Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Predicting Probability of Mortality in the Neonatal Intensive Care Unit

Dajie Zhou, Monique Frize, *Senior Member, IEEE*

***Abstract*—Artificial neural networks can be trained to predict outcomes in a neonatal intensive care unit (NICU). This paper expands on past research and shows that neural networks trained by the maximum likelihood estimation criterion will approximate the 'a posteriori probability' of NICU mortality. A gradient ascent method for the weight update of three-layer feed-forward neural networks was derived. The neural networks were trained on NICU data and the results were evaluated by performance measurement techniques, such as the Receiver Operating Characteristic Curve and the Hosmer-Lemeshow test. The resulting models applied as mortality prognostic screening tools are presented.**

I. Introduction & Background

NEONATAL mortality is defined by the World Health Organization as infant death occurring during 28 days after birth [1]. The estimation of the neonatal mortality risk in a neonatal intensive care unit (NICU) has highly correlated with provision of aggressive patient care and management of medical resources [2].

For this purpose, in addition to NICU scoring systems [3,4], artificial neural networks (ANNs) have been studied as clinical decision support tools for predicting mortality risk. An ANN can be trained using medical data to estimate clinical outcomes. Previous studies by the members of the Medical Information technology Research Group (MIRG) showed that feed-forward neural networks were successfully used to predict NICU mortality [5]. The performance of neural network mortality estimation was comparable to or outperformed the results of the 'SNAP' (Score for Neonatal Acute Physiology) scoring system or fuzzy-logic classifiers [6,7], etc. A new set of 13 variables was developed through ANN learning as the important predictors of neonatal mortality [6].

In MIRG's previous work, ANN-based mortality prediction tools could classify patients with a dichotomous outcome (i.e. death or survival). The results categorized the patients into high and low risk populations. However, the tools could not discriminate between the severity of illness within the same population. Incorporating information to stratify patients into more groups according to mortality risk will help to allocate medical resources and improve neonatal care. Probability, as a classical decision support tool, provides the numerical comparison of clinical outcomes. The probability of mortality represents the level of uncertainty of death, which is an indicator for the severity of illness. An ANN model with probability estimation would be of great interest and flexibility to decision makers to overview the population based on detailed levels of illness severity. In addition, a probability model can provide the flexibility to develop a good mortality-screening tool by assigning appropriate cutoff points within the range of predicted probabilities. The research addressed in this paper is to supplement MIRG's previous studies on using ANNs to predict NICU mortality. The paper includes the following: 1) introduction of an approach to train neural networks to approximate the probability of NICU mortality; 2) assessment of ANN probability estimation models by performance measurement tools (e.g. the area under the receiver operating characteristic curves (ROC), Hosmer-Lemeshow (H-L) goodness-of-fit test); 3) validation of the probability estimation models as a screening tool for babies who are expected to die.

A neural network trained by the mean square error or the minimum cross entropy criteria, when converging, can estimate the 'a posteriori probability' of class membership [8]. This property demonstrates the general conditions of a neural network classifier to predict the probabilistic feature of clinical outcomes. The method for training neural networks to predict the probability of mortality was based on the maximum likelihood approach. To illustrate this method, we first recall it from probability theory [9].

Suppose $X_1, X_2, \ldots, X_n$ are random samples from a distribution that depends on a set of parameters θ with the probability mass function (*p.m.f.*) or the probability density function (*p.d.f.*) denoted by $f(x; \theta)$; θ is restricted to a given parameter space Ω. The joint *p.m.f.* or *p.d.f.* of $X_1, X_2, \ldots, X_n$, is $L(\theta)= f(X_1,X_2 \ldots X_n)=\prod f(X_i; \theta)$, where $i=1 \ldots n$; $L(\theta)$ is called the likelihood function and the random samples $X_1, X_2, \ldots, X_n$ can take values from the observed data $x_1, x_2, \ldots, x_n$ respectively in a training set (supposing that there are n samples in the training set). The estimation of the function $f(X; \theta)$ depends on identifying θ. The reasonable estimator of the target θ is a θ' that maximizes the likelihood function $L(\theta)$ and results in maximizing the joint probability of getting the observed data. When $\theta= \theta'$, $L(\theta)$ will reach the maximum value. θ' is the maximum likelihood estimator of the parameter θ.

The maximum likelihood estimation is an approach of parametric estimation, where the function form of *p.d.f.* or

Manuscript received April 3, 2006.

Dajie Zhou is a master's student in the Program of Systems Science, University of Ottawa, 800 King Edward, Ottawa ,Ontario, Canada, K1N 6N5 (e-mail: dzhou@site.uottawa.ca).

Monique Frize is a professor with the Department of Systems and Computer Engineering, Carleton University and the School of Information Technology and Engineering, University of Ottawa, 1125 Colonel By Drive, Ottawa, Ontario, Canada, K1S 5B5 (e-mail: mfrize@connect.carleton.ca).

p.m.f. has been known. The purpose of parametric estimation is to approximate the true distribution by adjusting a set of parameters on a pre-defined function form (e.g. Gaussian or exponential distribution). On the other hand, the neural network approximation is in general a non-parametric estimation, where the function form of the *p.m.f.* or *p.d.f.* cannot be identified. However, we know that a three-layer neural network (with one hidden layer) can approximate any bounded continuous functions of the inputs when the number of hidden units is sufficiently large. Therefore, it is reasonable to apply the maximum likelihood estimation on an ANN to approximate a probability function by changing the network parameters such as the number of hidden neurons, weights and bias. In this case, the neural network learning based on the maximum likelihood estimation is to identify the set of network parameters that result in the approximation of a NICU mortality probability function.

ANN training has several stopping criteria that can be applied. In the previous ANN models developed by MIRG researchers, the maximum logarithmic sensitivity index was used [10]. To be consistent with the previous research, the logarithmic sensitivity index, defined in (1) was still adopted, although a maximum log-likelihood criterion could be employed. The logarithmic-sensitivity index attempts to achieve optimal sensitivity and specificity of a classifier while slightly favoring higher sensitivity [11].

$$log - sensitivity - index = - sensitivity^{n} * log10\,(1 - sensitivity * specificity) \quad (1)$$

II. METHODOLOGY

A. Database and variables

The patient data used in the research are from the Canadian Neonatal Network (CNN) database. The CNN data were collected by the Canadian Neonatal Network from January 8, 1996 to October 31, 1997. The database contains patient information from seventeen NICUs across Canada during the two-year period. The original database contained missing values in patient physiologic variables. Since an ANN cannot process missing values, the incomplete values were imputed [6]. These artificially imputed values could approximate the true clinical data and allow development of ANN mortality prediction models.

The original database contained 20488 patient records. In our experiments, only the admission data (Day 1), collected within the first 12 hours, were used. Moribund infants and missing data about mortality, birth weight, small for gestational age status and Apgar score at five minutes were deleted from the original database. Any patient records with more than five missing values were also excluded. After imputing the missing values, the new CNN database contained 19427 patient records with the mortality ratio of 3.74% (727 deaths). These records were randomly divided into a training set (2/3) and a test set (1/3).

Thirteen variables, a set of mortality risk indicators developed by Ennett in MIRG's previous research [6], were employed as neural network inputs. These variables include *lowest pO2/FiO2 ratio*, *lowest urine output*, *lowest serum pH*, *Apgar score at five minutes*, *lowest platelet count*, *small for gestational age status*, *highest sodium*, *highest respiratory rate*, *highest pCO2*, *birth weight*, *lowest glucose*, *lowest temperature*, and *highest mean blood pressure*. For the purpose of neural network training and testing, all the data on these variables were normalized. The death and survival outcomes in our experiments were coded as 1, 0 respectively.

B. The likelihood function and weight update formula

To train the neural network on the maximum likelihood estimation, we first needed to model the ANN cost function as the likelihood function with the mortality outcomes of the NICU patients. Some researchers discussed the modeling method for different scenarios [12][13]. Here, we summarize them according to the scenario for our problem. Suppose that the training data D composed of NICU patient records $\{<x_1, y_1>, <x_2, y_2>, \ldots <x_m, y_m>\}$. x_m represents the values of a set of input variables for the m_{th} patient and y_m corresponds to an outcome, which takes values of 1 or 0 (i.e. death or survival). We can rewrite the likelihood function in the following form:

$L(\theta)=f(<x_1, y_1>, <x_2, y_2>, \ldots <x_m, y_m>;\theta)= \prod f(x_i, y_i|\theta)$, where $f(x_i, y_i|\theta)$ denotes the probability of y_i, given the input as x_i in the i_{th} patient record.

The conditional probability that $y_i=1$ given x_i will be denoted as $P(y_i=1|x_i)$. This value can be given by the neural network output $O(x_i)$. It follows that $P(y_i=0|x_i)=1-O(x_i)$. We further rewrite the $f(<xi, yi>|\theta)$ in the likelihood function as

$$f(x_i, y_i|\theta) = O(x_i)^{yi}[1-O(x_i)]^{(1-yi)} \quad (2)$$

Since we assume the observations of patient records are independent, the likelihood function on the training data with m records is obtained as

$$\prod O(x_i)^{yi}[1-O(x_i)]^{(1-yi)} \quad (3)$$

We write the equation (3) in a log-likelihood form:

$$L(\theta)=ln[L(\theta)]=\sum y_i * lnO(x_i)+(1-y_i)*ln(1-O(x_i)) \quad (4)$$

Equation (4) is the logarithmic form of the likelihood function that is also the cost function of the neural network. The neural network training is a process to find a set of parameters that maximize the likelihood function. Traditional neural networks based on the gradient decent search on the error surface seek a minimum of the global error. In our case, on the contrary, the global maximum of the likelihood was the aim of the neural network learning. Therefore, a new weight update technique with the gradient ascent search for a three layer neural network with a single binary output unit was developed.

Equations (5) [13] and (6) updates the weights of a hidden to output layer and an input to hidden layer, respectively.

$$\Delta W_{kj} = \eta \sum_{n=1}^{m} (d_n - O(x_n)) x_{nkj}; \quad (5)$$

$$\Delta W_{ji} = \eta \sum_{n=1}^{m} (d_n - O(X_n)) * W_{kj} * \varphi'(V_j(n)) * Y_i(n); \quad (6)$$

X_{nkj} is the synapse input from hidden neuron j to output neuron k when processing the n_{th} input pattern;

$O(x_n)$ is the neural network output for the n_{th} pattern, the

predicted probability;
d_n is the desired outcome for the n_{th} pattern; for the case of death, the value is 1, otherwise, the value is 0;
$Y_i(n)$: the value of the n_{th} pattern on the input neuron i;
W_{ji}: synapse weight from input neuron i to hidden neuron j;
$V_j(n)$: $\sum W_{ji} Y_i(n)$ the sum of all weighted inputs of the n_{th} pattern from the input layer;
$\varphi'(*)$: the fist order derivative of the sigmoid activation function;
W_{kj}: synapse weight from hidden neuron j to output neuron k ($k=1$ in our experiments);
η is the learning rate.

C. ANN development and model evaluation

The three-layer feed-forward neural networks with the hidden neuron number from 2 to $2n+1$ nodes were trained; n is the number of input variables, which was 13 in the experiments. The maximum number of hidden units was defined according to the Kolmogorov superposition theory in an ANN context. It suggests that the number of hidden units should be no more than two times the number of inputs units plus one [14]. Therefore, 26 neural network models with different number of hidden nodes were developed. The activation function of the neurons was the sigmoid function in order to guarantee the output ranging from 0 to 1.

Discrimination and calibration were both measured on each model. Discrimination refers to the ability of a model to classify death and survival. A model with perfect discriminating ability assigns higher mortality probabilities to the infants who are going to die than the infants who are going to survive. The discriminating ability was assessed by the area under ROC. The area under ROC value of one means that the model perfectly classifies the death and the survival populations, whereas a value of 0.5 corresponds to a random guess. The calibration of a model represents how accurately the prediction of a model is over the entire range of mortality risk. It indicates if there is good correspondence between the observed and expected number of outcomes over all probability levels. For example, if the probability for a death outcome is 0.6 as calculated by a model, 60 deaths will be likely observed within a set of 100 samples of this type of patients. The calibration was evaluated by the H-L goodness-of-fit test [12]. The H-L test measures the calibration by calculating the Pearson Chi-square statistics across the number of patients grouped by the estimated probability of mortality. The sum of these resulting statistics follows a Chi-square distribution, with the degrees of freedom of the number of groups (usually 10 in most studies) minus 2. A P-value > 0.05 implies that there is no significant difference between a predicted distribution and a true distribution. That is, the goodness of fit is acceptable and the model is able to approximate the probability for mortality outcomes.

D. Cutoff point adjustment for screening outcomes

The cutoff point during the development of an ANN model was 0.5. Hence, an ANN output>0.5 would be classified as a death whereas an output<0.5 would be regarded as a survival case. The cutoff point was applied in order to calculate the sensitivity and the specificity in the logarithmic-sensitivity index equation (1) during training and testing phases. Since the training data was highly imbalanced (i.e. the mortality cases were much less than the survival ones), a neural network would tend to learn more from the survival cases than from the ones who died. Low sensitivity and high specificity could be obtained. Moreover, the initial purpose of these probability estimation models was to estimate the probability of mortality from group distributions rather than to make a direct statement on a specific outcome. Therefore, sensitivity and the specificity measures may not represent the capability of a model's probability estimation. This was validated in some experiments that ANN models with high test sensitivity fit poorly on the observed data. However, to some extent, we need a probability model to be a tool for mortality prognostic screening (e.g. classifying survival and death cases directly). Ideally, to ensure confidence of using a model for such a scenario, a specificity value of close to 100% and a sensitivity value of at least 50% would be acceptable by users [6]. For the development of a mortality-screening tool without affecting its probability estimation, we moved the value of the cutoff point between 0 and 0.5 to compare the resulting sensitivity and specificity.

III. Results

With the aforementioned set of 13 input variables, 26 ANN models with hidden units' number from 2 to 27 were trained. The performance of every model was assessed by the following measures: sensitivity(Sen.), specificity(Spec.), area under the ROC, Hosmer-Lemeshow Chi-square statistics (C-value), degree of freedom(Df) and P-value of the H-L test. The results of the ANN experiments based on the different numbers of hidden neurons are presented in Table 1. Table 1 only lists the models with good calibration results (i.e. P-value>0.05 in both training and test sets). Therefore, ANN models with 4, 5, and 15 hidden units are listed and the one with 15 hidden units fit the observed data best. Table 2 is the contingency table of the H-L test using the model with 15 hidden neurons in all patients' records.

Different cutoff points were selected. Table 3 is an example of the model in table 2 for comparing the resulting sensitivity and specificity after changing the cutoff point for classifying two outcomes. When the cutoff point is properly set, in addition to the probability prediction, the model can also be used as a mortality-risk-screening tool.

TABLE 1
PERFORMANCE MEASURES OF BOTH TRAINING AND TEST SETS

	4 Hidden Units	5 Hidden Units	15 Hidden Units
Train Sen.	0.297	0.282	0.283
Train Spec.	0.995	0.995	0.996
Test Sen.	0.257	0.248	0.252
Test Spec.	0.993	0.993	0.993
Train ROC	0.898	0.899	0.901

Test ROC	0.882	0.882	0.876
Train C-value	7.29	4.84	4.35
Test C-value	12.29	11.84	7.77
Train Df	8	8	8
Test Df	7	7	7
Train P-value	0.51	0.77	0.82
Test P-value	0.09	0.11	0.35

TABLE 2
H-L CONTINGENCY TABLE ON THE MODEL WITH 15 HIDDEN NEURONS IN ALL 19427 RECORDS

Probability Group	Total Records	Observed Death	Expected Death
<0.1	17746	219	219.2
0.1-0.2	716	106	101.1
0.2-0.3	317	74	77.3
0.3-0.4	192	71	67.3
0.4-0.5	150	57	66.9
0.5-0.6	114	60	61.9
0.6-0.7	75	46	48.2
0.7-0.8	57	41	42.7
0.8-0.9	41	35	34.4
>0.9	19	18	17.8

H-L C-value=4.27, Df=8, P-value=0.83

TABLE 3
ADJUSTMENT OF CUTOFF POINT IN THE MODEL OF TALBE 2

cutoff	0.5	0.4	0.35	0.25	0.15
Sen. %	27.5	35.4	41.1	50.4	61.9
Spec. %	99.4	98.9	98.6	97.7	95.8
CCR %	96.7	96.6	96.5	96	94.5

CCR is the correct classification rate.

IV. CONCLUSION

We presented a method for developing ANN models to estimate probability of mortality in infants admitted to the NICU using 13 variables readily obtained on admission of the infants. The three-layer ANNs trained on the maximum likelihood criterion and the weight update formulas via gradient ascent were able to provide an estimation of the probability of mortality. The resulting models were assessed and validated. A probability model is useful for quantitating the severity of illness in the discussions between patients' families and healthcare service providers [15]. These models used alone or in concert with the diagnosis of clinicians, may be useful in planning resource utilization and review in NICUs.

The cutoff point can be used to separate the groups of dichotomous outcomes. After setting the cutoff point appropriately, the probabilistic models are able to classify the death and survival of babies with high sensitivity and specificity, thus providing a mortality-screening tool.

V. FUTURE WORK

The method described in this paper can potentially be used for estimating many other NICU outcomes. Future work will add probability predictions to ventilation duration and to the occurrence of complications such as neuro-image abnormality, necrotizing entero-colitis, and broncho-pulmonary dysplasia. In addition, we will apply the maximum log-likelihood stopping criterion to the ANN training.

REFERENCES

[1] Health Canada. "Perinatal health indicators for Canada: A resource manual." Minister of Public Works and Government Services Canada, 2000.http://www.hc-sc.gc.ca/hpb/lcdc/brch/reprod/phic-ispc/pdf/indperie.pdf

[2] Stevens SM, Richardson DK, Gray JE, Goldmann DA, McCormick MC, "Estimating neonatal mortality risk: an analysis of clinicians' judgments, " *Pediatrics*, June 1994; 93(6 Pt 1):945-50

[3] Gray JE, Richardson DK, McCormick MC, Workman-Daniels K, Goldmann DA, "Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index, " *Pediatrics*, Oct 1992;90(4):561-7.

[4] Richardson DK, Corcoran JD, Escobar GJ, Lee SK, "SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. " *Pediatrics*, Jan 2001;138(1):92-100

[5] Walker CR, Ennett CM and Frize M, "Use of an Artificial Neural Network to Estimate Probability of Mortality and Duration of Ventilation in Neonatal Intensive Care Patients." in *Proc. Medinfo* 2001: 10(Pt 1):584.

[6] Ennett CM, "Imputation of missing values by integrating artificial neural networks and case-based reasoning, PhD thesis, Carleton University, Ottawa Ontario, Canada, 2003

[7] Frize M, Ibrahim D, Seker H, Walker RC, Odetayo MO, Petrovic D and Naguib RNG, "Predicting clinical outcomes for newborns using two artificial intelligence approaches, " in *Proc. IEEE EMBS-BMES* 2004, pp.3202-3205

[8] Bishop CM, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995

[9] Hogg RV, Tanis EA, *Probability and Statistical Inference*, Pearson Education, 2000

[10] Ennett CM, Frize M, Scales N, "Logarithmic-sensitivity index as a stopping criterion for neural networks, "in *Proc IEEE EMBS-BMES* 2002.

[11] Ennett CM, Frize M, Scales N, "Evaluation of the logarithmic-sensitivity index as a neural network stopping criterion for rare outcomes, " in *Proc. IEEE-ITAB*, 2003, pp. 207-210.

[12] Hosmer, DW, Lemeshow, S. *Applied Logistic Regression (2nd Edition)*, New York: Wiley. Long, JS 1997

[13] Mitchell, Tom M, *Machine Learning*, New York: McGraw-Hill, c1997

[14] Beiu V, "A novel highly reliable low-power nano architecture when von Neumann augments Kologorov," in *Proc. IEEE Application-Specific Systems, Architectures and Processors*, 2004, pp. 167-177.

[15] Groeger JS et al., "Probability of mortality of critically ill cancer patients at 72 h of intensive care unit (ICU) management Support Care Cancer, "Nov 2003, pp. 686-695.