Classification and Prediction Of A Health Status Of HIV/AIDS Patients: Artificial Neural Network Model

Chang W. Lee[a] and N. K. Kwak[b]

[a] *Chinju National University*
150 Chilam Dong, Chinju 660-758, Korea
Tel: +82-55-751-3454, Fax: +82-55-751-3459, E-mail:cwlee@chinju.ac.kr

[b] *Saint Louis University*
3674 Lindell Blvd, St. Louis, MO 63108, USA
Tel: +11-314-977-3878, Fax: +11-314-977-1483, E-mail: kwakn@slu.edu

Abstract

Artificial neural network (ANN) is known to identify relationships even when some of the input data are very complex, ill-defined and ill-structured. One of the advantages in ANN is that it can discriminate the linearly inseparable data. This study presents an application of ANN to classify and predict the symptomatic status of HIV/AIDS patients. Even though ANN techniques have been applied to a variety of areas, this study has a substantial contribution to the HIV/AIDS care and prevention planning area. ANN model in classifying both the HIV and AIDS status of HIV/AIDS patients is developed and analyzed. The diagnostic accuracy of the ANN in classifying both the HIV status and AIDS status of HIV/AIDS status is evaluated. Several different ANN topologies are applied to AIDS Cost and Services Utilization Survey (ACSUS) datasets in order to demonstrate the model's capability. If ANN design models are different, it would be interesting to see what influence would have on classification of HIV/AIDS-related persons.

Keywords: Artificial Neural Network, Health Services Research, HIV/AIDS Status

Introduction

Traditional parametric models are based on the assumption that the observed attributes of an object have the multivariate normal distribution. Some situations where data set contains unordered categorical and/or skewed continuous independent variables may not be able to make this assumption. In such situations, alternate approaches have a potential to outperform traditional parametric models. A variety of alternate approaches, including the traditional non-parametric methods such as Kernel method [1] and k-nearest-neighbor method [2], non-traditional method such as artificial neural network (ANN) [3][4], and mathematical programming methods [5][6][7], have been suggested.

The use of ANN models in classification issues of business, health services research, and others has been generally limited to the adoption of factors with continuous or ratio variables, rather than the categorized values found in socioeconomic or demographic variables [8][9][10][11][12]. Moreover, appropriate utilization of ANN model to implement a large-scale dataset is one of the most difficult issues in the ANN application discipline.

One of the important problems in the health-care research area is that of classifying and predicting a patient into a group based on the observed attributes of that patient. Even though successful applications of ANN models to health care area, little attention has been made to the prevention and care planning in HIV/AIDS area. The characteristic of the factors is that most of them are either unordered categorical or continuous with a skewed distribution. Most health-care data, especially HIV/AIDS related data, has unequal cases and distribution with survival data. Thus, these parameter properties result in utilization of ANN models adopting non-parametric mechanism.

The purpose of this study is to apply an ANN to present a good discriminant between HIV and AIDS status. An ANN model is developed based on the publicly available HIV/AIDS data of AIDS Cost and Services Utilization Survey (ACSUS) data as input and output variables.

Literature Review

In the 1950s and early 1960s, the idea of ANN was once viewed as a theoretical foundation of building machine learning systems. It was proven to have many limitations. Recent ANN research has overcome some early limitations. One of the advanced features is the development of a back propagation algorithm (BPA) in a learning mechanism to train multi-layer networks. The BPA using the hidden layer allows the data to be classified.

Since the development of ANN, it has received considerable attention from researchers and has been applied to a variety of problems in classification and prediction. ANN has been applied successfully for development of non-parametric statistical models. More reliable outcome research has been explored in the field of pattern classification and pattern prediction. ANN model is able to recognize and predict an existing pattern in data classification in different categories [13][14][15]. ANN in health-care applications have been used for clinical diagnosis [16][17], HIV-structure analysis [18], HIV/AIDS functional health status [19], image analysis [20], prediction of cancer [21][22], prediction of length of stay [23][24][25], sequence analysis [26], and speech recognition [27].

ANN is known to be able to identify relationships even when some of the input data are either unordered categorical or continuous with a skewed distribution. One of the potentials of an ANN is that it can discriminate linearly inseparable data. Even though ANN techniques have been applied to a variety of areas in business, public sector, and health services research, this study makes a practical contribution to the HIV/AIDS care and prevention planning area. If the appropriate methodologies in various ANN design models are different, it would be interesting to see what impact this would have on the classification of HIV/AIDS-related persons.

Research Design

Data Collection

This study utilizes the ACSUS data set [28], which is a longitudinal study data set of persons with HIV/AIDS-related diseases. Information has been gathered on 1,949 HIV/AIDS-related patients in a series of interviews over a total of six time periods with a quarterly follow-up survey. After collecting information on demographic and functional health status, interviewers contact clinical and other medical services providers identified by the study subject twice during the six time periods to collect the relevant information.

Table 1. Summary of Independent Variables

Variables[a]	Description	Type
Demographic		
SEX	Sex of patient	Unordered
RACE	Race of patient	Unordered
Exposure Route:		
EXPOS	Exposure route	Unordered
Health Services Utilization:		
ADM	Total admissions	Continuous
IPNGT	Total inpatient nights	Continuous
AMBVS	Total ambulatory visits	Continuous
ERVS	Total ER visits	Continuous
HCVS	Total hospital visits	Continuous
MDVS	Total PP visits	Continuous

[a] VIF values < 1.700 for multicollinearity diagnosis

Originally, 20 variables for this study are selected from the source data set. Ten variables are selected after controlling them by the variance inflation factor (VIF) method to detect multicollinearity between variables. After filtering cases based on the responses of specific patients (i.e., cases are excluded if their responses are Don't know, Refused, or any other inappropriate ones), 1,171 of 1,264 patients are selected as valid cases. Nine independent variables included in the variable set are listed in Table 1.

Major motivations of ANN for this study are (1) that this approach establish the potential to classify objects without making any assumptions on the distributional characteristics of the groups, and (2) that with its separating groups and distance measures the geometric interpretation of this approach clearly has intuitive appeal. Thus, this study applies ANN to classify and predict the symptomatic status of HIV/AIDS patients with three demographic characteristics, exposure route, and six health services utilization factors included in the ACSUS data set.

Patterns of use of health services and changes in these factors over the course of the disease can be analyzed for HIV/AIDS-related patients receiving care. 1,171 subjects selected as valid cases imply that a person had either on HIV status or on AIDS status. Table 2 presents descriptive statistics of all variables.

Table 2. Descriptive Statistics of Variables (n=1,171)

Variable*	Mean	SD	MIN	MAX
Health	0.43	0.49	0	1
Sex	0.84	0.37	0	1
Race	1.78	0.81	1	3
EXPOS	1.87	0.64	1	4
ADM	1.61	2.94	0	78
IPNGT	17.80	31.13	0	235
AMVS	25.50	22.74	0	218
ERVS	2.02	2.83	0	37
HCVS	15.60	18.31	0	218
MDVS	6.75	11.65	0	78

ANN Modeling

The model developed to classify a current symptomatic status of HIV/AIDS-related patients has the three-layer back-propagation algorithm. An input layer is used to represent a set of input variables. An output layer is used to represent the output variable. The hidden layer has an arbitrary number of hidden nodes. Thus, the numbers of hidden nodes are chosen arbitrarily and they derive different possible results.

For the purpose of this study, categories in output variables have been recoded as follows: (1) for the first node in the output layer, the output is zero if the current symptomatic status is HIV, and one if the current symptomatic status is AIDS; and (2) for the second node in the output layer, the code is assigned in the reverse way. Since no prior information is available on how the layers should be connected in the three-layer network, all nodes in the two adjacent layers are fully connected to each other.

Input pattern has 15 nodes: Sex male and Sex female, Race-White, Race-Black, Race-Hispanic, Expos-homo, Expos-IDU, Expos-IV, Expos- hetero, ADM, IPNGT, AMVS, ERVS, HCVS, and MDVS. Each of these variables is entered into the corresponding input layer of the network.

The values are multiplied by computer-generated random numbers resulted in the input values of the hidden layer. Each value is placed in a logistic function that computes the net activation of the hidden layer, becoming input values of the output layer. This value is entered into the same logistic function that computes the activation of the output layer, resulting in the output values: HIV status or AIDS status. Thus, in practice, the output values could be considered as representing the likelihood of HIV status or AIDS status in the current symptomatic status of each HIV/AIDS patient.

The network architecture is designed to be a three-layer BPA networks. The BPA has a linear approximation function for the input layer and a logistic function for the output layer. After configuring the network, a learning rate, initial weight, and momentum learning epoch are assigned to the model to initiate the training. Since assigning a learning rate, momentum, and number of epoch is arbitrary, a certain value as a default for each is assigned to the model. Once the model is designed, a certain percent over total pattern is extracted for the training set and the rest become the holdout set. An epoch is considered completed after the network examines all of the input and output patterns for all the training sets. Epochs for training set are repeated 200 times as a learning rate. In order to avoid over-fitting the network, the learning process was stopped when the total number of epoch repeats reached 20,000. NeuroShell®2 was utilized to conduct this study [29].

Model Analysis And Discussion

In order to analyze the model result, a network topology must be selected. Since there is no formal way to select a network topology, some trial experiments are performed to show different possible outcomes under different network topologies. A holdout set of 145 patients is used to examine the performance of the ANN model. A variety of tests are performed to analyze the model. The training ends the number of events that the minimum holdout set error has exceeded 20,000, as specified in the mode design. Table 3 presents a summary of model statistics with respect to different numbers of hidden nodes (i.e., H = 3, H = 5, and H = 7) in hidden layer.

Table 3. ANN Model Statistics with Different Nodes Topology

	Training set (1,026 training patterns)				Holdout set [a] (145 training patterns)		
	Best TLE	Epoch	LAE	MAE	Event	LAE	MAE
H_3	60.2	58	0.30	0.27	200	0.28	0.28
H_5	79.6	77	0.28	0.27	200	0.30	0.28
H_7	53.2	51	0.29	0.27	200	0.32	0.28

[a] : holdout sent is extracted about 10% of training set.
TLE: training learning event
LAE (last average error)
MAE (minimum average error)

In the case where the hidden layer has four nodes (H_5), the LAE and MAE in training set have the lowest values of 0.277 and 0.268, respectively. All models with different hidden nodes show the different best holdout set event, (H_3 = 60,200, H_5 = 79,600, and H_7 = 53,200) and the epochs are also different with a minimum average error (H_3 = 17, H_5 = 7, and H_7 = 10). Holdout set has the lowest LAE, 0.279, at the network model having three hidden nodes (H_3), but all network models have the same MAE, 0.275.

Table 4 illustrates the relative contribution between input and output variables. In this table, the BPA network model with different hidden nodes presents the similar results. The model shows seven variables with the higher relative contributions with ascending order (rectangular box along with numbers). Among these high contribution variables, IPNGT is the most significant factor to classify between HIV and AIDS status. Different hidden nodes result in different relative contributions among input variables. R^2 values range from 0.119 to 0.0101 and MSE from 0.216 to 0.219.

Two aspects of the objective tests are important in this study: validity and reliability. These indices are usually determined by administrating the test to one group which has the HIV symptomatic status and to another group which has the AIDS group, and then comparing the results. Thus, HIV positive is defined as the percent of those who have an HIV status and are so predicted by the network test. AIDS positive is defined as the percent of those who have an AIDS status and are so predicted by the network model.

Table 4. Relative Impacts of Output Variables

BPA Network	H_3	H_5	H_7
RW	-	-	-
RB	-	$\boxed{2.51}^7$	-
RH	$\boxed{1.06}^7$	-	-
SexM	$\boxed{1.07}^6$	-	$\boxed{2.62}^7$
SexF	-	-	-
Expr-homo	-	$\boxed{2.79}^6$	-
Expr-IDU	-	-	$\boxed{2.94}^6$
Expr-IV	-	-	-
Expr-hetero	-	-	-
ADMT	$\boxed{2.55}^5$	$\boxed{4.06}^3$	$\boxed{3.15}^3$
IPNGTT	$\boxed{8.46}^1$	$\boxed{14.01}^1$	$\boxed{9.88}^1$
AMSVT	$\boxed{2.90}^3$	$\boxed{3.86}^5$	$\boxed{3.04}^4$
ERVST	$\boxed{3.66}^2$	$\boxed{6.33}^2$	$\boxed{4.66}^2$
MDVST	-	-	-
HCVST	$\boxed{2.76}^4$	$\boxed{3.88}^4$	$\boxed{3.01}^5$
R^2	0.108	0.119	0.106
MSE	0.219	0.216	0.219

Another two indices are used to evaluate the reliability of a test: HIV predictive probability and AIDS predictive probability. HIV predictive probability is the probability of the HIV status being actually present, given that a symptomatic status of HIV is predicted as HIV. AIDS predictive probability is the probability of the AIDS status being actually present if a symptomatic status of AIDS is predicted as AIDS.

Table 5 presents a summary of results of correct classification with respect to each network and the relevant analysis. As table 5 indicates, sensitivity for actual HIV status over ANN classification is somewhat high, but specificity is rather low. This means that the HIV status can be identified easier than AIDS status. Since the AIDS status is very time-dependent, it is very difficult to identify AIDS status through socioeconomic variables and/or simple clinical records. Hidden nodes H_5 has the highest AIDS positive value with the highest classification rate.

Table 5. Correct Classification by Neural Network Model

Neural Network Topology	HIV positive (n=667)	AIDS positive (n=504)	HIV Predictive Prob.	AIDS Predictive Prob.
$H_3^{a,b}$	88%	34%	64	68
$H_5^{a,c}$	84%	42%	66	66
$H_7^{b,c}$	83%	39%	65	64

a, b: significant at p value < 0.000 by the paired t-test
c: no significant

Conclusion

This study presents an application of ANN on the prevention and care planning of HIV/AIDS patients. An ANN model is developed and analyzed. The diagnostic accuracy of the ANN is examined. Three different neural network topologies are applied to AIDS Cost and Services Utilization Survey (ACSUS) data sets in order to demonstrate the neural network's capability.

This study focuses on demonstrating the use of ANN to classify the symptomatic status of HIV/AIDS patients and to evaluate the potential benefit of ANN in terms of the classification accuracy. ANN model presents factors that are relatively more important to the classification of HIV/AIDS patient status, provide decision-makers with more accurate information to implement, reinforce prevention and care planning for HIV/AIDS patient, and provide strategies to meet more appropriately health-care policy and regulations.

References

[1] Hand, D. J., *Discrimination and classification*, New York, NY: John Wiley, 1981.

[2] Hand, D. J., *Kernel discriminant analysis*, Chichester, NY: Research Studies Press, 1982

[3] Lee, C. W. and J. A. Park, Assessment of HIV/AIDS-related health performance using an artificial neural network, *Information and Management*, 38 (2001) 231-238.

[4] Patuwo, E., M. Y. Hu, and M. S. Hung, Two-group classification using neural networks, *Decision Sciences*, 24 (1993) 825-845.

[5] Erenguc, S. S. and G. J. Koehler, Survey of mathematical programming models and experimental results for linear discriminant analysis, *Managerial and Decision Economics*, 11 (1990) 215-225.

[6] Joachimsthaler, E. A. and A. Stam, Mathematical programming approaches for the classification problem in two-group discriminant analysis, *Multivariate Behavioral Research*, 25 (1990) 427-454.

[7] Stam, A., Nontraditional approaches to statistical classification: some perspectives on LP-norm methods, *Annals of Operations Research*, 74 (1997) 1-36.

[8] Ebell, M. H., Artificial neural network for predicting failure to survive following in-hospital cardiopulmonary resuscitation, *Journal of Family Practice*, 36 (1993) 297-303.

[9] Faraggi, D., and Simon, R., A Neural network model for survival data, *Statistics in Medicine*, 14 (1995) 73-82.

[10] Hart, A., Using neural networks for classification tasks-some experiments on datasets and practical advice, *Journal of the Operational Research Society*, 44 (1992) 1129-1145.

[11] Sharda, R., Neural networks for the MS/OR analyst: an application bibliography, *Interfaces*, 24 (1994) 116-130.

[12] Wilson, R. L., Ranking college football teams: a neural network approach, *Interfaces*, 25 (1995) 44-59.

[13] Archer, N. P., and Wang, S., Application of the back propagation neural network algorithm with monotonicity constraints for two-group classification problems, *Decision Sciences*, 24 (1993) 60-75.

[14] Patuwo, E., Hu, M. H., and Hung, M. S., Two-group classification using neural networks, *Decision Sciences*, 24 (1993) 825-845.

[15] Wang, S., The unpredictability of standard back propagation neural networks in classification applications, *Management Science*, 41 (1995) 555-559.

[16] Baxt, W. G., Use of an artificial neural network for the diagnosis of myocardial infarction, *Annals of Internal Medicine*, 115 (1991) 843-848.

[17] Dorffner, G., and Porenta, G., On using feedforward neural networks for clinical diagnostic tasks, *Artificial Intelligence in Medicine*, 6 (1994) 417-435.

[18] Andreassen, H., Bohr, H., Bohr, J., Brunak, S., Bugge, T., Cotterill, R. M., Jacobsen, C., Kusk, P., Lautrup, B., and Petersen, S. B., Analysis of the secondary structure of the human immunodeficiency virus (HIV) proteins p17, gp120, and gp41 by computer modeling based on neural network methods, *Journal of Acquired Immune Deficiency Syndromes*, 3 (1990) 615-622.

[19] Kwak, N. K., and Lee, C. W., A neural network application to classification of health status of HIV/AIDS patients, *Journal of Medical Systems*, 21 (1997) 87-97.

[20] Dawson, A. E., Austin, R. E. Jr., and Weinberg, D. S., Nuclear grading of breast carcinoma by image analysis: classification by multivariate and neural network analysis, *American Journal of Clinical*

Pathology, 95 (1991) S29-37.

[21] Floyd, C. E. Jr., Lo, J. Y., Yun , A. J., Sullivan, D. C., and Kornguth, P. J., Prediction of breast cancer malignancy using an artificial neural network, *Cancer*, 74 (1994) 2944-2948.

[22] Ravdin, P. M., and Clark, G. M., A practical application of neural network analysis for predicting outcome of individual breast cancer patients, *Breast Cancer Research & Treatment*, 22 (1992) 285-293.

[23] Davis, G. E., Lowell, W. E., and Davis, G. L., A neural network that predicts psychiatric length of stay, *MD Computing*, 10 (1993) 87-92.

[24] Lowell, W. E., and Davis, G. E., Predicting length of stay for psychiatric diagnosis-related groups using neural networks, *Journal of the American Medical Informatics Association*, 1 (1994) 459-466.

[25] Tu, J. V., A comparison of neural network and logistic regression models for predicting length of stay in the intensive care unit following ·cardiac surgery, Unpublished *Masters Thesis*, University of Toronto, Canada, 1993.

[26] Fu, L., Polygenic trait analysis by neural network learning, *Artificial Intelligence in Medicine*, 6 (1994) 51-65.

[27] Kurogi, S., Speech recognition by an artificial neural network using findings on the afferent auditory system, *Biological Cybernetics*, 64 (1991) 243-249.

[28] Berk, M. L., C. Maffeo, and C. L. Schur, "Research design and analysis objectives," *AIDS Cost and Services Utilization Survey (ACSUS) Reports*, No. 1, AHCPR Pub. No. 93-0019, Rockville, MD: Agency for Health Care Policy and Research, 1993.

[29] *NeuroShell®2*, Frederick, MD: Ward Systems Groups, Inc., 1993.