Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp231-236)

A Comparative Study between Genetic Algorithms and Line Search Algorithm Optimization for HIV Predictions

BRAIN LEKE BETECHUOH*, TSHILIDZI MARWALA*,TARYN TIM*, MONICA LAGAZIO**
*Electrical Engineering Department
*University of Witwatersrand
*Private Bag 3, Johannesburg, Wits 2050
**University of Kent, United Kingdom

Abstract: - Neural Networks are used as pattern recognition tools in data mining to classify HIV status of individuals based on demographic and socio-economic characteristics. The data consists of seroprevalence survey information and contains variables such as age, education, location, race, parity and gravidity. The multilayer perceptron (MLP) neural network architecture was used for this study since as preliminary design showed this architecture to be the most optimal. The design of classifiers involves the assessment of classification performance, and this is based on the accuracy of the prediction using the confusion matrix. Two design approaches were implemented and a comparative analysis done. An accuracy of 84.24% was obtained for the genetic algorithms meanwhile an accuracy of 74% is obtained for the line search optimized network. The network structures for the different methodologies as well as the training and optimization times are also different. The gradient method proved to be the less computationally expensive but the most erroneous.

Key-Words: - Bayesian, Classification, Neural Networks, Multi Layer Perceptron, Confusion Matrix, Genetic Algorithms, AIDS, Conjugate Gradient Methods.

1 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. Since the identification of the virus and the disease, very little has been effective in stopping the spread. AIDS is now an epidemic, which at the end of 2003 had claimed an estimated 2.9 million lives.

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 disease [1]. 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]. The prevalence of infectious diseases is dependant on the nature of the disease transmission. 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].

Artificial intelligence techniques have been used successfully in medical informatics for decision making and prediction of outcomes (section 3). Neural networks are capable of non-linear pattern recognition without the need for an exact model. When applied to classification, neural networks are, firstly, used to discover which characteristics or combinations of characteristics are useful for distinguishing between classes. The second 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 [3] and then use the pattern classification system to classify new unseen cases. Based on the work that has been previously carried out in the field of HIV classification (Section 3), the objective of this study is as follows; Multilayer Perceptron (MLP) neural networks are trained to classify the HIV status of individuals by providing them with demographic factors, social and behavioral characteristics of the individuals. Preliminary work carried out by the authors of this paper, showed that compared to the Radial Basis Functions (RBF) neural networks, the MLP outperformed the RBF with respect to accuracy. The experimental data was obtained from antenatal seroprevalence surveys conducted in South Africa.

Genetic algorithms based on quasi-Darwinian evolutionary principles, can be used to implement efficient search strategies for optimal ANN configurations. In a genetic algorithm, neural

network parameters are represented by mathematical ''chromosomes'', which can be modified by mutation and by recombination in a process analogous to crossing over during meiosis. Genetic algorithms have been used with ANN to search for input variables [4], or to determine the number of nodes or connections in the network [5]. We used a genetic algorithm to search for optimal architectures, connectivity, and training parameters for ANN for predicting HIV status.

Section 2 provides a background on Biomedical Individualism; meanwhile Section 3 gives an overview on the use of Artificial Intelligence in HIV/AIDS predictions. Section 4 provides a background on the neural network architecture used in this study, MLP. Classification performance measures are explained in Section 5. The methodologies used to train and assess the networks are explained in Section 6 and the results obtained are presented in Section 7.

2 Biomedical Individualism

Biomedical individualism [1, 2] is defined as the basis of risk factor epidemiology and this is different from social epidemiology in that, in social epidemiology, social conditions are considered as the fundamental causes of diseases meanwhile in biomedical individualism, demographic and behavioral characteristics are considered. Poundstone et al. [1] related the demographic properties and their effect on HIV. They also related other demographic and social characteristics such as structural violence and discrimination, Race/ethnicity and racism [6], stigma and collective denial, legal structures, demographic change and policy environment, to the spread of the HIV. This thus justifies the use of such socio-demographic parameters in creating a model to predict the HIV status of individuals. The novelty in this design is the application of genetic algorithms in the HIV prediction from demographic data.

3 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 [6]. 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 [8]. The other applications of neural networks in AIDS research have been in bioinformatics where modeling of the virus has been done on a molecular level, such as the prediction of HIV-1 Protease Cleavage Sites. The above models concluded that ANN performed well in pattern recognition and signal processing. The methodology presented here aims at using other demographic and social factors, to predict the status of an individual.

4 Multilayer Perceptron

Multilayer perceptrons (MLP's) are feedforward neural networks [7]. 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. [7]
The network can be described as follows:

$$y_k = f_{outer}\left(\sum_{j=1}^{M} w_{kj}^{(2)} f_{inner}\left(\sum_{i=1}^{d} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right), \qquad (2)$$

Where y_k represents the k-th output, f_{outer} represents the output layer transfer function, f_{inner} represents the input layer transfer function, w represents the weights and biases, $^{(i)}$ represent the i-th layer.

The linear activation function will be used for the output and the hyperbolic tangent function will be used in the hidden layers [9]. 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 [7]:

$$a_j^{(2)} = \sum_{j=1}^{h} w_j^{(2)} \phi_j + b^{(2)} \qquad (3)$$

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.

5 Genetic Algorithms

Neural network structure and training parameters were represented by chromosomes consisting of ''genes'' of binary and continuous numbers. Each chromosome had four genes. The first gene representing the number of nodes in the hidden layers of the network, which could each range from 0 to 120 nodes. The second representing the learning rate which was confined between 0.001 and 1 (practical after pre-evaluation). The third gene was the number of training cycles between 100 and 1000. The fourth gene representing the threshold value. These mathematical chromosomes could be operated upon by quasi-genetic processes of crossing over and mutation. The genetic algorithm was started with 100 randomly generated chromosomes, with gene structures as described above. The genes were decoded, and neural networks with architecture and learning parameters represented by the decoded genes were trained on the training cohort. Fitness of each network was measured as the accuracy of the prediction. The chromosomes of propagated networks were then modified by crossing over and mutation, and the modified chromosomes were decoded to provide new parameters for the next round of network evolution. This process of fitness measurement, selection, crossover recombination, and mutation was iterated through 100 generations; and the network with the best accuracy in the final generation was designated as the optimal evolved network. All genetic algorithms [4, 5] and ANN were programmed in Matlab[10].

6 Quality of Classification

6.1. The confusion matrix

The MSE error is insufficient as a classification accuracy measure, as it indicates only the total number of correct classifications. In medical diagnosis in particular, it is necessary for a more detailed accuracy analysis, including the number of false positives, false negatives, true positives and true negatives. The confusion matrix shows the cross-classification of the predicted class against the true class. By splitting misclassifications into the different cells of the matrix, it is possible to assign a cost of making that particular misclassification [9]. The confusion matrix is given in (5).

$$C = \begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix} \quad (5)$$

Where: TN = True Negatives, FP = False Positives, FN= False Negatives and TP = True Positives.

The rows represent the true classes and the columns represent the predicted classes. The ideal solution has no false positives, nor false negatives, so the diagonal entries are at a maximum. Usually, as is true for this case, the cost of misclassification is difficult to determine. Using the quantities in the confusion matrix, it is possible to derive the *Receiver Operating Characteristic* or *ROC* curve. It is also possible to get the accuracy for the measurements from the confusion matrix which will be used to qualify the network and the results obtained.

6.2 Receiver Operating Characteristic and Accuracy

The neural network classifiers produce a continuous output indicative of the probability that the element belongs to a class. A threshold is applied to convert this output to predict class membership, and the value of the threshold affects performance. For an instance and a classifier there are four possible outcomes: true positive, where the instance is positive and is classified as positive; true negative, where the instance is negative and is classified as negative; false positive, where the instance is negative but is classified as positive; and false negative, where the instance is positive but is classified as negative. The True Positive Rate (hit rate or sensitivity) is defined in (6), and the False Positive Rate (false alarm rate), or specificity is defined in (7).

$$TruePositiveRatio = TP/(TP + FN) \quad (6)$$

$$FalsePositiveRatio = FP/(FP + TN) \quad (7)$$

The True Positive Ratio [11] is plotted against the False Positive Ratio for different threshold values. The accuracy in general is the number of correctly classified out of the total number of cases. The accuracy is obtained as follows:

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \quad (8)$$

6.2.1 Model Selection

Plotting the True Positive Ratio on the y-axis against the False Positive Ratio on the x-axis results in the ROC curve, which depicts the trade-offs between true positives and the costs (false positives). Perfect classification occurs at the point (0,1) on the ROC space, while (0,0) indicates that the classifier never issues positive classifications, and (1,1) represents a

classifier always issuing positive classification[11]. The threshold can therefore be selected according to the misclassification costs: a classifier in the lower left hand side should be selected for strong evidence.

7 Implementation

7.1 Data Processing

7.1.1 Data Source

Demographic and medical data came from the South African antenatal seroprevalence survey of 2001. The antenatal seroprevalence surveys are used as the main source of HIV prevalence data worldwide, reasons for this are that antenatal clinics are found throughout the world, and pregnant women are ideal candidates for the study as they are sexually active. Antenatal refers to the pregnant women and seroprevalence is the level of a pathogen in a population, measured in blood serum.

7.1.2 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.

7.1.3 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 [12]. 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. Parity is the number of times the individual has given birth, (for example, multiple births are counted as one) and this is not the same as gravidity. 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. The final number of input variables is 9. There is one output.

7.1.4 Outliers

Age is the only variable with outliers. The standard age bracket used in demographic studies relating to female fertility is 14-50 in African countries, and this was used to extract outliers in mother's age. The mean difference in age between mother and father is 7 years, and the upper limit on age of the father is thus 57. The data is partitioned into 3 sets, training, validation and testing.

7.1.5 Dataset Used

The dataset was divided into three sets; training, validation and testing sets. 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. The training set is balanced to consist of an equal number of positive outcomes as negatives, by duplicating the positive entries. An alternative to oversampling the minority class is to assign distinct costs to training examples, or by undersampling the majority class [6]. The original training set consisted of more negatives than positives with a ratio of 3:1. Neural networks are trained to model the statistical properties of the data, and had the neural network been trained on this biased dataset, the predicted outcome would always have been negative. This data was randomized and the inputs were scaled between 0 and 1.

7.2. Neural Network Architecture Design

7.2.1 Training

The scaled conjugate gradient optimization technique is used in error back-propagation to train the networks. Genetic algorithms are then used to train and optimize the networks. A line search optimization algorithm is also used to optimize the network.

7.2.2 Number of Hidden Nodes

With the number of input nodes for the primary RBF and MLP networks fixed at 1 and the input nodes for the secondary MLP network fixed at 9, the upper limit on the number of hidden nodes was 110, using a factor of 2 for the weights to data point's ratio. However, the ability of the neural network to model the data depends on the nature of the data, the types of inputs and not just the quantity of training examples available. Several networks of differing complexity between 2 and 110 hidden nodes were trained with a regularization coefficient of between 0.001 and 1. The number training cycles ranged from 100 to 1000. A polynomial minimization cost function was used to obtain the best suited parameters based on accuracy obtained from the confusion matrix. Genetic algorithm [4, 5] was then also used to optimize the number of hidden nodes, the regularization coefficient and the number of training cycles. The fitness criteria used was the validation accuracy. The optimized network from the genetic algorithm had 77 hidden nodes, a regularization coefficient of 0.24693 and required 144 training cycles. For the line search optimization there are 20 hidden nodes, regularization of 0.06 and 200 training cycles.

7.3 Threshold adjustment

The output is binary: 1 for HIV positive and 0 for HIV negative. The result from the neural network, however, is a continuous value between 0 and 1 and this needed to be hard limited to either 0 or 1. This was achieved by rounding the output to 1 if greater than a threshold and rounding it to 0 otherwise. The confusion matrix was calculated initially for a threshold of 0.5 on the training data, and this value was adjusted until the ratio of false positives to false negatives was approximately 1. The acceptance of the threshold was based on the accuracy obtained from the training and validation data sets. This threshold value was adjusted until an optimal value for the accuracy was obtained. Using this adjusted threshold, validation and final testing are limited at 0 or 1. The threshold was also optimized using the genetic algorithms. Threshold values of between 0.25 and 0.85 were tested and the optimal value yielded by the genetic algorithm was 0.80107. The optimal value yielded from the line search algorithm was 0.36. The best performing MLP network has a testing accuracy of 0.8424 (84.24%) for the genetic algorithms and 74% for the line search algorithm.

8 Results

The performance analysis is based on classification accuracy and training times. The most optimal network was the RBF network with 10 primary RBF networks, and 1 secondary MLP network. The optimal number of hidden nodes for the primary networks was 3; hence the structure was a 1-3-1 structure. The optimal number of hidden nodes for the secondary network was 77 for the Genetic algorithm optimized network; hence the structure was 9-77-1. The optimal number of hidden nodes for the secondary network was 20 for the line search optimized network.

TABLE 1
Classifier Confusion Matrix of the Two Algorithms

Confusion Matrix	Predicted Pos	Predicted Neg
Line Search Algorithm		
Actual Positive	594	444
Actual Negative	0	993
Genetic Algorithm		
Confusion Matrix	Predicted Pos	Predicted Neg
Actual Positive	680	313
Actual Negative	0	993

The genetic algorithm network combination gave an accuracy of 78.66% during training time and an accuracy of 84.24% on the test data sets meanwhile that of the line search optimized network yielded 75% during training and 74% on the test data. The accepted threshold value which gave the optimal value was 0.80107 for the genetic algorithm and 0.36 for the line search optimized network. The confusion matrix obtained for the genetic algorithm and line search algorithm network is as shown in Table 1. The accuracy of the above networks for the data sets is calculated using (8).

8.1 Analysis of two optimization approaches

The two approaches were analyzed for such aspect as accuracy, sensitivity, specificity, and duration. The first approach using genetic algorithms outperformed the other network vis-à-vis accuracy but was however computationally expensive. The classical network had a true positive rate of 100% and a true negative rate of 54.12%. It had a false positive rate of 46.73% and a false negative rate of 0%. The precision of this network was 68.15% with an accuracy of 76.64%. The genetic algorithm network had a true positive rate of 99.90 and a true negative rate of 68.58%. It had a false positive rate of 31.42% and a false negative rate of 0.10%. The precision of this network was 76.07% with an accuracy of 84.24%. The genetic algorithm network thus outperformed the line search network both with respect to accuracy, sensitivity, specificity and precision. The only significant performance of the classical network was computational time to optimize the parameters. The networks were also tested using two other data sets; one comprised of all positive cases; and the other comprised of all negative cases to test the ability of the network to generalize. Both networks performed averagely with the standard optimized network excelling for the first set comprised of all positives with an accuracy of 100% and the genetic algorithm excelled for the second data set with an accuracy of 69%. Both networks can thus be used to predict the HIV status of an individual from demographic data as they possess the ability to generalize appropriately. However based on accuracy, specificity, sensitivity, and precision genetic algorithm is recommended for the prediction and classification of HIV from demographic data.

9 Conclusion

Artificial intelligence methods can be used for classification of the HIV status of an individual, given certain demographic factors. The preliminary stage of the design proved that the multi layer perceptron networks were more accurate than the other possible architecture, the radial basis function, thus this was used as the optimal design architecture. In order to avoid over-fitting, a validation set is used during training to optimize the design of the

network. The training data set was used to choose the best network and this may lead to the network not being able to generalize. To overcome this, a validation set of data is required to optimize the design of the network. Optimizing the design involves setting the regularization parameters and the complexity of the network, set by the number of hidden nodes. Two approaches of choosing the optimal network are then analyzed. It was found that the genetic algorithms significantly outperformed the standard network optimization process vis-à-vis accuracy, specificity, precision and sensitivity and is recommendable for HIV classification. The optimal number of hidden neurons obtained from the genetic algorithm was 77 neurons for the secondary MLP network thus forming a structure 9-77-1 (with 9 inputs, 77 hidden nodes and 1 output node). The optimal number of training cycles was also optimized. This value was found as 144. The optimal regularization parameter was 0.24693. The optimal node for the second approach, line search optimization, was 20 with the number of training cycles being 200 and the regularization parameter being 0.06. Performance metrics such as the accuracy and the ROC curve are used to analyze the classification. Design of classifiers involves setting a threshold value to convert a probabilistic output to an actual classification. This value was determined by using genetic algorithms, and in this study, was set to the value that gives the best accuracy value, 0.80107. The line search optimization approach yielded 0.36. The results show that the MLP neural network architecture was quite efficient in predicting the HIV status from demographic properties to 84.24% accuracy compared to 74% from the second approach thus; demographic data is to an extent sufficient to accurately predict HIV status and genetic algorithms is more appropriate for the prediction and classification. Draghici and Potter [13] did a study on predicting HIV drug resistance using neural networks for classification and correctly classified unseen data with an accuracy of between 60% and 70%. The method employed in this design thus significantly increased on the accuracy. Further parameters such as; the financial standing of individuals, a more elaborate data set which spans the entire population and the inter-mobility level of individuals, will enhance the prediction and classification and possibly increase the accuracy significantly. It is recommended that different input features be tested, as well as automatic relevance detection to assess which inputs contribute to the output. By observing these features, it would be easier to find additional relevant input features.

References:

[1] K. Poundstone, S. Strathdee, and D. Celectano, The social epidemiology of human immunodeficiency virus/acquired Immunodeficiency syndrome, *Epidemiologic Reviews*, Vol. 26, 2004, pp. 22–35.

[2] E. Fee and N. Krieger, Understanding AIDS: historical interpretations and limits of biomedical individualism. *American Journal of Public Health*, Vol. 83, 1993, pp 1477 – 1488.

[3] D. L. Hudson and M. E. Cohen. *Neural Networks and Artificial Intelligence for Biomedical Engineering, ser. IEEE Press Series in Biomedical Engineering*. Piscataway, NJ: IEEE Press, 2000.

[4] M.N. Narayanan, S.B. Lucas, A genetic algorithm to improve a neural network to predict a patient's response to warfarin. *Meth Inform Med*, Vol. 32, 1993, pp. 55-8.

[5] M.E. Jefferson, N. Pendleton, C.P. Lucas, S.B. Lucas, M.A. Horan, Evolution of artificial neural network architecture: prediction of depression after mania. *Meth Inform Med*, Vol. 37, 1998, pp. 220–5.

[6] E.O. Laumann and Y. Youm, Racial/ethnic group differences in the prevalence of sexually transmitted diseases in the United States: a network explanation. *Sex Transm Dis*, Vol. 26, 1999, pp. 250– 61.

[7] I. Nabney, *Netlab: Algorithms for Pattern Recognition*. Springer Verlag, 2003.

[8] S. Sardari and D. Sardari Applications of artificial neural network in AIDS research and therapy, *Current Pharmaceutical Design*, Vol. 8, No. 8, 2002, pp. 659-670.

[9] N. Lavrac, Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine*, Vol. 16, 1999, pp. 3–23.

[10] MATLAB 7.1 Manual, *Matlab and Simulink for Technical Computing*, Release 13, Mathworks, 2004.

[11] T. Fawcett, *Roc graphs: Notes and practical considerations for data mining researchers*, Intelligent Enterprise Technologies Laboratory, HP Laboratories, Tech. Rep. HPL-2003-4, 2003.

[12] M. Rauner and M. Brandeau. AIDS policy modeling for the 21st Century: An Overview of Key Issues. *Health Care Management Science*. Vol. 4, 2001, pp. 165 – 180.

[13] S. Draghici and B. Potter, Predicting HIV drug resistance with neural networks. *Bioinformatics*. Vol. 19, No. 1, 2003, pp. 98 – 107.