Evaluation of On-Line Bradycardia Boundary Detectors from Neonatal Clinical Data

François Portet, Feng Gao, Jim Hunter, Somayajulu Sripada

Evaluation of On-Line Bradycardia Boundary Detectors from Neonatal Clinical Data

François Portet, Feng Gao, Jim Hunter, Somayajulu Sripada

Abstract—**this paper aims at investigating different methods for the detection of the start and end of bradycardias in heart rate signal of premature babies. We present two methods based on a disturbance detector and on a decision tree that are compared to classical thresholding approaches. Decision tree obtained the best detection results (Se=78.2%, PP=68.7%) against the disturbance detector (Se=90.2%, PP=61.3%) and the best thresholding method (Se=92.5%, PP=46.5%). Moreover, the decision tree exhibits better performance for the boundaries estimation (median delay = 7-5 seconds) than the disturbance detector (median delay = 8-5 seconds) with a better stability (STD=8.5 to 8.7s vs. STD=35.3 to 19.9s). These methods will be integrated to the BabyTalk project which aims at summarizing neonatal clinical data as text in order to improve data management in Neonatal ICU.**

I. Introduction

Interpretation of clinical data is an important task in the intensive care unit (ICU). Medical staff deals with high volumes of data which are so large (about 1 MB per patient per day), that attention overload and stress from looking after several patients can lead to mistakes. Although graphical presentation of the data is the norm, it has been shown that, in certain cases, textual descriptions of data can lead to better clinical decision [3]. Automatic textual summarization of Neonatal ICU (NICU) data is the aim of the BabyTalk project [2]. This project, try to automatically generate human-like clinical data descriptions such as the following:

- "There is a momentary bradycardia";
- "there are several significant bradycardias"; …

The inference of adjectives such as "momentary", "significant", and "several" requires correct detection of the start and end of the bradycardia events. To do so, we investigated the performances of different methods for the detection of the start and the end of bradycardias in the Heart Rate (HR) signal. However, we expect that the developed methods can be adapted to the detection and characterization of other clinical events such as desaturation, hypoxia, hypotension, and apnea.

The results of this study will be useful not only for characterization and summarization of data but also for improvement of clinical monitoring systems. Indeed, in clinical monitoring, a pair of thresholds is usually set in order to trigger an alarm when physiological parameters (such as the heart rate and the oxygen saturation) go outside the range defined by the thresholds [4]. This crude method leads to a high number of false alarms and a very poor specification of the event. A more accurate detection of bradycardias could leads to the computation of their degree of significance (the deeper and the larger the bradycardia, the more dangerous) that can be used to trigger an alarm only when it is clinically necessary [5].

Section II details the methods implemented. The dataset used for the evaluation is presented in section III and the evaluation methodology is introduced in section IV. Then, the training and the test results are given in sections V and VI. Finally, this paper ends with a short discussion.

This work was supported by UK EPSRC grant EP/D049520.

Authors are with the Department of Computing Science, University of Aberdeen, King's College, Aberdeen AB24 3UE, UK, (corresponding author's phone: +44 (0)1224 274173; fax: +44 (0)1224 273422; e-mail: fportet@ csd.abdn.ac.uk).

II. Methods For Bradycardia Detection

Several bradycardia detection schemes have been proposed [4] but the definition of bradycardia differs between different investigators with variations in amplitude as well as in duration. In our study, a bradycardia is defined as "*a sudden fall from the Heart Rate (HR) baseline followed by a rapid recovery*". In our case (BabyTalk project), long term HR variations are managed a trend detection stage. Four different methods have been investigated. They are based on (i) amplitude threshold; (ii) amplitude and duration thresholds; (iii) disturbance detection and (iv) decision tree. These methods have been selected for their simplicity and their ability to process large amounts of data in a short time (which is not the case for more sophisticated methods [1]).

As ICU data are characterized by artifacts and missing values, the HR channel is filtered by a 10-order low-pass FIR filter to remove high frequency noise.

A. Amplitude (A) Threshold

This is the crude threshold method used by the monitoring systems. It simply detects sample below a predefined threshold (set by the medical staff). In the following, this method is used as reference.

B. Amplitude and Duration (A-D) Thresholds

This method is an improvement of the crude thresholding method. A bradycardia is detected when "HR < 100 bpm (beats per minute) during 15s or HR < 60 bpm during 5s" [4]. The duration thresholds improve the precision, however, it is not able to deal with bradycardias above 100 bpm or short bradycardias.

C. *Disturbance Detection (Disturb.)*

One way to detect a bradycardia is to detect a sudden change (disturbance) in the heart rate. Extending the method presented in [6], the algorithm searches for a sudden change in a moving window. If Max (reps. Min) represents the maximum (resp. minimum) amplitude in a 30 seconds window, a sudden change is detected each time that Max-Min>TC, where TC is the sudden change threshold. Sudden-change neighbors are then merged to give a larger window. If this window is classified as a downward spike for which the Min exceeds the normal range then, it is a bradycardia.

To estimate start and end, a baseline is computed by a 60 seconds-width median filter applied to HR. Each bradycardia previously found is then expanded forwards and backwards until it meets the baseline. Using this baseline computation, the method should detect accurately the boundary of bradycardia episodes

D. *Decision Tree (DT)*

Another way of detecting bradycardias is to apply a machine learning technique to learn a decision tree. Decision trees have already been used by Tsien *et al.* [7] to detect artifacts in ICU data. Our method is thus an extension of this work to the detection of medical events.

The induction of a decision tree is based on the "divide and conquer" principle to partition a training set TS, composed of individuals described by several attributes, into homogeneous subgroups. Let the classes of the individuals be $\{C_1, C_2,\ldots, C_k\}$, there are four possibilities:

1. TS contains individuals belonging to only one class C_j. In this case, this is a leaf named C_j.
2. TS is empty. In this case this is a leaf for which the class is defined by information other than TS (e.g. the most frequent class of the parent).
3. TS contains individuals belonging to several classes. Thus, a test T is chosen, based on a single attribute that has the exclusive outcomes $\{O_1, O_2,\ldots, O_n\}$ which are used to partition TS into the subsets $\{TS_1, TS_2,\ldots, TS_n\}$ where TS_i contains all the individuals in TS that have outcomes O_i.The decision tree for TS consists of one branch for each outcome. This mechanism is then reapplied recursively on each subset TS_i.
4. TS contains individuals belonging to several classes but for which no test can be found. In this case this is a leaf for which the class is defined by information from TS (e.g. the most frequent class in TS) or other than TS.

The performance of the decision tree induction rests mainly on the choice of the test. In this paper, we use the well known C4.5 method derived by Quinlan [9] which uses the gain ratio to choose the test T. The gain ratio can be described as the gain of information (based on the entropy) for T normalized by the potential information of dividing TS into n outcomes. Therefore, the decision tree chooses the most discriminant tests (best separation of the information). So, the learnt decision tree will not only be used to detect bradycardias but will also give information on what the most discriminating attributes are. The attributes used to describe each sample of the signal are: raw value, area, linear slope, standard deviation, minimum amplitude, maximum amplitude, max derivative, mean, and median. Area is the sum of gaps between the amplitude of the samples and the mean. These attributes are computed for each data point in three moving centered windows, 5, 10 and 30 seconds, in order to take different change rates into account.

Once learnt, the decision tree is translated into rules to detect bradycardias. As the decision tree classifies each sample as belonging to a bradycardia or not, it does not output intervals. That is why a translation stage, consisting in smoothing the outputs of the decision tree by a 31-samples Blackman window is added. The output is consistent with probabilities and a threshold stage determines the intervals.

This method has several advantages: 1) regarding the mass of data present in ICU, the learning requires little preprocessing of the large volumes of ICU data; 2) the rules derived from decision trees are explicit and checkable by human experts; 3) the rules enable us to investigate the most significant attributes and relations; 4) the decision tree processes large amounts of data in a short time, which is required in ICU monitoring.

III. DATA

The data set consists of 13 24-hour records from premature babies receiving intensive care in the neonatal unit at the Royal Infirmary of Edinburgh [1]. Each record consists of heart rate, blood pressures, core and peripheral temperatures, oxygen saturation, TcPCO, TcPCO2 and humidity of the incubator all acquired with a sampling frequency of 1Hz. The periods of bradycardia were annotated by two clinical experts (as were other clinically significant events).

Figure 2 represents an excerpt of NICU data and shows the difficulty of the task.

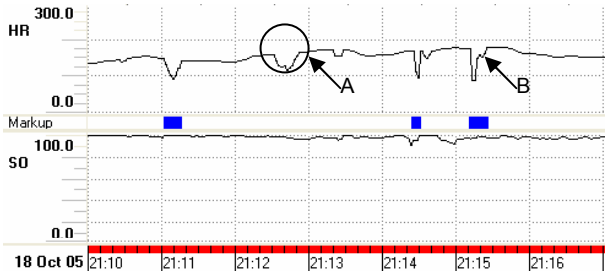

Fig. 2. Example of NICU data. HR signal (top chart) is followed by a markup channel that shows the periods of bradycardia annotated by two clinicians. HR: Heart Rate; SO: oxygen saturation.

Three intervals (Markup channel shown just below the HR) have been annotated as being bradycardias; however the discrimination of the first interval (21:11) from the downward spike (arrow A) is not straightforward. The last interval shows that the bradycardia covers the sudden fall from the baseline until the recovery. Even the period of high value (arrow B) is part of the bradycardia window. This kind of classification is impossible to do with a threshold based

method. The rest of the data shows that there is sometimes correlation with other channel (two momentary decreases in SO before the two last bradycardias) and sometime not (no obvious correlation with the first bradycardia). The dataset has been separated into a *training set* (5 records and 72 annotations) for optimization and learning and a *test set* (8 records with 174 bradycardias) for evaluation.

IV. Evaluation Method

Performances are assessed by computing the number of True Positives, TP (correct detections), False Negatives, FN (missed detections) and False Positives, FP (false alarms).

For each actual annotation of a bradycardia, the complete width is considered (the bradycardia window). Every detection of a bradycardia interval that overlaps a bradycardia window is a TP. If several detections overlap the same bradycardia window, only one is counted as TP, the others are ignored. Every detection that does not overlap a bradycardia window is a FP. Finally, every bradycardia window that does not overlap a detection is a FN. These values are used to compute three common criteria: Sensitivity (Se) = TP/(TP+FN), Positive Predictivity (PP) = TP/(TP+FP), and F-Measure (FM) = 2*Se*PP/(Se+PP).

Boundaries identification (start and end) are assessed by computing, for each TP, the start delay (resp. end delay) between the beginning (resp. end) of the detected interval and the beginning (resp. end) of the bradycardia window. When a bradycardia window contains several detections, the TP with the greatest delay is chosen.

V. Training

The training set was used to optimize each method and to learn the decision tree.

A. Optimization

The cut-off frequency *fc* of the input filter was chosen to remove only 15% of the bradycardia segments energy of the training set. *fc* was found to be 0.124Hz and led to the removal of 21% of the entire training set energy.

The threshold-based detector has been optimized on the training set by applying a set of thresholds and retrieving the one which leaded to the best FM value. The threshold found was 126bpm. A-D threshold technique has been used as described in Section II.B, except for the first threshold which has been optimized to 126bpm.

The disturbance detector was first evaluated without a thresholding stage. On the training set it demonstrated the following performance: Se=97.3% PP=44.44%. This shows that it is able to detect almost all bradycardias and that the next operation is to select the best one among all the candidates. To do so, the threshold of the disturbance detector was then set according to another study for which different HR values of babies in NICU were used to compute a linear regression. This model has been used to set the normal HR range according to the baby's gestation. The threshold was set to the lowest value of the normal range (here 129bpm).

B. Decision Tree Induction

The decision tree learner used was J48 - a C4.5 implementation from the Weka software toolbox [8]. The tree learning parameters were chosen to optimize the global size and the leaf size. Two classes where considered: the negative (not a bradycardia) and the positive (bradycardia). Fig. 3 gives the decision tree learnt from one record of the training set. With a size of 47 and with 24 leaves, it is possible for an expert to interpret it.

```
HRMovingMin_30 <= 124.225037
|   HRMovingDeviation_30 <= 27.484715
|   |   HRMovingMax_30 <= 183.449341
|   |   |   HRMovingMean_30 <= 147.056076
|   |   |   |   HRMovingMin_30 <= 79.825233: no (225/4)
|   |   |   |   HRMovingMin_30 > 79.825233
|   |   |   |   |   HRMovingMedian_10 <= 101.546494: yes (38/2)
|   |   |   |   |   HRMovingMedian_10 > 101.546494
|   |   |   |   |   |   HRMovingMaxDeriv_30 <= 0.494949
|   |   |   |   |   |   |   HRMovingMedian_10 <= 151.362595
|   |   |   |   |   |   |   |   HRMovingMin_30 <= 115.572456: yes (84/5)
|   |   |   |   |   |   |   |   HRMovingMin_30 > 115.572456: no (26/8)
|   |   |   |   |   |   |   HRMovingMedian_10 > 151.362595: no (29/3)
|   |   |   |   |   |   HRMovingMaxDeriv_30 > 0.494949
|   |   |   |   |   |   |   HRMovingBestSlope_10 <= 0.55661
|   |   |   |   |   |   |   |   HRMovingMax_10 <= 129.502609
|   |   |   |   |   |   |   |   |   HRMovingMax_5 <= 124.012764: no (46/9)
|   |   |   |   |   |   |   |   |   HRMovingMax_5 > 124.012764: yes (21/6)
|   |   |   |   |   |   |   |   HRMovingMax_10 > 129.502609: no (373/20)
|   |   |   |   |   |   |   HRMovingBestSlope_10 > 0.55661
|   |   |   |   |   |   |   |   HRMovingMax_30 <= 145.792984
|   |   |   |   |   |   |   |   |   HRMovingMedian_30 <= 125.16008
|   |   |   |   |   |   |   |   |   |   HRMovingMaxDeriv_30 <= 4.287994: yes (21/4)
|   |   |   |   |   |   |   |   |   |   HRMovingMaxDeriv_30 > 4.287994: no (26/6)
|   |   |   |   |   |   |   |   |   HRMovingMedian_30 > 125.16008: no (92/4)
|   |   |   |   |   |   |   |   HRMovingMax_30 > 145.792984
|   |   |   |   |   |   |   |   |   HRMovingMaxDeriv_30 <= 7.905991: yes (80/15)
|   |   |   |   |   |   |   |   |   HRMovingMaxDeriv_30 > 7.905991
|   |   |   |   |   |   |   |   |   |   HRMovingArea_30 <= 450.927032: no (23)
|   |   |   |   |   |   |   |   |   |   HRMovingArea_30 > 450.927032
|   |   |   |   |   |   |   |   |   |   |   HRMovingMaxDeriv_30 <= 10.476334: yes (21/2)
|   |   |   |   |   |   |   |   |   |   |   HRMovingMaxDeriv_30 > 10.476334
|   |   |   |   |   |   |   |   |   |   |   |   HRMovingMax_30 <= 154.33905: no (21)
|   |   |   |   |   |   |   |   |   |   |   |   HRMovingMax_30 > 154.33905: yes (23/10)
|   |   |   HRMovingMean_30 > 147.056076: no (281)
|   |   HRMovingMax_30 > 183.449341: yes (34/5)
|   HRMovingDeviation_30 > 27.484715
|   |   HRMovingDeviation_30 <= 57.645386
|   |   |   HRMovingArea_30 <= 527.860596: no (22/5)
|   |   |   HRMovingArea_30 > 527.860596
|   |   |   |   HRMovingMaxDeriv_30 <= 14.196518: yes (165/14)
|   |   |   |   HRMovingMaxDeriv_30 > 14.196518
|   |   |   |   |   HRMovingMaxDeriv_10 <= 16.057426: no (22/7)
|   |   |   |   |   HRMovingMaxDeriv_10 > 16.057426: yes (30/2)
|   |   HRMovingDeviation_30 > 57.645386: no (40)
HRMovingMin_30 > 124.225037: no (84644/15)
```

Fig. 3. Learnt decision tree. Every leaf is terminated by ": X (A/B)" or ": X (A)" where X is the class of the leaf (bradycardia or not), A gives the total number of instances that are classified by the leaf and B the number of instances incorrectly classified.

The first results to take from the decision tree are the attributes used. The main windows used have widths of 10 and 30 seconds. The 5 second window is used only once. This suggests that a sample need to be considered within a large neighborhood. The first test used is the minimum which separated a large part of the negative from the rest of the individuals. This first test actually reproduces the threshold method and its threshold (124 bpm) is very close to the optimal threshold (126 bpm). The standard deviation is then used to separate the very unstable (artifacts and bradycardias) from the more stable individuals (normal and

small bradycardias). The tree seems consistent with our expectations: small areas, high averages correspond to negatives and high slopes and derivative amplitudes correspond to positives and artifacts (negatives). The optimal threshold of the smoothing was found to be 0.24.

VI. RESULTS

The results of the evaluation performed on the test set are given table 1. The performance of each method is shown across a single row. The results include the median and the standard deviation of the start and end delays.

According to the F-Measure (FM), the results show the poor performance of the crude threshold detector (A Th.) which has high sensitivity (Se=92.5%) but a very low precision (PP=46.5%) thus a high number of false alarms. A-D Threshold shows that the duration thresholds lead to better PP (74.3%) but with a very low Se (43.1%). This method is thus unable to detect all the bradycardias. It has also the worst performance for the boundary detection.

TABLE I
BRADYCARDIA DETECTION AND BOUNDARIES ESTIMATION ON THE TEST SET

Methods	TP	FN	FP	start(s)	end(s)	Se	PP	FM
A Th.	**161**	**13**	185	$10^{\pm 6.4}$	$10^{\pm 6.5}$	**92.5**	46.5	61.9
A-D Th.	75	99	**26**	$11^{\pm 6.8}$	$11^{\pm 7.6}$	43.1	**74.3**	54.5
Disturb.	157	17	99	$8^{\pm 35.3}$	$5^{\pm 19.8}$	90.2	61.3	73.0
DT	136	38	62	$\mathbf{7^{\pm 8.5}}$	$\mathbf{5^{\pm 8.7}}$	78.2	68.7	**73.1**

The disturbance detector with an FM=73.0%, is the second best method slightly below the decision tree (DT) that exhibits the best FM value (FM=73.1%). However, although its median detection of the start and end are better than the threshold based one, the standard deviation of the estimation boundaries is high (35.3s and 19.8s). This suggests that its boundary estimation is not reliable. Finally, the decision tree demonstrates the best performance for boundary detection with median delays of 7 and 5 seconds only and is very reliable with standard deviations of 8.5 and 8.7 seconds only. Considering the inter-expert annotation variation of wave boundaries in other ICU areas [11], this is a very good result.

VII. DISCUSION

The results confirmed the already known poor performance of the threshold based detector. Our investigation of other techniques has led to better results and enabled us to emphasize further improvements.

The disturbance detector is promising as it is able to detect almost every bradycardia. However, a better method of classification, based on the area and the standard deviation could lead to important improvements. The detection of the start and end of the bradycardia by baseline estimation seems to suffer a lack of stability and needs to be refined.

The decision tree has shown a very good detection of start and end. This is understandable because the start and end of bradycardia have been explicitly learnt. This data mining technique has also given information on what kind of attributes can be most discriminating. It appears that large windows (10 and 30 seconds) are more useful than short windows (5 seconds) and a study including larger windows will be investigated to find what the best window size is.

The investigated methods even if they demonstrated better results than the thresholding methods, suffer from the corruption of data by high noise level episodes. The current filtering method also needs further investigation but anyway it will be unable to separate the part of noise that overlaps the frequency spectrum of the bradycardia. To face this problem, we plan to investigate the design of a Kalman filter and to study the modeling of the data by ARIMA as investigated by Quinn and Williams [1]. However, even this approach may not be the solution for all the cases of bradycardia. Indeed, the performance of every algorithm is related to the context in which it is used. One can perform better in the presence of high frequency noise; and another can be better in the presence of high HR variability. To benefit from the advantages of each detector a combined approach will be studied. Such an approach has already showed improvement in ECG analysis [10].

ACKNOWLEDGMENT

The authors would like to thank John Quinn for access to his dataset.

REFERENCES

[1] J.A. Quinn and C.K.I. Williams, "Known Unknowns: Novelty Detection in Condition Monitoring," to appear in *Proc IbPRIA2007*, 2007, Springer LNCS.

[2] F. Portet, E. Reiter, J. Hunter, J. Sripada, "Automatic Generation of Textual Summaries from Neonatal Intensive Care Data", to appear in *Proc of AIME 2007*, 2007, pp. 227–236.

[3] A.S. Law, Y. Freer, J.R.W. Hunter, R.H. Logie, N. McIntosh, J. Quinn, "A Comparison of Graphical and Textual Presentations of Time Series Data to Support Medical Decision Making in the Neonatal Intensive Care Unit," *J Clin Monit Comput.*, 2005, 19, pp. 183–194.

[4] JM. Di Fiore, "Neonatal cardiorespiratory monitoring techniques," *Semin Neonatol*. 2004, 9(3), pp. 195–203.

[5] M.-C. Chambrin, "Alarms in the intensive care unit: how can the number of false alarms be reduced?," *Crit Care*, 2001, 5(4), pp. 184–188.

[6] J. Yu, J. Hunter, E. Reiter and S. Sripada, "Recognising visual patterns to communicate gas turbine time-series data," in *Pro. of ES2002*, 2002, pp. 105–120.

[7] C.L. Tsien, I.S. Kohane and N. McIntosh, "Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit," *Artif Intell Med.,* 2000, 19(3), pp. 189–202.

[8] I.H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques," 2nd ed., Morgan Kaufmann, 2005

[9] J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, 1993

[10] F. Portet, A.I. Hernandez and G. Carrault, "Evaluation of real-time QRS detection algorithms in variable contexts," *Med Biol Eng Comput*, 2005, 43(3), pp. 379–385.

[11] R. Jané, A. Blasi, J. García and P. Laguna, "Evaluation of an automatic threshold based detector of waveform limits in Holter ECG with the QT database," in *Comp Cardio,* 1997, 24, pp. 295–298