ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Prediction of preterm infant mortality with Gaussian process classification

Olli-Pekka Rinta-Koski[1], Simo Särkkä[2], Jaakko Hollmén[1], Markus Leskinen[3] and Sture Andersson[3]

1- Aalto University - Department of Computer Science
Espoo - Finland

2- Aalto University - Dept. of Electrical Engineering and Automation
Espoo - Finland

3- University of Helsinki, and Helsinki University Hospital
Helsinki - Finland

Abstract. We present a method for predicting preterm infant in-hospital-mortality using Bayesian Gaussian process classification. We combined features extracted from sensor measurements, made during the first 24 hours of care for 581 Very Low Birth Weight infants, with standard clinical features calculated on arrival at the Neonatal Intensive Care Unit. We achieved a classification result with area under curve of 0.94 (standard error 0.02), which is in excess of the results achieved by using the clinical standard SNAP-II and SNAPPE-II scores.

1 Introduction

This article is related to the use of data-driven methods in the context of digital healthcare and health informatics [1, 2]. In particular, our aim is to develop machine learning methodology for integration of heterogeneous data sources in order to more accurately predict the survival chances of preterm infants during treatment in the Neonatal Intensive Care Unit (NICU). First, we combine the conventional scoring system used in clinical practice with data-driven prediction from raw sensor data. Second, we study the prediction accuracy when the clinical scores are completely replaced with measurement data. The development of new methods for predicting neonatal in-hospital mortality is important, because while the global under-five mortality rate has dropped 53% since 1990, the proportion of neonatal deaths is projected to increase from 45% in 2015 to 52% by 2030 [3]. Furthermore, data-only prediction is extremely important in clinical work, because the determination of the conventional scores is labour-intensive and requires that a specific set of diagnostic markers is available.

The machine learning methodology is based on the use of Gaussian process (GP) classification [4] with features extracted from raw cardiac, arterial and oxymeter sensor measurements in addition to the clinical scores, gestational age at birth, and birth weight. GP classifiers have been previously used in health data analysis in (adult) Intensive Care Units (ICU) [5, 6, 7] and machine learning methods have been applied to NICU data [8, 9].

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

The contribution of this paper is that using cross-validation we show that augmenting the staff-determined SNAP-II and SNAPPE-II [10] scores with sensor measurements improves prediction accuracy over standard clinical measures. We also show that a data-driven prediction from measurements alone can lead to better prediction accuracy than SNAP-II and SNAPPE-II. The proposed approach gives the area under the receiver operator characteristic curve (AUC) 0.94 (standard error (SE) 0.02) for mortality prediction, which compares favourably with AUC 0.9151 reported for logistic regression by Saria et al. [9], and AUC 0.913 for CRIB-II and AUC 0.907 for SNAPPE-II reported by Reid et al. [11]. Although it has previously been shown [10] that in-hospital mortality of preterm infants is strongly correlated with birth weight and gestational age at birth, we show that the prediction result achieved by using these two variables alone (Table 2) can be improved by adding features extracted from measurement time series.

2 Materials and methods

2.1 NICU database

The NICU at Helsinki University Hospital has been collecting patient data in a database since 1999. Data include measurements of clinical parameters such as oxygen saturation by pulse oximetry (SpO_2) and supplemental oxygen levels, observations made by staff, and clinical outcomes. Our study cohort includes 2059 Very Low Birth Weight (VLBW) infants (birth weight < 1500 g) admitted between 1999–2013. Median gestational age at birth was 202 days (H28+6 weeks) and median birth weight was 1102 g.

The NICU database contains data recorded from equipment interfaces, as well as notations made by hand. Automatically gathered data consists of 111 different variables taken from monitor outputs of equipment used in the NICU. As the monitoring equipment as well as clinical guidelines have varied during the 15 year period under which the data has been stored, not all data is available for all 2059 patients. For the experiment, we chose 581 patients in the dataset for whom there is data for each of these seven variables: gestational age at birth, birth weight, systolic, mean, and diastolic arterial blood pressure, heart rate measured by electrocardiography (ECG), and SpO_2. 53 (9%) of these patients died in the NICU. We also decided to look at the first 24 hours from delivery to see whether the data gathered during that time period has predictive power. First 24 hours is a clinically relevant time frame, as most in-hospital deaths occur within the first week; median in this dataset is 5 days.

2.2 Preprocessing and feature extraction

The data was preprocessed by removing out-of-range values caused by e.g. misplaced or removed sensors and monitoring equipment drifting out of calibration from the time series.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

For feature extraction, mean and standard deviation were calculated from each of the following time series for each patient: systolic, mean, and arterial blood pressure, ECG heart rate, and SpO_2. SNAP-II score, SNAPPE-II score, gestational age at birth, and birth weight were directly used as features.

2.3 Gaussian process classifier

We used a GP [4] classifier with a probit measurement model:

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')), \quad p(y_i \mid f(\mathbf{x}_i)) = \int_{-\infty}^{y_i\, f(\mathbf{x}_i)} \mathrm{N}(z \mid 0, 1)\, dz, \tag{1}$$

where the classes are labeled as $y_i \in \{-1, 1\}$. The kernel was a sum of squared exponential (or radial basis function) kernel, linear kernel, and constant kernel:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_{\mathrm{se}}^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^{\mathsf{T}} \Lambda^{-1} (\mathbf{x} - \mathbf{x}')\right) + \mathbf{x}^{\mathsf{T}} \Sigma\, \mathbf{x}' + \sigma^2, \tag{2}$$

where $\Lambda = \mathrm{diag}(l_1^2, \ldots, l_d^2)$ and $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. For training the classifier we used the GPstuff Toolbox [12] with Laplace approximation on the latent variables and circular composite design integration over the hyperparameters.

In order to evaluate the performance of the classifiers we used k-fold cross-validation (CV) with $k = 4$. Cross-validation was used to estimate the classification accuracy, precision, specificity, and sensitivity as well as receiver operating characteristics (ROC) curve and the area under the curve (AUC).

3 Results

3.1 GP classification with SNAP-II and SNAPPE-II scores

First, we tested the performance of the GP classifier using only SNAP-II and SNAPPE-II scores with gestational age at birth and birth weight. Although this information is equal to what the scores are traditionally computed from, as can be seen by comparing Tables 1 and 2, the GP classifier is able to achieve a better AUC of 0.93 (SE 0.02) than the clinical standard SNAP-II (AUC 0.78, SE 0.03) and SNAPPE-II (AUC 0.84, SE 0.02) scores.

Next, we used all available signals with the GP classifier in order to get an upper bound on the achievable performance. All the available features were used as classifier inputs, in other words, SNAP-II, SNAPPE-II, gestational age at birth, birth weight, and the mean and standard deviation of each of the following: systolic, mean, and diastolic arterial blood pressure, ECG heart rate, and SpO_2. Table 3 shows the best five AUC scores from GP classification with all the features. The best achievable AUC is 0.94 (SE 0.02), but the results indicate that there are several input combinations that give very similar results; in this case it seems that including the SNAPPE-II score is indeed beneficial, as well as gestational age at birth and SpO_2. Different combinations of blood pressures also appear in the best results and ECG can be found in the best performing input combination.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

Reference	Acc	PPV	Sens	Spec	AUC
Trivial	0.91(0.00)	1.00(0.00)	0.00(0.00)	1.00(0.00)	0.50(0.00)
SNAPII	0.90(0.01)	0.56(0.18)	0.11(0.05)	0.98(0.01)	0.78(0.03)
SNAPPEII	0.91(0.00)	0.60(0.14)	0.29(0.05)	0.97(0.01)	0.84(0.02)

Table 1: Reference results. Trivial = trivial classifier that assumes all patients survive, SNAPII = SNAP-II with optimal (cross-validated) thresholding, SNAPPEII = SNAPPE-II with optimal thresholding. Acc = accuracy, PPV = positive predictive value, Sens = sensitivity, Spec = specificity, AUC = area under the receiver operator characteristic curve. Values in parentheses indicate the associated standard error.

SN	PE	GA	BW	BP_S	BP_M	BP_D	HR	O_2	Acc	PPV	Sens	Spec	AUC
□	□	■	■	□	□	□	□	□	0.91(0.01)	0.59(0.14)	0.38(0.16)	0.96(0.02)	0.91(0.02)
■	■	□	□	□	□	□	□	□	0.91(0.00)	0.61(0.13)	0.27(0.05)	0.98(0.01)	0.89(0.01)
■	■	■	■	□	□	□	□	□	0.92(0.00)	0.60(0.06)	0.30(0.05)	0.98(0.01)	0.93(0.02)

Table 2: Results using only SNAP-II, SNAPPE-II, gestational age at birth, and birth weight. SN = SNAP-II, PE = SNAPPE-II, GA = gestational age at birth, BW = birth weight.

SN	PE	GA	BW	BP_S	BP_M	BP_D	HR	O_2	Acc	PPV	Sens	Spec	AUC
□	■	■	□	■	■	■	■	■	0.92(0.01)	0.60(0.10)	0.68(0.12)	0.95(0.02)	0.94(0.02)
□	■	■	□	■	■	■	□	■	0.93(0.01)	0.73(0.12)	0.62(0.13)	0.96(0.02)	0.94(0.02)
□	■	■	■	■	■	■	□	■	0.92(0.01)	0.61(0.10)	0.61(0.11)	0.96(0.01)	0.94(0.01)
□	■	■	□	■	■	□	□	■	0.91(0.01)	0.53(0.06)	0.49(0.07)	0.95(0.01)	0.94(0.02)
□	■	■	□	□	■	■	■	■	0.92(0.01)	0.62(0.10)	0.49(0.03)	0.96(0.02)	0.94(0.02)

Table 3: The best five AUC scores using all available features. $BP_{S,M,D}$ = systolic/mean/diastolic arterial blood pressure, HR = ECG heart rate, O_2 = oxygen saturation (SpO_2).

3.2 GP classification with reduced feature sets

As the next test, we tested how the GP classifier performs when SNAP-II and SNAPPE-II scores are not used (Table 4) and finally when only the sensor signals (dropping also gestational age at birth and birth weight) are used (Table 5). The best performance without SNAP-II/SNAPPE-II is AUC 0.94 (SE 0.01). Using only sensor signals we obtain AUC 0.88 (SE 0.03). Both of these AUCs are better than the reference results shown in Table 1.

3.3 Receiver operating characteristic (ROC) curves

Figure 1 shows the receiver operating characteristic (ROC) curves for GPs with the best AUC scores as analyzed in the previous subsection. The ROC curves with all features and with all features except SNAP-II and SNAPPE-II also confirm the good performance of the GP classifiers compared with the clinical standard scores. The rightmost ROC figure also confirms that GP with only sen-

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

GA	BW	BP_S	BP_M	BP_D	HR	O_2	Acc	PPV	Sens	Spec	AUC
■	■	■	■	■	□	■	0.92(0.01)	0.63(0.10)	0.51(0.04)	0.97(0.01)	0.94(0.01)
■	■	■	■	■	■	■	0.93(0.01)	0.73(0.13)	0.48(0.13)	0.97(0.01)	0.94(0.01)
■	■	□	■	□	□	■	0.92(0.01)	0.60(0.09)	0.38(0.05)	0.97(0.01)	0.94(0.01)
■	■	■	■	□	□	■	0.91(0.01)	0.54(0.08)	0.43(0.03)	0.96(0.01)	0.94(0.01)
■	■	□	□	■	□	■	0.92(0.01)	0.64(0.12)	0.32(0.04)	0.98(0.01)	0.94(0.01)

Table 4: The best five AUC scores when using all the available features except SNAP-II and SNAPPE-II.

GA	BW	BP_S	BP_M	BP_D	HR	O_2	Acc	PPV	Sens	Spec	AUC
□	□	■	■	□	□	■	0.92(0.01)	0.66(0.15)	0.27(0.10)	0.98(0.01)	0.88(0.03)
□	□	■	■	■	□	■	0.92(0.01)	0.69(0.13)	0.30(0.10)	0.98(0.02)	0.88(0.02)
□	□	■	□	■	□	■	0.92(0.01)	0.71(0.12)	0.32(0.10)	0.98(0.01)	0.87(0.02)
□	□	■	□	□	□	■	0.92(0.01)	0.65(0.14)	0.26(0.08)	0.99(0.00)	0.87(0.03)
□	□	■	□	■	■	■	0.92(0.01)	0.70(0.11)	0.29(0.12)	0.98(0.01)	0.87(0.02)

Table 5: Five best AUC scores using only time series data.

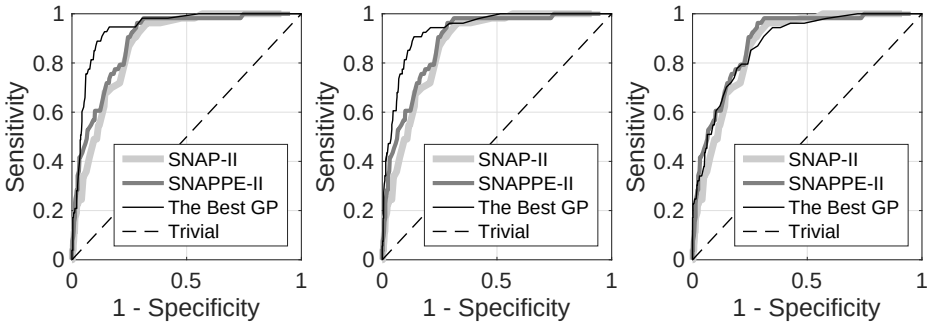

Fig. 1: ROC curves for GPs with the best AUC scores. Left: with SNAP-II & SNAPPE-II, Middle: without SNAP-II & SNAPPE-II, Right: sensors only. The ROCs of SNAP-II, SNAPPE-II, and the trivial classifier are also included.

sor measurements is also able to achieve the performance of the clinical standard scores.

4 Conclusions

Time series data from the first 24 hours of a preterm infant's intensive care unit stay can be used to improve the accuracy of existing methods for predicting in-hospital death. A Bayesian Gaussian process classifier can be used to create a predictive model. Combining features extracted from time series data with clinical scores calculated on arrival gives classification results in excess of clinical standards. Using only time series data gives results comparable with existing clinical standards.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

As current NICU patient data systems already collect sensory data used in this paper, predictive modeling could be included in the care process to give physicians advance warning of increased risk of in-hospital death. The model already outperforms existing methods in our retrospective cohort and with further refinement could prove to be a valuable clinical tool.

Acknowledgement. The authors would like to thank Professor Aki Vehtari for help in Gaussian process classifier design and implementation, and the Academy of Finland (projects 295505 and 266940) for financial support.

References

[1] N. Byrnes. Can technology fix medicine? *MIT Technology Review*, 2014.

[2] D. A. Clifton, K. E. Niehaus, P. Charlton, and G. W. Colopy. Health informatics via machine learning for the clinical management of patients. *Yearb Med Inform*, 10(1):38–43, 2015.

[3] D. You, L. Hug, S. Ejdemyr, and J. Beise. Levels and trends in child mortality. Report 2012. Estimates developed by the UN Inter-agency Group for Child Mortality Estimation. Technical report, United Nations Children's Fund (UNICEF), 2015.

[4] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[5] G. W. Colopy, Marco A. F. Pimentel, S. J. Roberts, and D. A. Clifton. Bayesian Gaussian processes for identifying the deteriorating patient. In *Proc. EMBC 2016*, 2015.

[6] M. Ghassemi, M. A. F. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *Proc Conf AAAI Artif Intell.*, pages 446–453, 2015.

[7] F Güiza, J. Ramon, and H. Blockeel. Gaussian processes for prediction in intensive care. In *Gaussian Processes in Practice Workshop*, pages 1–4, 2006.

[8] V. Gangadharan. *Automated multi-parameter monitoring of neonates*. PhD thesis, University College London (UCL), 2013.

[9] S. Saria, A. K. Rajani, J. Gould, D. L. Koller, and A. A. Penn. Integration of early physiological responses predicts later illness severity in preterm infants. *Sci Transl Med*, 2(48):48ra65, 2010.

[10] D. K. Richardson, J. D. Corcoran, G. J. Escobar, and S. K. Lee. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. *The Journal of Pediatrics*, 138(1):92–100, 2001.

[11] S. Reid, B. Bajuk, K. Lui, E. A. Sullivan, and NSW and ACT Neonatal Intensive Care Units Audit Group, PSN. Comparing CRIB-II and SNAPPE-II as mortality predictors for very preterm infants. *J Paediatr Child Health*, 51:524–528, 2015.

[12] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179, 2013.