Title: Artifact adjudication for vital sign step-down unit data can be improved using Active Learning with low-dimensional models

Authors: Fiterau M, Dubrawski A, Chen L, Hravnak M, Clermont G, Bose E, Guillame-Bert M, Pinsky MR

Affiliations: Auton Lab, Carnegie Mellon University;, School of Nursing and Department Critical Care Medicine, School of Medicine University of Pittsburgh, PA, USA.

INTRODUCTION: Artifactual (false) alerts in physiologically unstable monitored patients cause alarm fatigue in clinical staff. Training a machine learning classifier for automatic artifact adjudication requires that a subset of data must first be labeled by clinicians, which consumes precious time.
OBJECTIVE: Demonstrate the use of active machine learning to select which multivariate vital sign (VS) alerts should be labeled by experts for the purpose of training a classifier to distinguish true alerts from artifacts to reduce labeling effort yet still achieve highly accurate automated alert adjudication.
METHODS: We collected noninvasive VS data including ECG-derived heart rate (HR), respiratory rate (RR), systolic and diastolic blood pressure (BP), and pulse oxygen saturation (SpO_2). Our monitoring system alerts whenever any VS exceeds pre-set stability thresholds (HR<40 or >140, RR<8 or >36, systolic BP <80 or >200, diastolic BP>110, SpO_2<85%). 812 samples (10% of the available alerts) were annotated by two experts as artifact or true alerts, of which 240 corresponded to alerts related to SpO_2. The raw monitoring data were then processed to extract features independently from each VS during the alert time over threshold and 4 minutes preceding its onset (alert period). The features include common statistics (mean, standard deviation, minimum, maximum), and features inspired by domain expertise (data duty cycle [% of non-missing data during alert period], minimum and maximum of first order differences, slope of a linear fit to data, etc.). We used machine learning system called ActiveRIPR to predict SpO_2 alerts, treating the expert-labeled data as the pool of samples available for active learning. We performed 10-fold cross-validation, training the ActiveRIPR model on 90% of the samples and using the remainder to calculate the learning curve.
RESULTS: Figure 1 shows a comparison of how the model accuracy varies as the samples are labeled for SpO_2 alert adjudication using different sampling functions specific to active learning (Uncertainty, Query-by-Committee, Information Gain and Conditional Entropy). Table 1 shows the number of samples needed to achieve target AUC accuracies of 0.85 and 0.88, averaged over all cross-validation folds. An accuracy of 0.85 (0.88) can be achieved by labeling 18% (25%) of all available samples using the Uncertainty (Information Gain) scoring function.
CONCLUSIONS: Alerts issued by VS monitoring systems can be accurately classified as artifacts/real alerts by an automated classification system which requires that only a small fraction of available reference data be manually labeled to train the classifiers.
FUNDING: NIH NINR R01NR013912; NSF 0911032, 1320347

Percentage of samples needed for classification		Target Accuracy 0.85	Target Accuracy 0.88
Feature Selection Strategy	*Uncertainty*	**18%**	55%
	Query by Committee	46%	48%
	Information Gain	21%	**25%**
	Conditional Entropy	43%	46%