34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

A Data Mining Approach to Reduce the False Alarm Rate of Patient Monitors

Benedikt Baumgartner, Kolja Rödel and Alois Knoll[1]

Abstract— **Patient monitors in intensive care units trigger alarms if the state of the patient deteriorates or if there is a technical problem, e.g. loose sensors. Monitoring systems have a high sensitivity in order to detect relevant changes in the patient state. However, multiple studies revealed a high rate of either false or clinically not relevant alarms. It was found that the high rate of false alarms has a negative impact on both patients and staff.**

In this study we apply data mining methods to reduce the false alarm rate of monitoring systems. We follow a multi-parameter approach where multiple signals of a monitoring system are used to classify given alarm situations. In particular we focus on five alarm types and let our system decide whether the triggered alarm is clinically relevant or can be considered as a false alarm. Several classification algorithms (Naive Bayes, Decision Trees, SVM, kNN and Multi-Layer Perceptron) were evaluated. For training and test sets a subset of the freely available MIMIC II database was used. Alarm-specific classification accuracy was between 78.56% and 98.84%. Suppression rates for false alarms were between 75.24% and 99.23%. Classification results strongly depend on available training data, which is still limited in the intensive care domain. However, this study shows that data mining methods are useful and applicable for alarm classification.

I. INTRODUCTION

Patient monitors are indispensable in today's medicine. Especially in intensive care units (ICU) they are of significant importance. They support the staff in the assessment of a patient's health status and give acoustical warnings or alarms if the physiological state of the patient needs attention or if there is a technical problem. However, several problems concerning those alarms have been reported. The produced noise can have a negative influence on both patients and staff. Gabor et al. [1] reported sleep disorders due to the noise on ICUs which led to a slowed recovery. Also Novaes et al. [2] emphazise the negative impact of the increased noise level in ICUs and found that machine alarms seem to disturb the members of the professional team even more than the patients themselves. Topf et al. [3] reported similar results highlighting increased sound levels as an ambient stressor. Furthermore, high rates of alarms without any clinical relevance have been reported. Chambrin et al. [4] reported a false positive rate of 74.2%. Siebig et al. [5] stated that only 17% of the alarms are clinically relevant and Görges et al. [6] identified up to 94% of the alarms as false. The high frequency of alarms, and especially of false alarms, also leads to a desensitization of the staff as reported by the German Association for Electrical, Electronic and Information Technologies (VDE) [7].

Usually patient monitors display multiple signals such as ECG or blood pressures. Alarms are often triggered if one parameter is out of range, e.g. increased heart rate (HR). The motivation for this work is to reduce the rate of irrelevant alarms by not only looking at one parameter for its own, but to incorporate the knowledge from several sensors at the same time. This approach is driven by human decision making: doctors or nurses assess the state of a patient by looking at multiple features at once and by relating them among each other instead of just observing one parameter. Aboukhalil et al. [8] reduced the incidence of false critical ECG arrhythmia alarms from 42.7% to 17.2% by relating the ECG data with the arterial blood pressure curve (ABP). Also Apiletti et al. [9] showed the potentials of a data mining approach and used clustering methods for automated risk assessment. In this work we follow a general data mining process which includes data preprocessing, the selection of adequate signal features and the training and evaluation of classifiers.

II. METHODS

A. Data Source and Selection Criteria

For training and test sets a subset of PhysioNet's MIMIC II Waveform Database [10] was used. The database comprises 4458 measurement records from 4099 patients. The number of signals within each record can vary. However, all records have a "Waveform" part containing up to four high-resolution signals (125 Hz). Furthermore, most of the records have a "Numerics" part with signals of low resolution (one value per minute). In addition to waveform and numerics data, metadata is available, providing age and sex information of the patients. Alarm notifications of the monitors are also included. They consist of a timestamp, the alarm type, the channel that caused the alarm and threshold information if the alarm was caused by a parameter being out of range.

Aboukhalil et al. [8] selected a subset of the MIMIC II database for alarm labeling and classification that fulfilled two criteria: a critical ECG arrhythmia alarm was issued and one channel of ECG and an ABP waveform were present at the time of the alarm. They labeled five types of arrhythmia alarms: asystole (asystolic pause of 4 s), bradycardia (HR $<$40 bpm), tachycardia (HR $>$140 bpm), ventricular tachycardia (5 ventricular beats), and ventricular fibrillation

[1] Benedikt Baumgartner and Alois Knoll are with the Robotics and Embedded Systems Group, Department of Computer Science, Technische Universität München, Boltzmannstr. 3, 85748 Garching b. München, Germany, baumgarb@in.tum.de . Kolja Rödel did his Master's Thesis with the group.

This work was supported by the Graduate School of Information Science in Health (GSISH) and an unrestricted grant from Bayerische Forschungsstiftung.

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

TABLE I

DISTRIBUTION OF THE TRAINING SET OVER CLASS AND ALARM TYPE

Class	False Alarm	True Alarm	**Total**
Asystole	130	43	173
Brady	65	193	258
Tachy	169	802	971
V-Fib	92	33	125
V-Tach	315	282	597
Total	771	1353	**2124**

(fibrillatory waveform lasting for 4 s). Overall 5386 alarms were labeled and considered as ground truth. We chose their labeled subset for training and testing but imposed further requirements. Only records that contain the second Einthoven ECG derivation and the ABP signal were included. Furthermore the signals had to persist continuously over a time window (alarm scope) of 20 seconds around the alarm event (15 s before and 5 s after). If available, the pulmonary arterial pressure (PAP) was also part of the set. From the numerics part SpO2, the central venous pressure (CVP), age and sex of the patient were included. Only data with an associated alarm event and alarm label was included. The final set for training and testing contained 2124 labeled alarms. The set distribution with respect to the alarm type and the alarm class (true/false alarm) is displayed in Table I. Each alarm scope (20 s) in the set contains 7504 values: age, sex, CVP, SpO2 and from the waveform part ABP, ECG and PAP sampled at 125 Hz each. Fig. 1 exemplarily shows an alarm scope retrieved from the database.

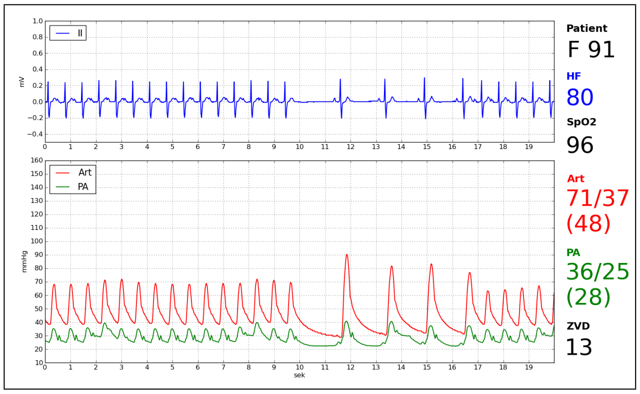

Fig. 1. Sample from the training set, showing ECG, arterial and pulmonary blood pressure as well as numerics data

B. Preprocessing and Feature Design

By the transformation to feature space the characteristics of a particular alarm scope are extracted. Moreover scaling issues are avoided and the dimension of the input data for classification is reduced. First, basic feature operations are implemented and then used in combination with several monitor signals.

For data from the numerics part of the set (Age, Sex, CVP, SpO2) there was no further processing necessary, since those values did not change during an alarm scope. A QRS detection on the ECG signal located the cardiac intervals (I, beat-to-beat distance) in the scope. Based on those intervals further time series were created: the heart rate (HR) in bpm, diastolic (BP_{dias}) and systolic (BP_{sys}) blood pressure time series (in mmHg) based on the min/max-values of the ABP and PAP data per interval:

$$HR(I) = \frac{60}{length(I)} \tag{1}$$

$$BP_{sys}(I) = max(BP(I)) \tag{2}$$

$$BP_{dias}(I) = min(BP(I)) \tag{3}$$

In the following 8 basic operations $f_{i=1,\ldots,8}$ are described to extract signal characteristics of the alarm scope X. Those features can be divided into three groups: scope-based, interval-based and sample-based features. For interval- and sample-based operations the results are aggregated such that the alarm scope is described by single values.

1) Scope-based features: The first two features are calculated directly from the samples $x_{i=0,\ldots,n}$ in the scope:

Mean value: An obvious feature of the alarm scope is the signal's mean value:

$$f_1(X) = \frac{1}{n}\sum_{i=0}^{n} x_i \tag{4}$$

Dispersion: The second feature we consider is the dispersion of the signal and illustrates how much the samples deviate from the average:

$$f_2(X) = \frac{\sum_{i=0}^{n} |x_i - f_1(X)|}{f_1(X)} \tag{5}$$

2) Interval-based features: The following features are calculated on the basis of cardiac intervals and aim at characterising the ABP and PAP signals:

Height: The height was computed as the difference of the maximum (systole) and the minimum (diastole) signal value of a cardiac interval.

$$f_3(I) = x_{sys} - x_{dias} \tag{6}$$

Width: The time difference between diastolic and systolic values represents the width of the interval.

$$f_4(I) = t(x_{dias}) - t(x_{sys}) \tag{7}$$

Area: The third feature extracted from the blood pressure signals is the area under the pressure curve and is calculated as the integral of each interval.

$$f_5(I) = (t(x_{dias}) - t(x_{sys})) \cdot \sum_{t=t(x_{sys})}^{t(x_{dias})} x_t \tag{8}$$

The interval-based operations f_3, f_4 and f_5 are aggregated for each alarm scope by only considering the minimum, maximum and mean value of all intervals in the scope:

$$\boldsymbol{f}_i(X) = \begin{pmatrix} min(f_i(I)) \\ mean(f_i(I)) \\ max(f_i(I)) \end{pmatrix}, \quad i = 3, 4, 5 \tag{9}$$

TABLE II

THRESHOLDS FOR THE NORMALITY RANGE FEATURE

Feature	c_-	c_+
HR (bpm)	40	160
ABP_sys (mmHg)	80	200
ABP_dias (mmHg)	40	110
PAP_sys (mmHg)	15	30
PAP_dias (mmHg)	4	12

3) Sample-based features: The last group of features is sample-based and includes a distance to normality measure, the slope and an offset measure of the signal. This last group of features is applied to the HR, BP_{sys} and BP_{dias} signals only. Those signals provide one sample per cardiac interval. To aggregate the feature values in the alarm scope 5 intervals were considered.

Normality Distance: Since many alarms are caused by exceedence of thresholds, we introduce a distance measure inspired by Apiletti et al. [9]. For each channel critical thresholds c_1 and c_2 are defined, and, if exceeded, the distance to those limits is calculated:

$$f_6(x_i) = \begin{cases} c_1 - x_i \Leftrightarrow x_i < c_1 \\ x_i - c_2 \Leftrightarrow x_i > c_2 \\ 0 \Leftrightarrow c_1 \leq x_i \leq c_2 \end{cases} \tag{10}$$

The signal-dependant thresholds were taken from [9] and are shown in Table II.

For the aggregation of the distance feature in the alarm scope, the sum and the maximum is considered.

$$\boldsymbol{f}_6(X) = \begin{pmatrix} sum(f_6(x_i)) \\ max(f_6(x_i)) \end{pmatrix} \tag{11}$$

Slope: The slope of a curve is determined by the difference between the values of two consecutive intervals and aims at detecting sharp changes which can indicate a dangerous situation for the patient.

$$f_7(x_i) = x_i - x_{i-1} \tag{12}$$

Offset: The offset feature, also inspired by [9], is similar to the slope. It does not regard a rapid change, but focuses on the trend over a few intervals. It calculates the difference between the current value and the average of the five previous intervals. For both slope and offset, several window sizes for the averaging were tried. A window of five intervals exhibited a good trade-off between balancing peak values and not blurring temporary signal aspects.

$$f_8(x_i) = |x_i - \frac{1}{5} \sum_{j=i-5}^{i-1} x_j| \tag{13}$$

The slope and offset features are aggregated again to represent the complete alarm scope. For that we consider the mean value and the maximum in the scope as well as the maximal sum of five consecutive intervals.

$$\boldsymbol{f}_j(X) = \begin{pmatrix} mean(f_j(x_i)) \\ max(f_j(x_i)) \\ maxsum(f_j(x_i)) \end{pmatrix}, \quad j = 7, 8 \tag{14}$$

TABLE III

DISCRETIZATION, FEATURE SELECTION AND CLASSIFICATION METHODS

Discretization Methods	Equal Width (6, 10 Bins), Equal Frequency (6, 10 Bins), Supervised Discrectization
Feature Selection Methods	Correlation-based Feature Selection (forward, backward), Supervised Feature Selection based on Information Gain, Relief Algorithm, Wrapper incl. Decision Trees, Wrapper incl. Naive Bayes
Classification Methods	Naive Bayes, kNN (1, 3, 5, 9 Neighbors), kNN + Distance Weighting (1, 3, 5, 9 Neighbors), Decision Tree (Confidence Level for Pruning: 0.2, 0.3, 0.5), Binary Decision Tree (Confidence Level for Pruning: 0.2, 0.3, 0.5), Unpruned Decision Tree, SVM incl. Polynomial Kernel (Classical, Standardized, Normalized, C=1, e=1), SVM incl. RBF Kernel (Classical, Standardized, Normalized, C=1, γ=0.01), Multi-Layer Perceptron (Learning Rate: 0.2, 0.3, 0.4, 40 nodes)

C. Classification and Evaluation Criteria

The Weka workbench [11] was used to design and evaluate the alarm classifiers. Combinations of 5 discretization, 6 feature selection, and 9 classification methods were applied on the test set. The algorithms that were used are enumerated in Table III. In addition to the algorithmic feature selection, feature sets with less than 3 or more than 20 features were excluded. They were considered as either too small to be representative or too voluminous with respect to computation time. We first evaluated classifiers on the complete training set before splitting the set with respect to the alarm types.

III. RESULTS AND DISCUSSION

Table IV lists the most successful combinations of discretization, feature selection and classification methods for the different sets with respect to classification accuracy. Due to the limited data set size all classifiers were evaluated by a 10-fold cross validation. The accuracy for the complete set was 84.70% with a 6-bin equal frequency discretization, a Naive Bayes wrapper for feature selection and a SVM with standardized RBF (C=1,γ=0.01). The suppression rate of false alarms was 71.73%. This comes along with a rather high suppression of true alarms (7.91%). The accuracy could be increased by splitting the data set with respect to the different alarm types with the best results for asystole alarms (98.84%). Surprisingly the accuracy for ventricular tachycardia alarms was worse than for the complete set (78.56%). Looking at the suppression rates, the results were also better with trained experts for each alarm type. The reduction rate of false asystole alarms was 99.23%. Again, ventricular tachycardia alarms were hard to classify with a rather low suppression rate of false alarms (75.24%) but a suppression of true alarms of 17.73%. Compared to the

TABLE IV

CLASSIFICATION RESULTS (10-FOLD CROSS VALIDATION) AND SUPPRESSION RATES FOR TRUE AND FALSE ALARMS

Training/Test Set	Discretization	Feature Selection	Classification Algorithm	TP	FN	FP	TN	Accuracy (%)	Suppresion Rate TA(%)	Suppresion Rate FA(%)
Complete Set	Equal Frequency (6 Bins)	Wrapper incl. Naive Bayes	SVM RBF Standardized	1246	107	218	553	84.70	7.91	71.73
Asystole	Supervised Discretization	Wrapper incl. Naive Bayes	Naive Bayes	42	1	1	129	98.84	2.33	99.23
Brady	Supervised Discretization	Wrapper incl. Decision Tree	kNN (1)	188	5	12	53	93.41	2.6	81.54
Tachy	Equal Width (6 Bins)	Wrapper incl. Naive Bayes	Binary Decision Tree (0.3)	789	13	33	136	95.26	1.62	80.47
V-Fib	Supervised Discretization	Correlation-based, forward	Unpruned Decision Tree	91	1	8	25	92.80	1.09	75.76
V-Tach	Supervised Discretization	Wrapper incl. Decision Tree	Binary Decision Tree (0.5)	232	50	78	237	78.56	17.73	75.24

work of Aboukhalil et al. [8] reduction of false alarms was significantly better for tachycardia and ventricular-related alarms. However, they report a 0% reduction of true alarms except for ventricular tachycardia alarms.

Training and testing classifiers on few instances can yield misleading results. In particular, this may apply to the classification of asystole (173) and ventricular fibrillation alarms (125). In addition, the class values of the training sets were unequally distributed (cmp. Table I). This was intensified by the limited size of the data set. E.g. asystole alarms naturally occur less frequently than bradycardia alarms. Using only a subset of the training data to include the same number of alarms in every set solves the distribution problem, but limits the overall data size. The available data has further implications: all alarm labels were obtained manually and even though they were declared gold standard by Aboukhalil, some labels are debatable as preliminary results of an online survey show (study under revision). Additionally the available data was a mix of true and false alarms but did not include missed alarm events or sections without a monitor triggered alarm event. Including such sections is part of future work.

Except for ventricular tachycardia alarms, the accuracies for a particular alarm type were above the one for the complete set. A distinct winner combination could not be determined. Yet, supervised discretization and correlation-based feature selection appeared most frequently among the top 5 combinations per alarm type (not shown). Regarding the classifiers, SVMs occured in the top 5 combinations for each alarm type, kNN in all but one. All evaluated algorithms achieved an accuracy rate of at least 71% except for V-Tach alarms (57%). The focus of this work was not to find an optimal method for alarm classification but to illustrate the applicability of data mining to the problem of false alarm rate reduction. A larger database with equally distributed alarm types is desirable to foster the results. Nevertheless one should be aware that patient safety is the primary goal in ICU monitoring and that an alarm classification system as presented suppresses true alarms.

ACKNOWLEDGMENT

The authors would like to thank Dr. Ulrich Schreiber, Dr. Stefan Eichhorn, Dr. Markus Krane, Matthias Kornek, Alejandro Mendoza and Christian Becker from the German Heart Center Munich for valuable discussions and their support of this study.

REFERENCES

[1] J. Y. Gabor, A. B. Cooper, S. A. Crombach, B. Lee, N. Kadikar, H. E. Bettger, and P. J. Hanly, “Contribution of the intensive care unit environment to sleep disruption in mechanically ventilated patients and healthy subjects,” *American Journal of Respiratory and Critical Care Medicine*, vol. 167, no. 5, pp. 708–715, Mar. 2003.

[2] M. A. F. P. Novaes, E. Knobel, A. M. Bork, O. F. Pavo, L. A. Nogueira-Martins, and M. Bosi Ferraz, “Stressors in icu: perception of the patient, relatives and health care team,” pp. 1421–1426, 1999-12-18.

[3] M. Topf, “Hospital noise pollution: an environmental stress model to guide research and clinical interventions,” *Journal of Advanced Nursing*, vol. 31, no. 3, pp. 520–528, 2000.

[4] M.-C. Chambrin, P. Ravaux, D. Calvelo-Aros, A. Jaborska, C. Chopin, and B. Boniface, “Multicentric study of monitoring alarms in the adult intensive care unit (icu): a descriptive analysis,” *Intensive Care Medicine*, vol. 25, no. 12, pp. 1360–1366, 1999-12-18.

[5] S. Siebig, S. Kuhls, M. Imhoff, J. Langgartner, M. Reng, J. Schlmerich, U. Gather, and C. E. Wrede, “Collection of annotated data in a clinical validation study for alarm algorithms in intensive care?a methodologic framework,” *Journal of Critical Care*, vol. 25, no. 1, pp. 128–135, Mar. 2010.

[6] M. Görges, B. A. Markewitz, and D. R. Westenskow, “Improving alarm performance in the medical intensive care unit using delays and clinical context,” *Anesthesia & Analgesia*, vol. 108, no. 5, pp. 1546–1552, 2009.

[7] VDE, Ed., *Alarmgebung medizinischer Geräte*, 2010.

[8] A. Aboukhalil, L. Nielsen, M. Saeed, R. G. Mark, and G. D. Clifford, “Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform,” *Journal of Biomedical Informatics*, vol. 41, no. 3, pp. 442–451, Jun. 2008.

[9] D. Apiletti, E. Baralis, G. Bruno, and T. Cerquitelli, “Real-time analysis of physiological data to support medical applications,” *Information Technology in Biomedicine, IEEE Transactions on*, vol. 13, no. 3, pp. 313–321, 2009.

[10] M. Saeed, M. Villarroel, A. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. Kyaw, B. Moody, and R. Mark, “Multiparameter intelligent monitoring in intensive care ii: a public-access intensive care unit database.” *Critical care medicine*, vol. 39, no. 5, pp. 952–960, May 2011.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: an update,” *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.