

Using Inverse Neural Networks for HIV Adaptive Control

B. Leke Betechuoh, T. Marwala and T. Tetley

School of Electrical and Information Engineering,
University of Witwatersrand, Johannesburg,
South Africa
{b.leke, t.marwala, t.tetley}@ee.wits.ac.za

Abstract: Neural Networks are used in this paper, in an inverse configuration, for the adaptive control of HIV status of individuals. In the paper, a control mechanism, to understand how demographic properties (in this case, educational level) affect the risk of being HIV positive, is implemented. Preliminary design showed inverse neural networks outperforms the other methodology. The moral behind this implementation is to understand whether HIV susceptibility can be controlled by modifying some of the demographic properties such as education. The proposed method is tested on the HIV data set. It is found that the proposed method is able to predict the educational level of individuals to an accuracy of 88% if the HIV status of individuals and other demographic properties are known. It is thus possible to understand how the educational level of individuals can be modified to control the proneness of individuals to HIV contraction.

Keywords: Multilayer perceptron, Feedforward neural networks, Inverse neural networks, Genetic algorithms, HIV, Adaptive control.

I. Introduction

Acquired Immunodeficiency Syndrome (AIDS) was first defined in 1982 to describe the first cases of unusual immune system failure that were identified in the previous year. The Human Immunodeficiency Virus (HIV) was later identified as the cause of AIDS. Epidemiology examines the role of host, agent and environment to explain the incidence and transmission of disease. Risk factor epidemiology examines the individual (demographic and social) characteristics of individuals and attempts to determine the factors that place an individual at risk of acquiring a disease¹. In this study, the demographic and social characteristics of the individuals and their behaviour are used to determine the risk of HIV infection; this is referred to as "biomedical individualism" [1, 2]. By identifying the individual risk factors that lead to the disease, it is possible to modify social conditions which give rise to these factors, and thus design effective HIV intervention policies [1]. The

model is then used to control the disease using adaptive control.

Adaptive control theory provides a possible way to solve many problems. Two distinct approaches—direct adaptive control and indirect adaptive control—can be used to control a plant adaptively. In the direct control, the parameters of the controller are directly adjusted to reduce some norm of the output error. In the indirect control, the parameters of the plant are estimated as the elements of a vector at any instant k , and the parameters vector of the controller is adapted based on the estimated plant vector.

At each sampling instant, the input and output of the generating unit are sampled and a plant model is obtained by some on-line identification algorithm to represent the dynamic behavior of the generating unit at that instant in time. The required control signal is computed based on the identified model. Various control techniques can be used to compute the control. All control algorithms assume that the identified model is the true mathematical description of the controlled system. In this case study, the control algorithm is artificial intelligence (AI). The controlled system model will also be created using neural networks (AI).

Artificial intelligence has been used successfully in medical informatics for decision making, clinical diagnosis, prognosis, and prediction of outcomes [3, 4]. It has been defined as a term that in its broadest sense will indicate the ability of a machine or artifact to perform the same kinds of functions that characterize the human thought [5]. Neural networks when applied to classification discover which characteristics or combinations of characteristics are useful for distinguishing between classes. Another objective of a pattern classification system is to find a separator that will divide the classes, placing as many samples into the correct classes as possible [6].

II. Background

A. Neural Networks

Neural computation is introduced as an intelligent system

relating the processing parameters to the process responses. Such a system is based on artificial neural network (ANN) which is an inter-connected structure of processing elements called neurons. The ANN structure consists of; the input pattern representing the processing parameters. In this application, the size of the input pattern is 9; the output pattern representing the HIV status; the hidden layers describing implicitly the correlations between the processing parameters and the output characteristics. The connection between a couple of neurons is described by a number called weight, translating the strength of the connection.

Three steps are required to optimize the ANN structure [7]. These are training, validation and testing steps. There are several types of neural network architectures and in this paper we focus on the MLP which will be discussed further in section B.

B. Multilayer Perceptron

Multilayer perceptrons (MLP's) are feedforward neural networks [8]. They are supervised networks, so they require a desired response to be trained. They learn how to transform input data into a desired response, so they are widely used for pattern classification. With one or more hidden layers, they can approximate virtually any input-output map. MLP's are probably the most widely used architecture for practical applications.

The network input, x to output, y relationship can be described mathematically as follows [8]:

$$y_k = \sum_{j=0}^M w_{kj}^{(2)} \tanh\left(\sum_{i=0}^d w_{ji}^{(1)} x_i\right) \quad (1)$$

Where, $w_{ji}^{(1)}$ and $w_{kj}^{(2)}$ indicate weights in the first and second layer, respectively, going from input i to hidden unit j , M is the number of hidden units and d is the number of output units.

Because of its efficiency, the scaled conjugate gradient method is used as the optimization technique used to train the networks. For a two class classifier one output node is sufficient, so there is only one activation at the second layer. The outputs from the hidden layer are connected via weighted connections to the output node and biased to form the second layer activation. The output y , is a continuous scalar bounded between 0 and 1, thus to use y as the indicator of class membership it needs to be converted to binary values using a threshold. The MLP network is trained to find the weights, biases. Overfitting or underfitting, which arises when the network does not generalize statistical patterns, is catered for by dividing the data sets into training, validation and testing.

C. Artificial Intelligence in HIV/AIDS Predictions

Artificial neural networks (ANNs) have been used to classify and predict the status of HIV/AIDS patients from symptoms

[7]. The data used were all the complete entries from a publicly available AIDS Cost and Services Utilization Survey performed in the United States of America. Multilayer perceptron architecture, with 15 linear inputs and 3 hidden logistic nodes and one output, being the HIV status or AIDS status, was trained using 200 epochs with a learning rate of 0.1 and momentum of 0.1. 1026 cases were used for training and 667 HIV cases were used for testing. The best accuracy obtained was 587 correct. A study was also performed to predict the functional health status of HIV and AIDS patients defined as well or not well, using neural networks [9]. The methodology presented here aims at using other demographic and social factors, to predict the status of an individual.

D. Genetic Algorithms

A genetic algorithm (GA) is an optimization method deriving its behavior from processes of evolution in nature, inspired by Darwin's theory of natural evolution [10, 11]. This is done by the creation within a machine/computer of a population of individuals. The individuals then go through the process of evolution. GA uses fitness-proportionate or tournament selection to select the missing entries (individuals) probabilistically that yields the right HIV status for the individuals. Although not guaranteed to provide the globally optimum solution, GA has been shown to be highly efficient at reaching a very near optimum solution in a computationally efficient manner [10, 11].

III. Methodology

The process of design of the adaptive controller was as follows:

1. Generate a mathematical model to appropriately represent the HIV prediction from demographic property using genetic algorithms.
2. Generate a model to predict an input given the output and other inputs, thus modeling output-input relationship.

A. Generating the Model for HIV Prediction

1) Data Source

Demographic and medical data came from the South African antenatal seroprevalence survey of 2001. This is a national survey, and any pregnant women attending selected public health care clinics participating for the first time in the survey were eligible to participate.

2) Missing Data

Out of the total data set cases, 12945 complete cases were selected, out of 13087 cases (98.91%) and the incomplete entries (142 cases – 1.09%) were discarded.

3) Variables

The variables obtained in the study are: race, region, age of the mother, age of the father, education level of the mother, gravidity, parity, province of origin, race, region of origin and HIV status. The qualitative variables such as race and

region are converted to integer values. The age of mother and father are represented in years. The integer value representing education level represents the highest grade successfully completed, with 13 representing tertiary education. Gravidity is the number of pregnancies, complete or incomplete, experienced by a female, and this variable is represented by an integer between 0 and 11. Parity is the number of times the individual has given birth. Both these quantities are important, as they show the reproductive activity as well as the reproductive health state of the women. The HIV status is binary coded; a 1 represents positive status, while a 0 represents negative status. There is one output.

4) Dataset Used

The dataset was divided into three sets; training, validation and testing sets. The sets were created by dividing the huge dataset into three equivalent small datasets of 1988 entries each. The inputs used were; age of the mother, age gap, educational level of the mother, gravidity, parity, province of origin, race, and region of origin.

5) Model

The NETLAB toolbox was used to create and train the MLP architecture. Three data sets were used for training, validation and testing. Also, since the determination of network architecture involves the optimization of the number of hidden nodes, it is incorrect to use the testing set to compare results before determining the final number of hidden nodes, and thus the final architecture. Genetic algorithm was used to optimize the network (hidden nodes, α , and training cycles). The network implemented comprised of 9 primary RBF neural networks each accepting a different demographic input and mapping that to the output (HIV). The outputs of these 9 primary RBF networks are then fed into a secondary MLP network and they are related again to the output. The primary RBF networks thus contained one input node, three hidden nodes and one output node. The secondary MLP network was composed of 9 input nodes, 77 hidden nodes (optimal) and 1 output node. A threshold value of 0.80107 was used since as this was found as the most optimal threshold during training and validation.

B. Generating the Inverse Neural Networks Model

The same datasets were used to create a network which relates the output and other inputs, to a particular input. Genetic algorithms was also used to optimize the network (hidden nodes, α , and training cycles). The network comprised of an MLP network comprising of 9 inputs (the educational level in the demographic input database being replaced by the HIV status) and 1 output (in this case the educational level). The educational level was chosen as the control variable because there was a correlation coefficient of about 0.75 between the educational level and the HIV status, which was the highest correlation. The educational level was also chosen as the controllable parameter because in the dataset used, it was the only modifiable parameter. Other parameters such as race, place of origin, region of

origin, province of origin, age of male and age of female, are inherent to an individual and are not modifiable. The network with 9 inputs and 1 output was then optimized using genetic algorithms. The number of hidden units returned by the genetic algorithm was 18 with an alpha of 0.0012421 and 984 training cycles, which was found after training, validation and testing. Inverse neural network models have been used for the implementation of adaptive controllers. Neural networks have been applied in adaptive control [12 – 14].

The inputs are used in the feed forward neural network to predict the HIV status. If the HIV status is positive then the inverse neural network model is used to predict the input parameter value required to make the status negative. This is then sent to the forward neural network again to make sure the prediction yields a negative.

C. Generating the Model for HIV Control

The datasets were then used to generate the overall model using genetic algorithms and neural networks and inverse neural networks, implemented in Matlab[®] simulink [15] to predict the educational level from the other demographic characteristics and HIV status. This model was used to assess the educational level that an individual whose status is predicted as positive requires to have been less prone to contracting HIV. For this model, an input dataset, comprising of the demographic input characteristics, is sent into a prediction model, which was implemented in section III-A of this paper. The prediction model classifies the HIV status of the individual with the demographic characteristics. If the classification yielded by the prediction network at this stage is 1 (indicative of an HIV positive class), then the educational level in the demographic input characteristics is replaced by a zero (representative of a negative HIV status, which is required) and the new input data is passed to the estimation model, which is the inverse model implemented in section III-B. The educational level thus required for the HIV status to be classified as zero is then obtained from the inverse neural networks model. The results are presented in the next section.

IV. Results

The data in section III was used with 9 demographic parameters. A feedforward and inverse neural network comprising of 9 inputs and 1 output was constructed and genetic algorithm was used to choose the optimal number of hidden units. The genetic algorithm yielded 77 hidden units as the optimal network that gives the best classification of the HIV status of individuals from the demographic characteristics for the feedforward neural network model. The genetic algorithm also yielded 18 hidden units as the optimal network that gives the best prediction of the educational level of individuals from the HIV status of the individuals and the other demographic characteristics for the

inverse neural network model. The networks were trained using scaled conjugate gradient method [16] with error backpropagation algorithms.

The first experiment was to use the data set for HIV modeling, thus predicting the HIV status of individuals from demographic characteristics. As earlier stated, genetic algorithms was used to obtain the optimal structure. The performance analysis for the HIV prediction model is based on classification accuracy and training times. Computation was performed on a computer with a processing speed of 3 GHz. The optimal number of hidden nodes for the neural network was 77 hence the structure was 9-77-1. This network gave an accuracy of 84.24% on the test data sets. The accepted threshold value, which gave the optimal value for conversion from continuous to discrete, was 0.80107. The training time was 352.8s.

The second experiment used the same dataset; however, the educational level of individuals was replaced by the HIV status of the individuals. The educational level was thus used as the output meanwhile the HIV status was used as one of the inputs. The inverse neural network model was thus created. The optimal number of hidden nodes for the inverse neural network model obtained by the genetic algorithm was 18 thus yielding a 9 – 18 – 1 structure. This network gave a mean square error of 5.9956 on the training dataset, thus about 94% accuracy and 11.573 on the validating data set thus about 88% accuracy. The training time was 46.9980s.

The adaptive controller was then implemented using the inverse neural network model. The inverse neural network model converged fast and yielded better results compared to the genetic algorithm model during preliminary design. The inverse neural network model however has a problem of singularity of solution due to the fact that the possible outcomes might be above the bounds of the input set. The genetic algorithm model yields a solution within the bounds due to the fact that it is a bounded optimization method. The status of the individuals could be controlled adaptively.

The sources of errors in the experiment are mainly data set related errors due to the data being biased towards one class and neural network training. To minimize the data set errors reliable the data was replicated for the class with less data. To minimize the neural network training errors, standard procedures were used for training, generalization and testing of neural networks. The effects of these errors were however minimal on the overall predictability [17]. The inverse neural network model also depends entirely on the correctness of the model and this may have also been a source of error.

V. Conclusion

In this study, a method based on neural networks is proposed to control the HIV status of individuals, by changing the status from positive to negative (or as the case may be). An adaptive control module was generated and implemented in Matlab simulink. This procedure was tested on a set of

demographic characteristics. The proposed method is able to estimate the parameters using GA and inverse neural and even though the genetic algorithms are computationally expensive, they yield bounded results, which are more realistic in terms of this application. Inverse neural network models yield results faster but these are however unrealistic since as these models yield unbounded results for some inputs, for example, an educational level of 14 will be unacceptable but could be yielded by the inverse neural network model. An accuracy of 88% was obtained by the inverse neural network model which was significantly accurate. Genetic algorithm was used to optimize the network parameters such as hidden units and training cycles. The authors thus conclude that a model can be developed to control the HIV status of an individual and also recommend that genetic algorithms be used for such a control system if time constraints permit, however based on accuracy obtained the authors do recommend inverse neural networks.

Acknowledgment

The authors thank the National Research Foundation (NRF) for their financial support.

References

- [1] K. Poundstone, S. Strathdee, D. Celestano. "The social epidemiology of human immunodeficiency virus/acquired Immunodeficiency syndrome", *Epidemiologic Reviews*, 26, pp. 22–35, 2004.
- [2] E. Fee, N. Krieger. "Understanding AIDS: historical interpretations and limits of biomedical individualism". *American Journal of Public Health*, 83, pp 1477 – 1488, 1993.
- [3] R. Tandon, S. Adak, J.A. Kaye. "Neural Network for longitudinal studies in Alzheimer's disease", *Artificial Intelligence in Medicine*, 36(3), pp. 245-255, 2006.
- [4] T. Sawa, L. Ohno-Machado. "A neural network-based similarity index for clustering DNA microarray data", *Computers in Biology and Medicine*, 33(1), pp. 1-15, 2003.
- [5] S.A. Kalogirou. "Artificial intelligence for the modeling and control of combustion processes: a review", *Progress in Energy and Combustion Science*, 29, pp. 515–566, 2003.
- [6] D.L. Hudson, M.E. Cohen. *Neural Networks and Artificial Intelligence for Biomedical Engineering*, ser. IEEE Press Series in Biomedical Engineering. Piscataway, NJ: IEEE Press, 2000.
- [7] E.O. Laumann, Y. Youm. "Racial/ethnic group differences in the prevalence of sexually transmitted diseases in the United States: a network explanation". *Sex Transm Dis*, 26, pp. 250– 61, 1999.
- [8] S. Guessasma, G. Montavon, C. Coddet. "On the neural network concept to describe the thermal spray deposition process: an introduction", In *Proceedings of the International Thermal Spray Conference and*

- Exposition*, Düsseldorf, DVS-Verlag GmbH, pp. 435–439, 2002.
- [9] T. Takano, K. Nakamura, M. Watanabe. “Urban residential environments and senior citizens; longevity in megacity areas: the importance of walkable green spaces”. *Journal of Epidemiology Community Health*, 56, pp. 913 – 918, 2002.
- [10] J. Holland. *Adaptation in Natural and Artificial Systems*, Ann Arbor: University of Michigan Press, 1975.
- [11] D.E. Goldberg. *Genetic algorithms in search optimization and machine learning*. Reading, MA: Addison-Wesley; 1989.
- [12] R. Salman. “Neural networks of adaptive inverse control systems”. *Applied Mathematics and Computation*, 163 (2), pp. 931 – 939, 2005.
- [13] H. Arab – Alibeik, S. Setayeshi. “Adaptive control of a PWR core power using neural networks”. *Annals of Nuclear Energy*, 32 (6), pp. 588 – 605, April 2005.
- [14] L. Chen, S.K. Narendra. “Nonlinear adaptive control using neural networks and multiple models”, *Automatica*, 37 (8), pp. 1245 – 1255, August 2001.
- [15] Appendix MATLAB 7.1 Manual, *Matlab and Simulink for Technical Computing*, Release 13, Mathworks, 2004.
- [16] M. Møller. “A scaled conjugate gradient algorithm for fast supervised learning”. *Neural Networks*, 6(4), pp. 525-533, 1993.
- [17] T. Marwala, S. Chakraverty. “Fault classification in structures with incomplete measured data using autoassociative neural networks and genetic algorithm”. *Current Science Journal*, 90 (4), pp. 542 – 549, 2006.

Author Biographies

Brain Leke Betechuoh, born in 1982 in Cameroon, is a PhD student at the University of Witwatersrand, Johannesburg, South Africa. He received a BSc. in December 2003 and received an MSc. degree in May 2005, both in Electrical Engineering at the University of Witwatersrand. His research interests include financial modeling, HIV/AIDS modeling, and neural networks.

Prof Tshilidzi Marwala is the head of the Control and Artificial Intelligence research group at the University of Witwatersrand, Johannesburg, South Africa. He holds a PhD from the University of Cambridge, UK. His research interests include neural networks, evolutionary computing, agent based modeling, dynamic model updating, structural health modeling, and complexity modeling.

Thando Tettey is a masters student at the University of Witwatersrand, Johannesburg, South Africa. He received a BSc. in Electrical Engineering in December 2005 at the University of Witwatersrand. His research interests include neuro-fuzzy modeling and interstate conflict management using computational intelligence.