Novelty Detection for Identifying Deterioration in Emergency Department Patients

David A. Clifton[1], David Wong[1], Susannah Fleming[2], Sarah J. Wilson[3,4], Rob Way[4], Richard Pullinger[4], and Lionel Tarassenko[1]

[1] Institute of Biomedical Engineering, University of Oxford, Oxford, UK
david.clifton@eng.ox.ac.uk
[2] Department of Primary Health Care, University of Oxford, Oxford, UK
[3] Heatherwood and Wexham Park Hospitals NHS Foundation Trust, Wexham, UK
[4] Oxford Radcliffe Hospitals NHS Trust, Oxford, UK

Abstract. This paper presents the preliminary results of an observational study into the use of novelty detection techniques for detecting physiological deterioration in vital-sign data acquired from Emergency Department (ED) patients. Such patients are typically in an acute condition with a significant chance of deteriorating during their stay in hospital. Existing methods for monitoring ED patients involve manual "early warning score" (EWS) systems based on heuristics in which clinicians calculate a score based on the patient vital signs. We investigate automated novelty detection methods to perform "intelligent" monitoring of the patient between manual observations, to provide early warning of patient deterioration. Analysis of the performance of classification systems for on-line novelty detection is not straightforward. We discuss the obstacles that must be considered when determining the efficacy of on-line classification systems, and propose metrics for evaluating such systems.

Keywords: Novelty Detection, Support Vector Machines.

1 Introduction

1.1 Early Warning Scores

Adverse events in acutely ill hospital patients occur when their physiological condition is not recognised or acted upon early enough [1]. Clinical guidance in the UK [2] recommends the regular observational recording of certain vital signs[1], combined with the use of EWS systems. The latter involve the clinician applying univariate scoring criteria to each vital sign in turn (e.g., "score 3 if heart rate exceeds 140 beats per minute"), and then escalating care to a higher level if any of the scores assigned to individual vital signs, or the sum of all such scores, exceed some threshold.

[1] heart rate (HR) measured in beats per minute, respiration rate (RR) measured in breaths per minute, blood oxygen saturation (SpO_2) measured as a percentage, systolic blood pressure (SysBP) measured in mmHg, etc.

H. Yin, W. Wang, and V. Rayward-Smith (Eds.): IDEAL 2011, LNCS 6936, pp. 220–227, 2011.

EWS systems have a number of disadvantages. (i) The scores assigned to each vital sign, and the thresholds against which the scores are compared, are mostly determined heuristically. However, a large evidence base of vital-sign data was used to construct the EWS proposed in [3]. (ii) EWS systems are used with periodic observation of vital signs, which may be made as infrequently as once every few hours in some wards. Patients may deterioriate significantly between observations. (iii) There is a significant error-rate associated with manual scoring, especially in the high-workload setting of the ED. (iv) Each vital sign is treated independently and correlations between vital signs are not taken into account.

2 Novelty Detection

This paper takes a novelty detection approach, in which a model of "normal" patient physiology (for adult in-hospital patients) is constructed. Novelty detection is typically performed in preference to a multi-class approach to classification when there are insufficient data to model abnormal states with any accuracy.

2.1 Manual Clinical Methods

For the purposes of this study, we will consider the performance of the heuristic EWS system that was in place in the ED at the time of data acquisition, which is summarised in table 1. We will also consider the "evidence-based" EWS system described in [3], which is summarised in table 2.

Table 1. EWS system used in the ED at the time of data acquisition

Score:	3	2	1	0	1	2	3
HR		≤ 40	41 - 50	51 - 100	101 - 110	111 - 129	≥ 130
BR	≤ 8			9 - 18	19 - 24	25 - 29	≥ 30
SpO_2	≤ 92			≥ 93			
SysBP	≤ 90	91 - 99		100 - 179			≥ 180

Table 2. Evidence-based EWS system proposed in [3]

Score:	3	2	1	0	1	2	3
HR	≤ 42	43 - 49	50 - 53	54 - 104	105 - 112	113 - 127	≥ 128
BR	≤ 7	8 - 10	11 - 13	14 - 25	26 - 28	29 - 33	≥ 34
SpO_2	≤ 84	85 - 90	91 - 93	≥ 94			
SysBP	≤ 85	86 - 96	97 - 101	102 - 154	155 - 164	165 - 184	≥ 185

2.2 Automated Methods - Estimation of the Joint Density

Previous work [6] has modelled the joint pdf $f(\mathbf{x})$ of vital signs $\mathbf{x} \in \mathbb{R}^4$, for the vital signs shown in table 1. Each vital sign was standardised with respect to

its own mean and variance, $x' = (x - \mu)/\sigma$. The joint distribution of the (normalised) training data was estimated using a mixture of Gaussian distributions, obtained as a Parzen window estimate with 400 components. The process used to estimate this distribution involved first summarising (using the k-means clustering algorithm) a set of approximately 2.3×10^6 data, corresponding to over 3,000 hours of vital-sign data acquired from acutely-ill hospital patients. The width parameter σ shared by all of the isotropic Gaussian distributions was set using an independent validation set [6].

The likelihood $f(\mathbf{x}|\boldsymbol{\theta})$ of previously-unseen test data $\mathbf{x}$ is then evaluated with respect to the Parzen window estimate (parameterised by $\boldsymbol{\theta}$) and used to generate a corresponding novelty score, $z(\mathbf{x}) = -\ln f(\mathbf{x}|\boldsymbol{\theta})$. This novelty score takes high values when the test data are "abnormal" with respect to f, and which thus take low probability densities as $f \to 0$.

A threshold κ is defined on z such that test data $\mathbf{x}$ are deemed "abnormal" with respect to the joint pdf if $z(\mathbf{x}) > \kappa$. In order to avoid false-positive alerts caused by transient noise and other artefact of short duration, this method only generates a novelty alert when $z(\mathbf{x}) > \kappa$ for four minutes out of any five-minute window of data. The value of κ was similarly selected using an independent validation set, selected from over 18,000 hours of vital-sign data acquired from acute patients [6].

2.3 Automated Methods - One-Class Support Vector Machine

We also consider the use of a one-class support vector machine (SVM), trained using the same data as that from which the density estimate described above was obtained. We used the method proposed by [7], in which the objective function is defined by separating the training data from the origin in the feature space defined by the SVM kernel, for which we use the Gaussian distribution[2].

The degree to which the SVM objective function is penalised by misclassifications (and thus the flatness of the decision boundary) is controlled by the C parameter, the value of which, along with the width parameter σ shared by all of the isotropic Gaussian kernels in the model, was selected using cross-validation, and was performed using the same independent validation set as was used for the density estimate described above [6].

The SVM produces a novelty score $z(\mathbf{x})$, which represents the distance between test data $\mathbf{x}$ and its decision boundary in the feature space defined by the Gaussian kernel[3]. One-class classification is performed according to the sign of z; i.e., test data $\mathbf{x}$ are classified "abnormal" if $z(\mathbf{x}) < 0$, and "normal" otherwise. To avoid false-positive alerts due to transient noise and artefact, as with the probabilistic method, an alert was generated if test data were classified "abnormal" for four minutes in any five-minute window of test data.

[2] This method typically performs similarly to the other popular one-class SVM formulation, the *support vector data description*, as proposed by [8].

[3] where that distance is normalised by the distance of the support vectors to the boundary [9].

3 Clinical Study

3.1 Overview

Vital-sign data were acquired from 472 adult patients during their stay in the ED of the John Radcliffe hospital, Oxford, using existing hospital bed-side monitors. These monitors provide measurements of HR, RR, and SpO_2 at a sampling interval of approximately 20 secs, and measurements of BP whenever the patient's blood-pressure cuff is inflated. Following [6], it was assumed that a blood-pressure measurement was valid for a period of 30 minutes after acquisition.

The total amount of data acquired from the 472 patients was approximately 1,708 hours. Patients were admitted to study (on a random basis) between January, 2009 and January, 2010, and patient consent was gained in accordance with approval from the Medical Research Ethics Committee (MREC).

3.2 Clinical Labels

To evaluate the performance of novelty detection, we would ideally have accurate labels of "normal" and "abnormal" episodes of data. The "gold standard" in classification problems is often a set of labels provided by domain experts - here, ED clinicians. However, such exhaustive labelling is typically not possible in practice due to the size of the datasets and the difficulty in determining patient abnormality from retrospective review of the vital signs. Furthermore, intra- and inter-expert variability makes the labelling process inaccurate.

An approach in which clinical experts are asked to review only vital signs from periods of suspected patient abnormality is often adopted. Here, *clinical escalations* have been taken as being indications of patient abnormality. These escalations are events that took place during the patient's stay in the ED, and they were identified retrospectively from the patient's written clinical notes.

There are many reasons for which a patient's care may be escalated in practice, only some of which will be associated with abnormal vital signs, and which could therefore be expected to be identified by an automatic method. Two clinical experts independently reviewed the patient notes and identified those periods during the patient's stay in the ED that corresponded to escalations that would be expected to be associated with abnormal vital signs. Any differences in opinion between the two experts were resolved by a third clinical expert, independently.

4 Methodology and Results

4.1 Obstacles to Evaluating Classifier Performance

The evaluation of the performance of classifiers with respect to discrete events within a time-series is a complex topic, which is reviewed in this section.

Independence. The classical method for evaluating classifier performance is to construct a confusion matrix, which quantifies the number of true and false, positive and negative classifications (TP, FP, TN, and FN) made by the classifier, with respect to the "ideal" classification. The sensitivity and specificity of the classifier may then be plotted as a function of some variable of its operation (typically a parameter that controls the decision threshold), to give the receiver operating characteristic (ROC) curve. Some EWS systems [10] have been constructed by maximising the area under this curve (AUROC). A *loss function* may be defined to assign different weights to false-positive and -negative errors if one is deemed more costly than the other. However, such weights are typically difficult to assign in practice.

The equivalent Bayesian methodology for evaluating performance in this manner is to integrate over the loss function, and select the classification parameters that minimise the expected loss ("risk") for each decision.

This approach is appropriate in the diagnostic context, as in mammography or blood chemical analysis, but are problematic when the classifier is used to analyse time-series data, as samples and events are not independent, but correlated. The results could be biased, for example, by a small number of "abnormal" patients with long hospitals stays (which is a common example, given that length-of-stay often correlates with abnormality), which contribute a large proportion of data to the set of "positives", but which are largely dependent. The performance of the classifier would be skewed towards how well it performed on this small number of patients.

We suggest that there is no simple answer to this problem, and that it is inappropriate to reduce the evaluation of classifier performance to a single metric (such as accuracy / AUROC).

Patient-Based Analysis. To use ROC-based performance metrics in time-series analysis, we must select a basic unit of analysis other than individual samples. To avoid breaking the independence assumption between basic units, we perform the analysis on a per-patient basis. In this study, we adopt the following convention:

"Event" patients: This group comprises all patients with one or more "events", the latter defined according to the criteria given in section 3 (i.e., those events with associated changes in vital signs). This corresponds to 34 (7%) of the 472 patients in our study[4].

"Normal" patients: This group comprises all patients who had no clinical escalation of any kind, and corresponds to 217 (46%) of the 472 patients in our study.

We define a TP classification to be an "event patient" for whom the first event was successfully detected; conversely, we define a FN classification to be an "event patient" for whom the first event was not successfully detected. The first event is used because the number of events from "event" patients varies between

[4] 75 (16%) of the 472 patients had no corresponding vital-sign data.

1 and 4 in our dataset. We define an event to have been detected if the method under evaluation generates an alert within some time w ahead of that event. We will consider the performance of our novelty detection systems as w is varied from [0 60] minutes, representing "early warning" up to an hour ahead of the event in the context of patient vital-sign monitoring in the ED.

We define a TN classification to be a "normal patient" for whom there were no alerts generated; conversely, we define a FP classification to be a "normal patient" for whom one or more alerts are generated by the classification system under test.

4.2 Results

Figure 1 shows the TP and FP results for candidate novelty detection methods, when evaluated on the per-patient basis described above. We compare (i) the density-based estimation method, (ii) the SVM, (iii) the heuristic EWS system that was used in the hospital at the time of the study, and (iv) the "evidenced-based" EWS system proposed in [3].

The two EWS systems (used with paper charts, in practice) were evaluated when applied to continuous data at frequencies of 30 minutes and 2 hours. The SVM and density-based methods were both trained on data obtained from a previous clinical study, as described in section 2. These training data were acquired from another set of acutely ill hospital patients. The density-based method allows the possibility of adapting its normalisation parameters (μ, σ for each vital sign) to those observed in the ED population, while still retaining the parameterisation $\boldsymbol{\theta}$ obtained from the original training set.

The results shown in the figure indicate that this "localisation" of the density-based method (whereby local population normalisation coefficients are used with an existing model) causes the performance of the model to improve (TPs increasing from 20 to 33). This increase in sensitivity is, however, matched by an increase in FPs from 34 to 66 for the original and "localised" density-based methods, respectively.

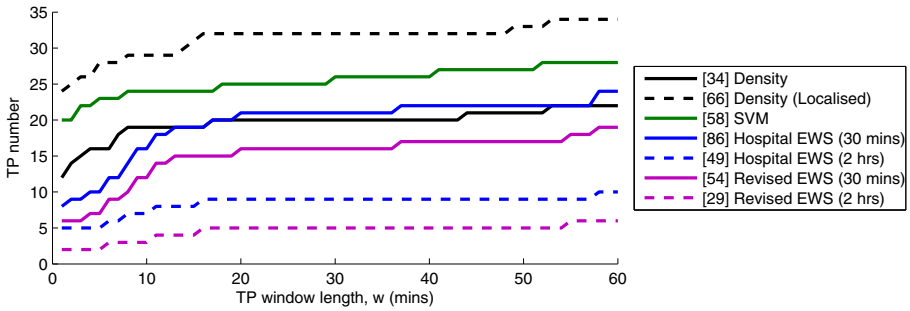

Fig. 1. TP numbers for each novelty detection method, shown as a function of w, the window-length used for determining if a physiological event was "detected". Numbers in square brackets (in the legend) indicate the number of FP classifications for each method when applied to 217 "normal" patients.

By comparison, the one-class SVM method outperforms the original density estimate, and approaches the performance of the "localised" density estimate, while having a lower FP rate, even without "localisation".

It may be seen from the figure that the manual EWS systems perform poorly in comparison with the more principled techniques described above. To match the number of TPs obtained with the density-based method, the hospital EWS system must be applied very frequently (every 30 minutes) - not a practical proposition in the clinical environment, due to the workload of clinical staff. Furthermore, at this level of TP classification, the number of FPs is particularly high (86). The FP number may be decreased by taking vital-sign observations less frequently: increasing the hospital EWS observation interval from 30 minutes to 2 hours reduces the FP number to 49, but this results in a very low sensitivity.

The effect of the "evidence-based" EWS system proposed in [3] is to significantly reduce the number of FPs in comparison to the hospital EWS system, with a small decrease in sensitivity as a result.

5 Conclusions

Analysing the performance of classifiers operating on time-series data in which discrete "abnormal events" occur is difficult. We have proposed a method to provide evaluation of classifier performance on a per-patient basis, and presented preliminary results in the context of a clinical study described in the ED.

We have demonstrated that paper-based EWS systems can be improved upon significantly by using automated methods when patients are continuously monitored with bed-side monitors (as in the high-acuity areas of the ED). We have examined the performance of density-based and one-class SVM approaches, and have shown that both provide an increase in sensitivity and specificity over existing EWS systems. Based on a training set acquired from a previous study, the SVM method outperforms the density-based method, although the latter can be improved by "localising" its normalisation coefficients to the ED population (whereas the SVM method must retain the existing normalisation coefficients of the input data, in order to retain the validity of its decision boundary in the high-dimensional feature space associated with the kernel).

It is possible that an on-line approach, which adapts to the new training data observed in the ED, would outperform both methods. Moving from a population-based to a patient-based modelling approach may also improve sensitivity to abnormalities. However, smaller quantities of training data would be available if patient-specific models were constructed, introducing significant model uncertainty, in which case a Bayesian approach is likely to be more appropriate.

Acknowledgements. The work described in this paper was funded by the NIHR Biomedical Research Centre Programme, Oxford. Dr. David Clifton was supported by the Wellcome Trust and the EPSRC under grant number WT 088877/Z/09/Z.

References

1. Safer Care for Acutely Ill Patients: Learning from Serious Accidents. Technical Report. National Patient Safety Association (2007)
2. Recognition of and Response to Acute Illness in Adults in Hospital. Technical Report. National Institute for Clinical Excellence (2007)
3. Tarassenko, L., Clifton., D.A., Pinsky, M.R., Hravnak, M.T., Woods, J.R., Watkinson, P.J.: Centile-Based Early Warning Scores Derived from Statistical Distributions of Vital Signs. Resuscitation (2011), doi:10.1016/j.resuscitation.2011.03.006
4. Williams, C.K.I., Quinn, J., McIntosh, N.: Factorial Switched Kalman Filters for Condition Monitoring in Neonatal Intensive Care. In: Advances in Neural Information Processing Systems, vol. 18, pp. 1513–1520. MIT Press, Cambridge (2006)
5. Tarassenko, L., Hann, A., Young, D.: Integrated Monitoring and Analysis for Early Warning of Patient Deterioration. Brit. J. Anaesthesia 98(1), 149–152 (2007)
6. Hann, A.: Multi-parameter Monitoring for Early Warning of Patient Deterioration. Ph.D. Thesis. University of Oxford (2008)
7. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the Support of a High-Dimensional Distribution. Neural Computation 13, 1443–1471 (2001)
8. Tax, D.M.J., Duin, R.P.W.: Data Domain Description using Support Vectors. In: Proc. ESANN, pp. 215–256 (1999)
9. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
10. Prytherch, D.R., Smith, G.B., Schmidt, P.E., Featherstone, P.I.: ViEWS - Towards a National Early Warning Score for Detecting Adult In-Patient Deterioration. Resuscitation 81, 932–937 (2010)