MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Influence of Missing Values on Artificial Neural Network Performance

Colleen M. Ennett[a], Monique Frize[a,b], C. Robin Walker[c]

[a] *Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada*
[b] *School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada*
[c] *Department of Paediatrics, Children's Hospital of Eastern Ontario (CHEO) and University of Ottawa, Ottawa, Canada*

Abstract

The problem of databases containing missing values is a common one in the medical environment. Researchers must find a way to incorporate the incomplete data into the data set to use those cases in their experiments. Artificial neural networks (ANNs) cannot interpret missing values, and when a database is highly skewed, ANNs have difficulty identifying the factors leading to a rare outcome. This study investigates the impact on ANN performance when predicting neonatal mortality of increasing the number of cases with missing values in the data sets.

Although previous work using the Canadian Neonatal Intensive Care Unit (NICU) Network's database showed that the ANN could not correctly classify any patients who died when the missing values were replaced with normal or mean values, this problem did not arise as expected in this study. Instead, the ANN consistently performed better than the constant predictor (which classifies all cases as belonging to the outcome with the highest training set a priori probability) with a 0.6-1.3% improvement over the constant predictor. The sensitivity of the models ranged from 14.5-20.3% and the specificity ranged from 99.2-99.7%. These results indicate that nearly 1 in 5 babies who will eventually die are correctly classified by the ANN, and very few babies were incorrectly identified as patients who will die. These findings are important for patient care, counselling of parents and resource allocation.

Keywords

Neural Networks (Computer); Intensive Care, Neonatal; Decision Support Systems, Clinical; Infant Mortality; Outcomes Estimation; Missing Values

Introduction

Objective

To evaluate the classification performance of an artificial neural network (ANN) in predicting mortality in a neonatal intensive care unit (NICU) as the number of cases with one or two missing values replaced by normals are added to the database of cases with no missing values. This means that the mortality rate of the data sets will drop as more cases with missing values replaced by normals are added to the data sets, because babies with measurements for all of the tests are assumed to be very sick, and this group of infants had the highest mortality rate. If values were missing from the list of variables, it is most likely that these tests were not run, and therefore, the babies were less ill. The performance will be compared based on the classification rate, sensitivity and specificity.

Background

Data collection in hospital settings has become a common occurrence. Typically, the type of data collected depends on the medical environment and the prediction models presented in the literature. The Score for Neonatal Acute Physiology (SNAP) [1] was based on the adult intensive care unit (ICU) prediction model, the Acute Physiology and Chronic Health Evaluation (APACHE) [2], which assigned a weight ranging from 0 to 4 indicating a deviation from normality to each of 34 possible physiological measurements to reflect the patient's severity of illness. The SNAP Score weights ranged from 0 to 5 and had 37 possible measurements chosen based on clinical expertise. Logistic regression analysis using the SNAP Score and data from seventeen Canadian NICUs (members of the Canadian NICU Network) with slightly more than 20 000 babies led to the development of SNAP-II [3]. SNAP-II achieves the same performance as SNAP but requires only six of the 37 variables (lowest blood pressure, lowest serum pH, lowest temperature, lowest PO2/FiO2 ratio, seizures, and urine output – all variables were collected in the first 24 hours from admission to the NICU).

These scoring systems (APACHE, SNAP, SNAP-II) classify patients according to their severity of illness, which is closely correlated with mortality. Death is a rare outcome in the NICU given that the mortality rate is less than 4 percent. The variables collected for SNAP and SNAP-II often have missing values because certain tests or measurements are only done or recorded if deemed

necessary. This means that no baby has test results and measurements for all of the SNAP variables, therefore there are lots of missing values for SNAP. The SNAP Guidelines state that missing values are to be replaced with zeros (categorical variables with zeros used to indicate normal readings) for the scoring system because it is assumed that if the values were important for clinical management, the data would have been recorded [1]. Replacing missing values with normal values generally does not work as well with artificial neural networks (ANNs). This approach tends to skew the data towards normality. The database becomes too homogeneous and the ANN has difficulty identifying the features that lead to neonatal death. SNAP-II only has six variables, so some babies have data for all six variables. These babies are believed to be among the sickest as evidenced by a higher mortality rate for this group (9.7% versus a 3.9% overall mortality rate for all babies in the NICU).

Advantages of Modelling with Artificial Neural Networks

The advantage that ANNs offer over statistical techniques is that the model does not have to be explicitly defined before the experiments begin. There are no preconceived ideas about the model. ANNs can grasp the relevant data to develop the model, whereas to derive a statistical model, prior knowledge of the relationships between the factors under investigation is required [4]. One data set is used to train the network, and then the model is tested on a new data set to verify its performance.

Disadvantages of Modelling with Artificial Neural Networks

Unfortunately, ANNs are unable to work with incomplete data. This is an issue when working with medical databases because there are often missing values for a variety of reasons. It may be that the medical procedures were not needed or that the procedure was not available, or that the physiological measurements were taken but not recorded by the nurse perhaps due to time constraints. There are several methods to deal with this issue, as outlined below.

Missing Values

Statistics suggest some approaches to deal with missing values in medical databases. When investigating the influence of specific variables on the estimation of an outcome, the easiest way to deal with missing values is to simply delete all cases with missing values for the variables under consideration. This technique, however, may lead to the loss of potentially valuable information about patients whose values are missing. Also, the resultant database may be biased depending on the reason that the values were not entered into the database as, for example, when the values were not recorded because the test was not deemed necessary.

A second approach is to replace all missing values with the means (in the case of continuous variables). Due to the fact that in the NICU database variables with values are for tests run when necessary, thus replacing the missing values with the means might bias the database towards the sicker infants.

Finally, the third technique is to replace all missing values with normals. It is difficult to determine "normal" values for neonates because of the number of factors that might affect the different parameters such as birth weight and gestational age. Using the SNAP Guidelines we know that "normal" values are actually a range of values coded as 0. This can account for some of the variability in the "normal" range for neonates.

These three approaches were attempted by Tong [5] with the Canadian NICU Network database using the continuous values of the variables without much success. Each time the network parameters were changed in an attempt to optimize the network, the ANN classified every case as a survivor. In other words, the ANN became a constant predictor (a simple statistical benchmark where all cases are classified as belonging to the class with the highest training set *a priori* probability).

In the experiments for this study, the variables were classified as categorical values according to the SNAP Guidelines. Since the SNAP Guidelines state that missing values should be replaced with normal values (categorized as zeros), this is the approach selected for the study.

The goal was to determine the "breakpoint" for the ANN's performance using databases with missing values replaced by normal values. In other words, how many cases with missing values can be added before the ANN becomes a constant predictor?

Materials and Methods

Canadian NICU Network Database

The database of the Canadian NICU Network contains patient cases from seventeen NICU member centres from across Canada. Data was collected on Day 1, Day 3, Day 14, and Day 28 or discharge to allow for time-varying analysis in future work. The centres collected this information from January 8, 1996 to October 31, 1997. Infants who stayed in the NICU for less than 24 hours were excluded from the study.

This study used only the Day 1 database, which contained 20,001 cases excluding moribund babies (mortality rate = 3.9%). Only 5196 cases (mortality rate = 9.7%) had no missing values for the six variables under investigation. Of the remaining 14,806 cases with missing values (mortality rate = 1.8%), there were 6341 cases (mortality rate = 2.8%) that were only missing values for lowest PO2/FiO2 ratio and/or urine output.

Network Architecture

We used a three-layer feed forward network with two hidden nodes and the hyperbolic tangent transfer function trained with the back propagation algorithm written in MATLAB/C++. The weight-elimination cost function [6] was used to minimize overfitting. Our ANN was automated to optimize the nine network parameters (weight-elimination scale factor, learning rate, learning rate increment, learning rate decrement, momentum, error ratio, weight-decay constant, weight-decay constant increment, weight-decay constant decrement – defined in Appendix A) for the best test set correct classification rate [7]. The nine parameters were optimized one at a time while the other eight parameters were set to default values.

Table 1 outlines the different data sets that were used in this study. The first data set used only cases with no missing values for the six variables under consideration. The subsequent data sets included these cases with no missing values (n = 5196) plus a random sample of a specific percentage of patients from the data set with cases missing only urine output and the lowest PO2/FiO2 ratio (n = 6341). Two-thirds of the data set was used as the training set, and the remaining one-third became the test set.

Table 1 – Number of patients in data sets

Data set	Patients in data set	Training set	Test set
Only complete	5196	3426	1770
Complete + 30%	6755	4441	2314
Complete + 50%	7794	5163	2631
Complete + 85%	9613	6414	3199
Complete + 100%	10392	6914	3478

Table 2 describes the *a priori* statistics of each data set. Note that the survival rate of the data sets increase as more patients who are missing one or two values were added to the database.

Table 2 – A priori statistics of data sets (i.e., survival rate in each set)

Data set	Entire set (%)	Training set (%)	Test set (%)
Only complete	90.3	90.0	90.8
Complete + 30%	91.9	92.1	91.4
Complete + 50%	92.7	92.6	92.9
Complete + 85%	93.3	93.4	93.1
Complete + 100%	93.8	94.0	93.3

Results

The experimental test set results for the prediction of neonatal death in the NICU are outlined in Table 3. The correct classification rates (CCRs) of the ANN are compared with the constant predictor (CP) for the particular data set. In Table 3, we identify the percentage improvement of the ANN's performance over the constant predictor. The sensitivity and specificity of the classification results are presented to show the performance of the networks. Sensitivity is the number of patients who died that were correctly identified out of all patients who died. Specificity is the number of patients who survived that were correctly classified out of all patients who survived.

Table 3 – ANN test set experimental results for neonatal mortality prediction

Data set	CP (%)	ANN CCR (%)	% improvement over CP	Sens (%)	Spec (%)
Only complete	90.8	91.8	+ 1.0	18.5	99.2
Complete + 30%	91.4	92.7	+ 1.3	20.7	99.4
Complete + 50%	92.9	93.6	+ 0.7	20.3	99.2
Complete + 85%	93.1	93.9	+ 0.6	14.5	99.7
Complete + 100%	93.3	94.3	+ 1.0	14.7	99.7

Discussion

The expected result of these experiments was to find a "breakpoint" for the ANNs to show at what percentage of cases with missing values replaced by normal values would the network become a constant predictor. Given the ANN's previously identified difficulties with classifying highly skewed data [5], we expected that at one point the ANN's classification performance would be exceeded by the constant predictor. In fact, this was not what we found. In every case, the ANN achieved a 0.6-1.3% improvement over the constant predictor. This means that the network was correctly classifying individual patients better than the constant predictor. The high specificity (99.2-99.7%) also means that few babies are incorrectly identified as infants who will die. This is important when counselling parents about their child's prognosis, as well as for appropriate patient and resource management. At first glance, the sensitivity results look disappointing, but we must remember that the ANN is predicting death for *individual* patients. Although the sensitivity is low (number of babies who died that were correctly identified) ranging from 14.5-20.3%, this means that nearly 1 in 5 babies who die are identified. The cut-off point for these networks is 0.01 (the hyperbolic tangent transfer function has a range of –1 to 1). Certainly the cut-off point could be altered to get a higher sensitivity at the expense of a lower specificity.

There was not an obvious and consistent drop in the sensitivity of the ANN's performance as had been expected. The sensitivity of the data sets with a higher replacement of missing values is lower than for sets which had fewer missing values. The drop is not consistent and may be

related to the sampling method. This aspect will be investigated in the future.

Another interesting discovery was that the statistics for the mortality rates of the various databases confirmed that the patient data sets with fewer missing values also had a higher mortality rate. This is evident from the fact that the overall mortality rate of the database was 3.9%, but for the babies with no missing values for the six variables under consideration, the mortality rate was 9.7%. In the case of patients missing at least one value, the mortality rate was only 1.8%, and for those missing only urine output and/or lowest PO2/FiO2 ratio, the mortality rate was 2.8%.

Limitations

As noted previously, a limitation of this analysis is the exclusion of cases missing more than two values of the six variables investigated in this study. Another challenge is that a lower mortality rate makes it more difficult for the ANNs to correctly classify the outcome "death". These limitations may be particular to the NICU environment and data collection approach as set out by the SNAP Guidelines.

Conclusion

Since the ANN performance was always better than the constant predictor, the experimental results were very encouraging. We discovered that the ANN was able to classify the patients better than the constant predictor by 0.6-1.3%. The low sensitivity (ranging from 14.5-20.3%) and high specificity (ranging from 99.2-99.7%) mean that nearly 1 in 5 patients who die are correctly identified, and that very few patients who survive are incorrectly classified. These may be useful findings for clinicians when counselling parents, and for appropriate patient care and resource allocation.

Future Work

These experiments should be repeated several times to determine whether the results are fairly stable or whether they are influenced by the sampling method used to replace missing values with normal values.

As a follow-up to this study, cases missing more than just two values will be added to the data sets in incremental steps to continue to observe the impact of missing values on ANN performance. As well, we will look more closely at the fact that more missing values translates into a lower risk of death. Experiments with data sets that simply identify the presence or absence of a recorded result for each variable may lead to important findings for predicting survival in the NICU.

Acknowledgments

We wish to thank the members of the Canadian Neonatal Network and particularly Dr Shoo Lee of the University of British Columbia for allowing us to use the Canadian NICU Network database for this research. This work was completed with funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Medical Research Council of Canada (MRC).

References

[1] Richardson DK, Gray JE, McCormick MC, Workmann K, and Goldmann DA. Score for neonatal acute physiology: a physiologic severity index for neonatal intensive care. *Pediatrics* 1993: 91(3): 617-623.

[2] Knaus WA, Zimmerman JE, Wagner DP, Draper EA, and Lawrence DE. APACHE – acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med* 1981; 9(8): 591-597.

[3] Richardson DK, Corcoran JD, Escobar GJ, Lee SK, Canadian NICU Network (Walker CR, Member), Kaiser Permanente Neonatal Minimum Data Set Area Network, SNAP-II Study Group. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. *J Pediatr* 2000; 137 (In press).

[4] Livingstone DH, Manallack DT, Tetko IV. Data modelling with neural networks: Advantages and limitations. *J Comp-Aided Molecular Design* 1997; 11: 135-142.

[5] Tong Y, Frize M, and Walker R. Estimating ventilation using artificial neural networks in intensive care units. *Proc BMES-EMBS Conf* 1999.

[6] Weigend AS, Rumelhart DE, Huberman BA. Back-propagation, weight-elimination and time series prediction. In: Tourestzky DS, Elman JL, Sejnowski TJ, Hinton GE, eds. *Proc 1990 Connectionist Models Summer School*. San Mateo: Morgan Kaufmann, 1990; pp. 105-116.

[7] Frize M, Ennett CM, Charette E. Automated optimization of neural networks in estimating medical outcomes. *Proc ITAB-ITIS* 2000.

Address for correspondence

Colleen M. Ennett, BSc(Eng), MASc
Department of Systems and Computer Engineering
Carleton University
1125 Colonel By Drive
Ottawa, ON K1S 5B6
Canada
Email: Ennett@canada.com

Appendix A

Learning rate (lr): Determines the speed at which the network attains a minimum in the criterion function so long as it is small enough to insure convergence. If the learning rate is too high, it may oscillate around the global minimum, unable to converge.
Learning rate increment (lr_inc): The learning rate's incremental value.
Learning rate decrement (lr_dec): The learning rate's decrement value.
Weight-decay constant (λ): Determines how strongly the weights are penalized.

Weight-decay constant increment (λ_inc): Weight-decay constant's incremental value.
Weight-decay constant decrement (λ_dec): Weight-decay constant's decrement value.
Weight-elimination scale factor (w_0): Defines the sizes of "large" and "small" weights. When w_0 is small, small weights will be forced to zero resulting in fewer large weights (i.e., weight-elimination). A large w_0 causes many small weights to remain and limits the size of large weights (i.e., weight-decay).
Momentum (momentum): Adds a proportion of the previous weight-change value to the new value, thereby giving the algorithm some "momentum" to prevent it from getting caught in local minima.
Error ratio (err_ratio): Controls how back propagation makes adaptive changes in the learning rate, weight-decay constant, and momentum.