expert disagreement or at least one of the experts being uncertain, 4 artifacts could not be filtered out using the projections, while only one alert was missed. The remaining 17 samples could not be adjudicated even through the use of the time series. As shown in Table 3, combining the predictions of the two experts results in a more conservative estimate of the label, but it also decreases the number of mistakes compared to single-expert prediction.

Effective training of automatic alert adjudication systems, calibrated on selective human annotation, improves accuracy of automated adjudication, reducing clinician effort. The system adjudicated 75% of samples with high confidence. 32% of the unlabeled samples were classified as artifacts.

Table 3: Success of projection-assisted labeling compared to ground truth (VS-based labeling).

Number of samples	Correct IP-assisted classification	Inconclusive IP-assisted classification	Incorrect IP-assisted classification	Sample is ambiguous
Reviewer 1	31	34	10 (4FN [1], 5 FP)	5
Reviewer 2	43	25	11 (1 FN, 7 FP)	1
Final	27	31	5 (1 FN, 4 FP)	17

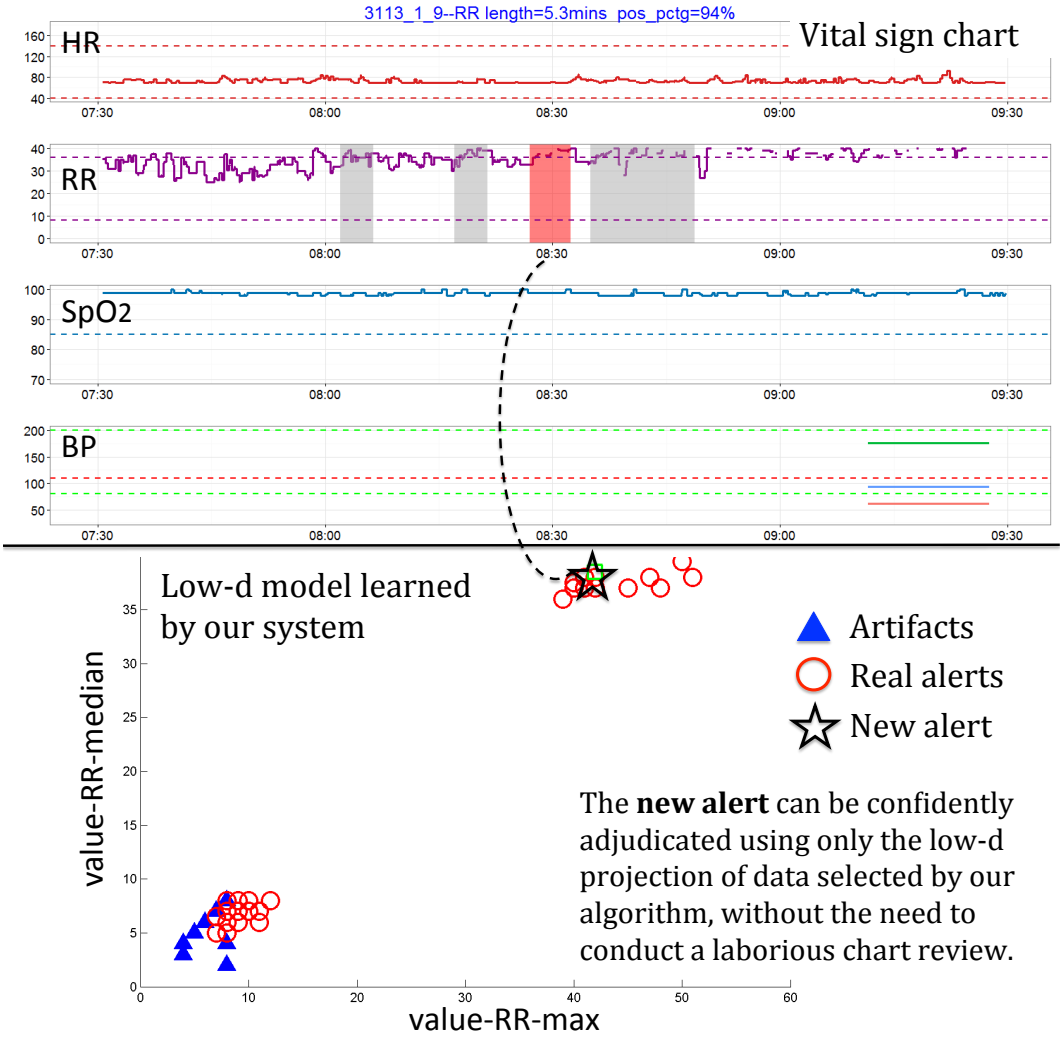

Figure 1: Example of projection-assisted annotation. Original vital sign chart (top) and informative projection (bottom). The RR alert that needs to be adjudicated, identified with a star symbol, is projected on the features value_RR_max and value_RR_median, amid previously labeled data. It is located in a cluster of data that were labeled as artifacts. Based on this informative projection, it was labeled as a real alert by both clinicians. Based on the time series corresponding to this alert, which is also represented at the top of Figure 1, the alert was also labeled as real. In this case, the outcome of using the compact representation using Informative Projection is the same as that of using the full time series representation with the added benefit that adjudication can be performed faster/easier by domain experts and that the labels can be automatically assigned by the system.