Proceedings of the 25th Annual International Conference of the IEEE EMBS
Cancun, Mexico • September 17-21, 2003

Validation of a Hybrid Approach for Imputing Missing Data

Colleen M. Ennett[1], Monique Frize[1,2]
[1]Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada
[2]School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada

Abstract—**A hybrid system has been constructed to impute missing values in a neonatal intensive care unit database using artificial neural networks and case-based reasoning. This paper presents the preliminary test results of a system using the connection weights of a linear neural network as the match weights in a case-based reasoner to find the closest-matching cases. The means of the ten closest-matching cases then replaced the missing values in the queries. The hybrid approaches were compared to mean and random imputations, and showed slightly better performance.**

Keywords—**Artificial neural networks, case-based reasoning, feature extraction, hybrid system, neonatal intensive care, missing values**

I. Introduction

The problem of missing values in databases is very common. In the past, the recommended way to deal with missing values was to avoid them. Missing data may be beyond the control of those collecting the data or those using the data for analysis. Current methods of working with missing values are unsatisfactory. Either valuable information is discarded or values based on the entire data set are imputed, which may be inappropriate. A new approach that attempts to incorporate specific information about the characteristics of the cases with missing values is essential to impute relevant values.

In the past decade, an increasing interest in the use of artificial neural networks (ANNs) in place of statistical approaches has developed. An ANN is a parallel, distributed, adaptive system that processes information in order to learn patterns as a result of exposure to a set of representative training cases [1]. The main advantage of neural networks over statistical techniques is that the model does not have to be explicitly defined before beginning the experiments to develop the model. ANNs can recognize the relevant data and patterns, whereas a statistical model requires prior knowledge of the relationships between the factors under investigation [2]. Also, with statistics, it is difficult to integrate data of different formats (i.e. working simultaneously with continuous, binary, ordinal and nominal data), but this can easily be achieved using ANNs.

Previous work by our research group [3] has shown the usefulness of ANNs for estimating: (1) resource use such as duration of mechanical ventilation or length of hospital stay [4,5], which is valuable for strategic planning; and (2) clinical outcomes such as mortality [6,7] to aid clinical decision-making. The previous barrier that discouraged many from using neural networks was their image as "black boxes." ANNs use complex algorithms for calculating the weights on each input and node. Our research group can extract the weights of the inputs and hidden nodes to assess the importance of each variable with respect to the outcome [8]. Weight-extraction is somewhat controversial, however, there is definite support in the field for this concept [9].

Our research group developed a case-based reasoner (CBR) to identify similar cases using an inference matching engine [10]. This system determines how closely related cases are by calculating the absolute "distance" between the cases in a multi-dimensional space (usually using a *k*-nearest neighbours algorithm); the smaller the distance is, the more similar the cases are. The CBR can match cases with missing data, however, it is only considering the variables that are present in each of the particular cases [11]. Therefore, by exploiting the CBR's matching ability and weighting only the variables present for the query, it is possible to determine which cases in the database most closely resemble the query, despite missing values.

II. Objective

In the past, the match weights in a CBR can be assigned uniform weights (all weights are equal), expert-chosen weights (i.e. weights assigned by a physician [10]) or discovered through trial and error. The objective of these experiments was to use the weights of a linear ANN as the matching weights of the CBR to find the closest-matching cases.

This work involved evaluating the ANN-CBR hybrid system's ability to impute missing values in an artificial database. This was necessary to validate the use of the linear ANN weights as the matching weights of the CBR to find the closest-matching cases. Removing known values to create an artificial database provided an opportunity to see how closely the hybrid system was able to estimate the missing values compared to what the actual values were. The hybrid system outputs were compared to two benchmark approaches: replacing with mean values and random values.

The database used to test this approach had 5102 complete patient cases from the Canadian Neonatal Network's neonatal intensive care unit (NICU) database with the input variables from SNAPPE-II (Score for Neonatal Acute Physiology, Version 2 with Perinatal Extension: lowest values for blood pressure, temperature, pO2/fiO2 ratio, serum pH and urine output, presence of seizures, birth weight, small for gestational age (SGA) and Apgar score at 5 minutes –measured within 12 hours of NICU admission) [12]. The outcome under investigation

0-7803-7789-3/03/$17.00 ©2003 IEEE

was in-hospital mortality. The SNAP-II variables are a subset (found via logistic regression) of the original SNAP score that contained 37 inputs [13]. This data set of nine inputs provided an adequate starting point to test the ANN-CBR hybrid system's ability to impute missing values. The SNAPPE-II score contains commonly collected data so a sufficient number of complete cases were available. As well, these variables were found through statistical analysis to be relevant for predicting neonatal mortality in the Canadian Neonatal Network's database [12].

Neonatologists use birth weight, SGA and Apgar score at 5 minutes like demographic variables to quickly assess to which risk category the infant may belong. To accommodate the clinical expertise of neonatologists and to compare with uniformly weighting the input variables, three different sets of CBR weights were tested:

1. ANN-extracted weights for all nine inputs (hereafter referred to as "ANN weights");
2. ANN-extracted weights for the six SNAP-II variables and weights = 100 for birth weight, small for gestational age status and Apgar score at 5 minutes ("clinician weights"); and
3. Uniformly weighting all nine input variables: all weights = 100 ("uniform weights").

III. Methodology

Creation of the Artificial Data Set. To evaluate how accurately the ANN-CBR hybrid technique imputed missing values compared to what the actual values were, it was necessary to create an artificial test set with missing data by deleting known values. The percentage of missing values (and which variables would have missing data) was based on the statistics of the entire database to determine a realistic percentage of missing data for each of the input variables. Due to the importance of birth weight, SGA and Apgar at 5 minutes (and mortality as the outcome), cases missing these variables were excluded from consideration. There were few or no missing values for seizure and lowest temperature, but the frequencies of missing values in the entire database for lowest blood pressure, pO2/fiO2 ratio, lowest urine output and lowest serum pH were 16%, 64%, 53% and 40%, respectively. The artificial data set was created by randomly selecting 16% of the database and deleting their values for lowest blood pressure. Then the database was reshuffled, and 64% of the database was selected so that the pO2/fiO2 ratio values for those cases could be deleted, and so on. This approach lead to the creation of an artificial missing values database that approximated the true distribution of missing values in the entire NICU database.

Weight-extraction from the ANN. To determine the appropriate weights for the CBR to identify matched cases, ANN experiments using only complete cases were performed. The feed forward backpropagation ANN with the hyperbolic tangent transfer function was trained (3358 cases) and tested (1744 cases) on independent datasets. All networks for this series of experiments were trained and tested with complete cases only (i.e. no missing data in the SNAPPE-II variables) since the neural network cannot deal with missing values. As well, it was necessary for the preliminary experiments to use only complete cases to test how close the imputed values were to the known values. A two-layer (linear) network was used with the weight-elimination cost function to prune the network and the logarithmic-sensitivity index as a stopping criterion that optimizes for both sensitivity and specificity [14,15].

From these preliminary experiments, the weights of the linear network were extracted and used as the weights for the CBR to determine the importance of each input variable for matching the cases to the query. The absolute values of these nine ANN weights were recalibrated to a maximum of 100 for use in the CBR.

Data Imputation Using the CBR. The CBR was used to impute the missing values using the ANN weights. For each case with missing data, the CBR identified the ten closest-matching cases from the remaining complete cases in its particular match set. Then the missing value in the query was replaced with the calculated mean from these closest-matching cases for the variable whose value was originally missing. Cases were matched using only input information to simulate a clinical setting where the patient's outcome would not be known.

Measures of Performance. To establish a benchmark for how accurately the ANN-CBR hybrid system imputed missing values, it was necessary to replace the missing values using mean and random statistics for comparison with the hybrid system. When the imputed values were compared to the known values in the complete data set, the percent error was calculated according to (1). Eq. (2) was used to calculate the average error. These errors were then compared between the true and imputed values in the original test set database for the three hybrid methods, mean and random approaches. The objective was to minimize the difference between the imputed and actual values.

$$\text{percent error} = \frac{(\text{imputed} - \text{known})}{\text{known}} * 100 \quad (1)$$

$$\text{average error} = \frac{\sum \text{percent error}}{\#\ \text{cases with missing values}} \quad (2)$$

Impact of Imputed Data. The next step was to test the impact of the imputed values on the neural network model's performance. Each "complete" set was divided into the same two-thirds training and one-third test sets, so their classification performance could be adequately observed. Each set had the same cases with missing values that were imputed according to the specific approach. The neural network classification performance of the initial experiments with the known/true values was then compared to the

TABLE 1
RECALIBRATED MEAN CONNECTION WEIGHTS

Input variable	Variable name	Recalibrated CBR weights
Lowest pO2/fiO2 ratio	*po2fio2r*	100
Lowest urine output	*lurine*	65
Apgar score at 5 minutes	*apgar5*	43
Lowest temperature	*ltempf*	42
Small for gestational age (SGA) status	*sga*	32
Lowest serum pH	*lserum*	28
Birth weight	*bthwt*	19
Lowest blood pressure	*lbloodp*	19
Presence of multiple seizures	*seizure*	19

TABLE 2
AVERAGE VALUES FOR TEST SETS WITH IMPUTED VALUES IN ARTIFICIAL MISSING VALUE DATABASE

Imputation approach	*lbloodp*	*lserum*	*lurine*	*po2fio2r*
True average	32.41±8.08	7.30±0.12	0.85±0.75	1.83±1.30
Hybrid: ANN	32.55±5.40	7.30±0.05	0.85±0.29	1.88±0.53
Hybrid: Clinician	32.71±5.57	7.30±0.05	0.84±0.29	1.88±0.50
Hybrid: Uniform	32.69±5.50	7.30±0.05	0.84±0.30	1.89±0.55
Mean	33.11±0.00	7.30±0.00	0.84±0.00	1.85±0.00
Random	33.77±7.97	7.30±0.11	1.14±0.67	2.11±1.22
Percent missing	16%	40%	53%	64%

TABLE 3
AVERAGE ERROR FOR IMPUTED VALUES IN ARTIFICIAL MISSING VALUES DATABASE

Imputation approach	*lbloodp*	*lserum*	*lurine*	*po2fio2r*
Hybrid: ANN	15.73%	1.14%	174.23%	94.62%
Hybrid: Clinician	15.97%	1.16%	172.21%	97.66%
Hybrid: Uniform	15.78%	1.12%	176.62%	93.40%
Mean	22.60%	1.16%	191.86%	107.07%
Random	30.45%	1.72%	298.56%	146.02%

hybrid-imputed data sets, and the mean and randomly imputed data sets.

IV. RESULTS

Overall, the ANN weights (shown in Table 1) tended to be similar to values chosen by clinical intuition. The weights for birth weight, SGA and Apgar score at 5 minutes did not come out as the most important, which was surprising because of their perceived significance by clinicians.

The average values for the variables with imputed data are shown in Table 2; this compares the true average values with the three hybrids, mean and random approaches. Table 3 shows the average error associated with the imputed values for each variable with missing data for the same imputation approaches. The mean errors for the hybrid-imputed values are approximately the same, whereas there is often greater error with the mean and randomly imputed values. Note also that the lowest error occurs for imputed values of *lbloodp* and *lserum*. These two variables also had the smallest number of missing values, although *lserum* was missing 40% of its data.

TABLE 4
TEST SET RESULTS OF LINEAR ANN EXPERIMENTS WITH IMPUTED DATA (N = 5102); TRAINING STOPPED AT EPOCH 846

	True	ANN	Clinician	Uniform	Mean	Random
Sens (%)	25.7	15.1	14.5	14.5	15.6	16.2
Spec (%)	97.6	98.2	98.2	98.3	98.1	98.1
CR (%)	90.2	89.7	89.6	89.7	89.7	89.7
ROC	0.8290	0.7372	0.7353	0.7404	0.7280	0.7292

TABLE 5
TEST SET RESULTS OF LINEAR ANN EXPERIMENTS WITH IMPUTED DATA (N = 5102); BEST TEST SET PERFORMANCE

	True	ANN	Clinician	Uniform	Mean	Random
Sens (%)	25.7	24.0	24.0	23.5	22.3	22.9
Spec (%)	97.6	96.8	96.8	97.0	97.4	96.9
CR (%)	90.2	89.3	89.3	89.0	89.7	89.3
ROC	0.8290	0.7786	0.7697	0.7814	0.7815	0.7701

Another factor affecting the model's ability to accurately impute a value may be the number of missing values the case had. Since the CBR was only matching on complete data, fewer values offer less information for the CBR to assess how closely matched a case is to the query. There are two factors that influence the percent error: (1) the number of missing values in the case; and (2) the number of cases missing values for that particular variable. In general, the more missing values a case has, the less information the CBR has to find the close-matching cases, so there is a greater possibility for imputation error. As well, the more cases missing values for a particular variable, the greater the room for error. Overall, the percent errors were similar amongst the three hybrid weight combinations (table not shown here) and it is difficult to identify one approach as being significantly better than the others, at least in this particular test and data type.

The ANN experiments were executed to observe any changes in classification performance with the imputed data. Two-layer (linear) networks were used for a direct comparison with the optimized networks that were used to extract the weights for CBR matching. The results of those experiments are displayed in Table 4 and are measured by the sensitivity, specificity, classification rate (CR) and area under the ROC curve.

When the neural network was stopped at the same point that the best test set classification was found using the true values, there was an equivalent drop in classification performance for all of the hybrid imputation approaches, as well as the mean and random imputation (approximately a 10% drop in sensitivity and a drop in the area under the ROC curve from approximately 0.83 to 0.73). If the ANNs were allowed to train to find the best test set classification rate for the imputed data, similar classification performance as with the known data was achieved, as shown in Table 5. The sensitivity for the hybrid approaches was slightly higher than that of the mean and random approaches. Again, none of the techniques clearly outperformed the others at this point, although the hybrid system's imputation techniques performed slightly better than the mean and random approaches.

These experiments involved a relatively homogenous population of patients. Patients with complete data for the SNAPPE-II variables (N =5102) have a mortality rate of 9.5% (N = 484 deaths) whereas the remaining 14325 cases (which have at least one missing value among the SNAPPE-II variables) have a mortality rate of 1.7% (N = 243 deaths). Clearly, these are two very different patient populations. Although the mean and random imputation approaches appeared to work quite well with the SNAPPE-II complete case patients, it was suspected that when the less ill patients were included in the database, their classification performance would degrade. This degradation was not anticipated for the hybrid imputation approach because the missing values were gradually imputed and were based on the mean of the ten closest-matching cases which was expected to be better than imputing the mean of the entire database or simply a random value.

V. Discussion

We used the extracted weights from the ANN as the weights of a matching algorithm in a case-based reasoner to find the closest-matching cases. From this set of matched cases, the statistical mean of the variable missing in the query was calculated and replaced the query's missing value. Since there were no complete cases for the original SNAP score (37 inputs) because no patient would have all tests and parameters recorded, this new approach for imputing missing values is quite valuable. Starting from the complete SNAPPE-II dataset (9 inputs), the database can be expanded either "vertically" or "horizontally."

Vertical Expansion suggests imputing missing values for the SNAPPE-II variables into the remaining incomplete cases to broaden the spectrum of infants by including some less sick babies in the database. We know that the fewer missing values that an infant has, the sicker the infant is [7]. The imputed values would be drawn from the sickest population and may not reflect the actual value for a baby that was less ill. To overcome this, the first imputed cases would be missing only one of the six SNAP-II variables, assuming that an infant only missing one variable would be similarly ill to a baby not missing any. Then groups of these infants would be added to the match-database to gradually lean the mean of the matched cases away from the bias of the sickest babies. Similarly, the next set would be the infants missing two values, etc.

Horizontal Expansion means working with infants with similar severity of illness, so the true values for their unrecorded data may be similar. The objective is to expand the NICU dataset to include more of the 37 SNAP variables for the 5102 cases, then the neural network with weight-elimination can determine the most important variables for predicting mortality and find the minimum number of variables that predicts mortality (minimum data set). We are interested to see if the ANN model returns the same six input variables as the SNAP-II variables or if the ANN minimum data set contains a different set of variables.

VI. Conclusion

We have validated that the variable weights for the linear ANN are consistent and reasonable. This finding can strengthen the confidence in neural networks as a decision support tool for strategic and clinical decision-making. The weights from the ANN were useful for replacing missing values using a matching algorithm in a case-based reasoner, so that an appropriate minimum data set can be developed to estimate a variety of medical outcomes. The hybrid system imputation approaches performed slightly better than mean and random imputation of missing values.

References

[1] W. Penny and D. Frost, "Neural networks in clinical medicine," *Med Decis Making* 1996;16:386-398.

[2] A. Blum, *Neural Networks in C++*, John Wiley & Sons, Inc.: New York, 1992.

[3] Medical IDEAS Research Group (Intelligent Decision Aid Systems), http://www.sce.carleton.ca/faculty/frize/MIRG_2001/mirg.html

[4] M. Frize, C.M. Ennett, M. Stevenson and H.C.E. Trigg, "Clinical decision-support systems for intensive care units using artificial neural networks," *Med Eng Phys*, 2001 Apr;23(3):217-225.

[5] Y. Tong, M. Frize and R. Walker, "Extending ventilation duration estimations approach from adult to neonatal intensive care patients using artificial neural networks," *IEEE Trans Info Technol Biomed*,2002 Jun;6(2):188-191.

[6] C. M. Ennett and M. Frize, "Weight-elimination neural networks applied to coronary surgery mortality prediction," *IEEE Trans Info Technol Biomed*, In press.

[7] C. M. Ennett, M. Frize and C. R. Walker, "Influence of missing values on artificial neural network performance," *Medinfo* 2001;10(Pt 1):449-53.

[8] M. Frize and C. M. Ennett, "Improving the potential clinical significance of decision-support systems using artificial neural networks," *Proc AMIA Symp* 2000:1011.

[9] T. Masters, *Advanced Algorithms for Neural Networks: A C++ Sourcebook*, John Wiley and Sons, Inc.: New York, NY, 1995.

[10] M. Frize and R. Walker, "Clinical decision-support systems for intensive care units using case-based reasoning," *Med Eng Phys* 2000 Nov;22(9):671-677.

[11] B. U. Haque, R. A. Belecheanu, R. J. Barson, K. S. Pawar, "Toward the application of case based reasoning to decision-making in concurrent product development (concurrent engineering)," *Knowledge-Based Systems* 2000; 13(2-3):101-112.

[12] 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," *J Pediatr* 2001;138:92-100.

[13] D. K. Richardson, J. E. Gray, M. C. McCormick, K. Workmann and D. A. Goldmann, "Score for neonatal acute physiology: a physiologic severity index for neonatal intensive care," *Pediatrics* 1993 Mar;91(3):617-23.

[14] C. M. Ennett, M. Frize and N. Scales, "Logarithmic sensitivity index as a stopping criterion for neural networks," *Proc IEEE-EMBS/BMES* 2002.

[15] C. M. Ennett, M. Frize and N. Scales, "Evaluation of the logarithmic-sensitivity index as a neural network stopping criterion for rare outcomes," *Proc IEEE-ITAB* 2003.