International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 9 (2015) pp. 22969-22984

http://www.ripublication.com

Prediction of HIV/AIDS Disease Using Multilayer Perceptron and Radial Basis Function With Dempster-Shafer Theory of Evidence

A.M. Saravanan*
Asst. Professor of Computer Science
Muthurangam Govt. Arts College, Vellore- 632 002
Tamil Nadu, INDIA.
amsara1@rediffmail.com
Dr. C. Jothi Venkateswaran
Associate Professor of Computer Science
Presidency College, Chennai – 600005
Tamil Nadu, INDIA.
jothivenkateswaran@yahoo.co.in

Abstract

The historical data about medical diagnosis and Demographic attributes are collected from annual antenatal surveys done in and around Vellore district, Tamilnadu. This information is used to predict status of acquiring HIV. The predicting system is designed with a neural network which is trained to obtain the classification of status of HIV. The surveillance information contains many attributes and those attributes are filtered using Dempster-Shafer theory of Evidence. Such filtered discrete variables are age, gravidity, parity, as well as the quantitative variables race and location, making up the input to the neural network. The output of the NN is the HIV status. A multilayer perceptron with a logistic function is trained with a cross entropy error function, providing a probabilistic interpretation of the outcome. The performance of the prediction is measured, and the sensitivity and specificity are illustrated on the Receiver Operating Characteristic. An auto-associative neural network is trained on the full datasets and if there is any missing entries then they are approximated with global optimization methods. The effect of the imputed data on the network prediction is investigated.

Keywords: Dempster-Shafer theory, Multilayer perceptron, Radial Basis Function, Threshold, Receiver Operating Characteristic

Introduction

Acquired Immunodeficiency Syndrome (AIDS) was first defined in 1982 to describe the first cases of unusual immune system failure. During 2006, around four million adults and children are estimated to have been infected with HIV in the world. By the end of the 2006 [1], an estimated 10 million people were living with HIV/AIDS. India is having localized epidemics of HIV infection, National Family and Health Survey III (NFHS) data generated through population based survey. One of the states in India, Tamil Nadu has shown a considerable change in HIV scenario over a period. This is evident from the HIV sentinel surveillance (HSS), Behavior Surveillance Survey (BSS), AIDS Prevention And Control (APAC) and recently from NFHS-III and Integrated Behavioral and Biological Assessment surveys (IBBA). Government of Tamil Nadu has initiated aggressive programme against the infection. Initial activities are focused on awareness generation.

To effectively manage this epidemic, accurate information on prevalence, improved understanding of the socio demographic factors in which the epidemic occurs and the relative impact of interventions is required to construct and improve behavior and treatment interventions. This is obtained by creating a model of the HIV epidemic. The aim of this study is to predict the HIV status of an individual, given readily available demographic data. This knowledge will be used to construct health and social policies for HIV/AIDS prevention. A few national population based surveys have been conducted to determine behavioral and social factors influencing the prevalence of HIV.

Knowledge discovery and data mining in HIV related demographic are to be used to obtain the model. Artificial intelligence has been used successfully in medical informatics for decision making, and is used in this paper. To successfully perform knowledge discovery, an understanding of the data and application is necessary.

Knowledge Discovery Process

The steps involved [2] in the KD Process are

1. Develop an application with prior knowledge, understanding of problem and goals.
2. Create a target data set to be used for discovery.
3. Clean and preprocess data
4. Reduce the number of variables and find invariant representations of data if possible.
5. Select the suitable the data-mining task .
6. Use appropriate data-mining algorithm.
7. Search for patterns of interest (this is the actual data mining).
8. Interpret the pattern mined and if there is a need then repeat the steps from 1 to 7.
9. Consolidate knowledge discovered and prepare a report

The general goal of data mining is to uncover relationships within the data and to predict outcomes. Data mining involves fitting models to or determining patterns from data. The general algorithm for data mining consists of three parts.

1. The model: The function of the model and its representational form, i.e. the data mining task
2. The preference criterion: The basis for which a model set of parameters is given preference for the particular data set. This is usually a goodness of fit function to the model. In terms of optimization, this can be seen as the function to judge the quality of the fitted models on observed data.
3. The search algorithm: the algorithm for finding particular models and parameters, given the data, model and preference criterion.

Classification and Regression

The goal is to develop a model allowing one variable to be predicted from known values of other variables. If the predicted value is categorical, the function is classification, and if the predicted value is quantitative, the function is regression [3].

- Classification: Classifies the data into predefined categories: i.e., HIV positive and negative
- Regression: It maps the data item into a real value variable

Numerous soft-computing algorithms exist for data mining, including: neuro fuzzy computing, genetic algorithms, neural networks, rough sets, decision trees and hybridizations. The choice of algorithm is based on the data mining task, as different algorithms are better suited to different tasks.

Demographic, Behavioral and Social Risk Factors For HIV

The most comprehensive population-based survey is the study of HIV/AIDS carried out in the past years, which provides more accurate HIV-related sexual behavior risk profiles. The effects of race, age, locality, province, history of diagnosis of STI, and education levels are investigated. People living in informal urban areas were significantly more likely to be HIV positive than those living in urban formal areas. There is no simple relationship between HIV infection and levels of education. School attendance may increase access to both information and potentially to prevention interventions. However, the improvements in socioeconomic status and lifestyle changes that go with higher educational attainment may be associated with behaviors that increase the risk of HIV infection [4]. International studies have identified key behavioral factors such as age, condom usage, median age at first sex, and knowledge about AIDS [5]. Risk factors influencing prevalence are related to the risk behavior of the male partner: marital status, having multiple sexual partners, having a male sexual partner who drank alcohol or who had higher income[6;7]

Neural Network Tools For Classification In Medical Research

The use of neural networks in the medical field in general, as well as their suitability for classification. Artificial intelligence has been used in medicine for the clinical functions of diagnosis, prognosis and survival analysis, and decision support. It has been used in a wide variety of medical domains such as oncology, critical care, and tuberculosis, cardiovascular and renal transplantation. Artificial Neural Networks (ANNs) perform well in pattern recognition, and are suitable for signal processing (EEG, ECG, and hemodynamic signals), as well as image processing (mammography,

chest radiographs, tomography, nuclear medicine imaging, magnetic resonance). A common task in medicine is thus classification using predictive models.

The major strength of neural networks in modeling multidimensional spaces, is their ability to scale with increasing dimensionality of the data. The ability of a polynomial to model a non-linear function is limited by the number of terms in it, and hence the order of the polynomial allows it to model higher dimensions. To accurately determine the numerous parameters that would arise in a high order polynomial, large data sets would be required. In contrast to polynomials, which use one function to model the relationship, neural networks use superposition of many functions of a single variable each. These functions are known as the hidden functions, and adapt as the complexity of the model grows, not simply with dimensionality.

Artificial intelligence methods in HIV research

Despite the numerous applications of artificial neural networks to classification in medicine, very little attention has been made to the HIV/AIDS prevention and planning [8]. Artificial neural networks have been used to classify and predict the symptomatic status of HIV/AIDS patients. The inputs are: sex, race, exposure rate (homosexual, IV drug user, heterosexual), medical records. The output is HIV status. A study was performed to predict the functional health status of HIV and AIDS patients defined as well or not well, using neural networks [9]. The inputs were medical care access, such as number of emergency room visits and inpatient nights.

2. Dempster Shafer Theory

The Dempster-Shafer theory of evidence is based on the belief functions and subjective probability [10]. But in Bayesian theory uses degrees of belief for one question on probabilities for a other question. The degrees of belief may or may not have the mathematical properties of probabilities; how much they differ from probabilities will depend on how closely the two questions are related.

The Dempster-Shafer theory is based on two ideas: the idea of obtaining degrees of belief for one question from subjective probabilities for a related question [11], and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence. There are three important functions in Dempster-Shafer theory: the basic probability assignment function (bpa or m), the Belief function (Bel) [12], and the Plausibility function (Pl). The basic probability assignment (bpa) is a primitive of evidence theory.

The basic probability assignment (bpa) is denoted by the letter m, it defines value is between 0 and 1, and the summation of bpa is equal to 1. The value of the bpa for a given set A (m(A)) which expresses the proportion of relevant evidence to support the claim that a particular element of X (the universal set) belongs to the set A but to no particular subset of A . The value of m(A) pertains only to the set A and makes no additional claims about any subsets of A. The other evidence on the subsets of A would be represented by another bpa, i.e. B Ì A, m(B) would the bpa for the subset B. Formally, this description of m can be represented with the following three equations:

$$m: P(X) \rightarrow [0,1] \quad (1)$$

$m(\Phi) = 0$ (2)

$\sum m(A) = 1$ (3)

$A \in P(X)$

Where P(X) represents the power set of X, Φ is the null set, and A is a set in the power set ($A \in AP(X)$).

Using the basic probability assignment, the upper bound Plausibility and lower bound Belief interval contains the probability of a set of interest. The lower bound Belief for a set A is the sum of all the basic probability assignments of the proper subsets (B) of the set of interest (A) ($B \subset A$). The upper bound is the sum of all the probability assignments of the sets (B) that intersect the set of interest (A) ($B \cap A = \Phi$). Formally, for all sets A that are elements of the power set ($A \in P(X)$),

$Bel(A) = \sum m(B)$ (4)

$B|B \subseteq A$

$Pl(A) = \Sigma m(Bi), B \mid B \cap A \neq \varnothing$ (5)

The two measures, Belief and Plausibility are non additive. This can be interpreted as is not required for the sum of all the Belief measures to be 1 and similarly for the sum of the Plausibility measures [11]. It is possible to obtain the basic probability assignment from the Belief measure with the following inverse function:

$$m(A) = \sum_{B|B \subseteq A} (-1)^{|A-B|} Bel(B) \quad (6)$$

where |A-B| is the cardinality difference of the sets.

In addition to deriving these measures from the basic probability assignment (m), these two measures can be derived from each other. Plausibility can be derived from Belief [13] using the formula:

$$Pl(A) = 1 - Bel(\bar{A}) \quad (7)$$

where $\bar{A}$ is the complement of A. This definition of Plausibility in terms of Belief comes from the fact that all basic assignments must sum to 1.

$$Bel(\bar{A}) = \sum_{B|B \subseteq \bar{A}} m(B) = \sum_{B|B \cap A = \varnothing} m(B) \quad (8)$$

$$\sum_{B|B \cap A = \varnothing} m(B) = 1 - \sum_{B|B \cap A = \varnothing} m(B) \quad (9)$$

From the definitions of Belief and Plausibility, it follows that Pl(A) =1-Bel(A) . As a consequence of Equations 6 and 7, given any one of these measures (m(A), Bel(A), Pl(A)) it is possible to derive the values of the other two measures. The precise probability of an event (in the classical sense) lies within the lower and upper bounds of Belief and Plausibility.

$$Bel(A) = P(A) = Pl(A) \qquad (10)$$

Problem Statement

Using neural network model a classification system is to be developed and involves the selection of demographic attributes as parameters and optimization is used to classify an outcome in the given data. HIV classification using demographic factors, and the developed algorithm should be applicable to any complex system with sufficient data.

HIV Status Classification and Incomplete Data

The primary objective is to extract the features using Dempster-Shafer theory and to study is disease risk analysis with artificial intelligence methods like neural networks to perform knowledge discovery and data mining on HIV clinical and demographic data, resulting in a classifier of HIV status of a patient based on demographic inputs.
The problem with using neural networks is the need for a complete set of data. But the medical data often contains many missing entries and these missing data be replaced by a method was developed by Abdella and Marwala [14]. This is to be used to complete the data set. For the classifier, a multilayer perceptron was trained using the Bayesian framework, and a Genetic Algorithm was used to obtain the neural network architecture parameters.

Methodology

The artificial neural network (ANN) originated as a model for biological neural networks. Neural networks are viewed as a computational method of representing the non-linear functional mapping of an input to an output in this paper. In classification, they are function approximators, where the function approximated is the probabilities of membership to classes as a function of the input variables. Neural networks are thus used as predictive data models. The functional performance of the network depends on three factors: the selection of inputs, the network structure and the training of the network. The design and training of the network involves presenting it with a set of corresponding inputs and outputs, and training it to adapt to provide the outputs given a specific input.

Inputs and structure of Network

The inputs are selected using **Dempster-Shafer theory of evidence** [18] and the inputs are converted into binary value if they are continuous. There are many different ANN models; they all have the same basic structure. An ANN is a network of many simple processors (”neurons” or nodes), each with an associated memory. The nodes are connected in layers, most commonly three layers: input, hidden and output layers. The networks considered in this study are feed-forward: outputs of nodes in each layer are connected to nodes in the next successive layer. No information is fed back from the outputs to the input layers.

Training of The Network

Using the relationship between the inputs and outputs the interconnections are made in the network. Training is the process of adjusting the weights of the connections according to the data. This adjustment can either take place after one example is presented to the network. The training method depends greatly on the structure of the network and radial basis functions are trained in a very different way from the way multilayer perceptrons are trained.

Multilayer Perceptrons

The most widely used neural network architecture is the multilayer perceptron. The multilayer perceptron structure provides a non-linear mapping from a real-valued input vector x to a real valued vector y. It can thus be used as a non-linear model for both regression as well as classification, depending on the interpretation of the output(s). The functionality of the MLP is based on its multi-layer structure and activation functions. The main idea in MLPs is that the input vector is successively modified through multiplication by weight matrices in the different layers, and the products are transformed by non-linear activation functions.

Network Structure and Their Weights

The architecture of the MLP consists of an input , hidden and output layers. It has been proven that a single hidden layer feed-forward neural network is capable of approximating uniformly any continuous multivariate function to any desired degree of accuracy, provided that the number of hidden units is sufficiently large [15; 16]. A network with no hidden layer, that is, with only one layer of nodes with activation functions is sufficient for two class classification, only if the data points are linearly separable. This network is unfeasible, as for a set of N points in d dimensional space, a network with N/d nodes is needed to correctly separate the points into two classes [16].

The inputs are fed directly into the input layer, and each input is connected to every node in the hidden layer via a feed forward connection. The output from each hidden node is connected to every output node.

Mathematical Function

Every hidden node has an associated activation due to its weighted connections to each input xi. There is also a bias bj at each hidden node, thus the first layer activations are defined in [16] by:

$$a_j = \sum_{i=1}^{N} w_{ij}^{(1)} x_i + b_j^{(1)}$$

Where i is the number of inputs and j is the number of hidden nodes.

The significance of the number of hidden nodes in a classifier is that each hidden unit divides the input space with a hyper plane, so that activation z = 1 is on one side of the hyper plane, and z = 0. Sigmoid functions are S-shaped, mapping the interval $(-\infty,\infty)$ onto the interval $(-1, 1)$) for hyperbolic tangent functions, and onto the interval (0, 1) for the logistic sigmoid function. They are differentiable, which is a

necessary property for back-propagation, and are able to represent smooth mappings between continuous variables. These types of functions are used as activation functions as they are monotonic. In the hidden layer, hyperbolic tangent functions are used as they are equivalent to the logistic function through a linear transformation, but in addition to all the properties offered by the logistic function, also give rise to a faster convergence of training algorithms than the logistic functions [16]. The outputs of the hidden nodes zj are thus:

$$z_j = \tanh(a_j^{(1)})$$

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 [21]:

$$a^{(2)} = \sum_{j=1}^{h} w_j^{(2)} z_j + b^{(2)}$$

This second layer activation is transformed by the logistic output activation function, as it operates in the range 0 to 1, and it allows the output to be given a probabilistic interpretation, since it is derived using Bayes theorem to represent the posterior probabilities of membership to classes. Additionally, the sigmoidal function is able to represent both non-linear functions as well as linear functions (if |a| is small).

$$y = \frac{1}{1+e^{-a(2)}}$$

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. Since the resultant model is non-linear, when applied to classification, the decision boundary between the classes produced by the network is also non-linear. This is an advantage over most other classification methods such as trees which have linear decision boundaries. Non-linearity allows for highly flexible decision surface shapes, but since non-linear estimation of the parameters is not straightforward iterative techniques are used (training). The process of obtaining the weights that produce the model is called training.

Training and Decision Surface

The weights and biases that result in the optimal decision surface are determined through the minimization of the error produced by the network. The error function is the sum of squares error in the target outputs t, and those produced by the network y. The optimization methods used are gradient based, and these require the derivative of the error function with respect to the weights. Sigmoid activation functions are used because they are differentiable, and this is a necessary requirement in error back-propagation. The error back-propagation algorithm solves for the weights through the propagation of the errors backwards throughout the network.

The output activation function determines the decision surface in the input plane, in this case a logistic output activation function is used, since the range is 0 to 1, a typical value for a threshold would be 0.5, and thus y = 0.5 defines a decision surface in the output space that can separate the classes.

Generalization

The aim of the network is to capture the statistical properties of the data to be able to make accurate predictions for new inputs, that is, to generalize. Poor generalization can either arise if the network is not sufficiently complex and is unable to accurately represent the process, or it may be too flexible, and be fitted to the noise of the training data, or the data is too difficult to be mapped using neural networks. There needs to be an optimization between these two points, to find a model that represents the data accurately enough to correctly predict the new data..

Radial Basis Function

The major difference between radial basis function networks and MLPs is that the activation of the hidden unit is determined by the distance between the input vector and a weight vector. Their major attractive feature is that they train much faster than MLPs, and this is attributed to the two stage training procedure.

Network Structure

The structure of the radial basis function network is similar to that of the MLP, however instead of hidden nodes with activation functions, the RBF has basis functions. Also, there is usually only one hidden layer in an RBF network. Like the MLP, the nodes are feed-forward, fully connected and have weights and biases.

Mathematical Function

Radial basis functions use a combination of supervised and unsupervised learning techniques. Learning in the hidden layer is unsupervised, using methods like k-means clustering. These clusters are used as the starting points from which supervised learning in the outer layer takes place. A least mean squares method is used for the training and the Gaussian function is

$$\phi(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

where σ controls the smoothness of the properties of the interpolating function. The parameter σ is the width , and the value is found in training. This function is localised, as $|x| \rightarrow \infty$, $\varphi \rightarrow 0$. The thin plate spline function has also been used, and can be written as follows: $\phi(x = x^{2\ln(x)})$

The output is represented as the sum of the product of the basis functions with the biases added: $y(x) = \sum_{j=1}^{M} w_j \phi_j(x) + w_{10}$

In a similar way that the biases were absorbed in the MLP as another input fixed at 1, the biases can be absorbed into the summation as another basis function with the activation set to 1.

The basis function is $\phi_j(x) = \exp\left(-\frac{|x-\mu_j|^2}{2\sigma_j^2}\right)$

where x is the d-dimensional input vector, μj is the centre vector of the basis function φj

Decision Surface and Training

The activations of the basis functions can be interpreted as the posterior probabilities of the presence of the corresponding features in the input space, and the weights can be interpreted as the posterior probabilities of class membership, given the presence of features. Instead of the hyper-planes in the MLP approach to classification, there are hyper-spheres separating the classes in the RBF.

Training an RBF takes part in two stages. The first stage is an unsupervised learning procedure, where the input data set is used to determine the parameters of the basis functions. These parameters are the basis function for centre and width are denoted by μj and σj respectively. The input data is used, and no target information is used. The parameters are set such that the network models the unconditional data density. Selecting the centre's can be done in a few ways: the simplest is to randomly select data points as the basis function centres. **Generalization**

In a similar way that limiting complexity improves generalization in MLPs, the network complexity is used to improve generalization in RBFs. A reduced number of hidden nodes, or in this case, basis functions, limits the networks ability to model the data exactly, and a smoother mapping is achieved

Implementation

Neural Network Classifier

The steps involved in designing a classifier are:

1. Data Collection and Processing: The survey data used was collated and processed to make it suitable for neural network training
2. Feature Choice: Since there were few distinct features present in the data, all features were used. This could be done because although the neural networks do not eliminate unnecessary features altogether, they adjust to the data by assigning larger weights to more relevant features.
3. Model Choice: This step involves the design of the neural network architecture. Several different architectures (MLP and RBF) and training methods (Maximum likelihood training, and variants of Bayesian Regularization) were investigated in this study. A GA was used to select the parameters for training and the architecture of an MLP.
4. Training Evaluation: Performance of the classifier depends on its ability to model the data correctly with the ability to generalize. Assessment of

classification accuracy was made with the Receiver Operating Characteristic. The data was gathered from TANSAC, Tamilnadu. in spreadsheet format, and was processed in MATLAB.

Data Processing

Data Source and Variables

Demographic and medical data came from the TANSAC surveillance survey of 2010. 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. Anonymity is guaranteed. The antenatal seroprevalence surveys are used as the main source of HIV prevalence data , reasons for this are that antenatal clinics are found throughout the country, 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. Information was obtained using a questionnaire, and the HIV status of the patient was measured using a enzyme linked immunosorbent assay (ELISA) test.

The variables used in this paper are: race, region, age of the mother, age of the father, education level of the mother, gravidity, parity and HIV status [1]. The qualitative variables such as race and region are converted to binary values in order to prevent placing an incorrect importance on these variables had they been coded numerically. The age of mother and father are represented in integer. The value 13 represents highest education level. Gravidity is the number of pregnancies is represented by numbers from 0 to 5. Parity is the number of times the individual has given birth with multiple births are counted as one and it shows the reproductive activity as well as the reproductive health state of the women. The HIV status is 1 for positive and 0 for negative.

Table of Input and Output Variables		
Variable	**Type**	**Range**
Input Variable		
Region A	Binary	0 or 1
Region B	Binary	0 or 1
Region C	Binary	0 or 1
Region D	Binary	0 or 1
Age	Integer	0 or 1
Race : Indian	Binary	0 or 1
Race : Coloured	Binary	0 or 1
Race : White	Binary	0 or 1
Race: Black	Binary	0 or 1
Education	Integer	0-10
Gravidity	Integer	0-5
Parity	Integer	0-5

Age of father	Integer	21-50
Mother's Age	Integer	16-40
Output Variable		
HIV Status	Binary	0 or 1

Dataset biasing

The training set is balanced to consist of an equal number of positive outcomes as negatives. To oversampling the minority class is to assign distinct costs to training examples, or by under sampling the majority class [17]. Due to the limited size of the dataset, over sampling the positive cases was used rather than under sampling the negative cases. The original training set consisted of 310 positives and 295 negatives. The NN is trained to model the properties of data, and had the neural network been trained on this biased dataset, the predicted outcome is negative. There are 1416 entries in this set. The validation and testing sets each contain 575 entries.

Maximum Likelihood Neural Network Architecture Design

Training and Number of Hidden Nodes

The scaled conjugate gradient optimization technique is used in error back-propagation to train the networks. With the number of input nodes fixed at 14, the upper limit on the number of hidden nodes was 12, using a factor of 5 for the weights to data points 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.

Weight Decay Constant

The effect of α on validation error is that an increase in α results in the weights being restricted and hence regularization of the training. Using the smallest network that consistently produced the lowest training and validation errors, alpha was optimised. The best performing network was selected based on the area under the ROC curve for training and validation. The value of α is adjusted to ensure that the validation error decreases as much as possible, before reaching an approximate constant value.

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. Since the data set consisted of equal positive and negative outcomes, the classifier should produce equal false positive and false negative results, such that it is not biased toward predicting either case. The networks were optimized for a unity false positive to false negative ratio. The threshold was adjusted on the training data to ensure that there was no bias to either outcome, i.e. specificity and sensitivity have equal importance.

Using this adjusted threshold, validation and final testing are limited at 0 or 1. It is incorrect to adjust the threshold on the validation set, since this set is not balanced. One hundred different neural networks were trained and the best classifier on validation data was selected based on accuracy of classification. The best performing MLP network has a validation ROC of 0.5648. With an initial threshold of 0.5, the accuracy is 54.48% on training, but the number of false negatives (151) exceeds the number of false positives (130). The number of true positives is 205 and there are 89 true negatives. Shifting the threshold to 0.53 results in a better balanced classifier, with 130 false negatives and 140 false positives.

The best performing RBF network has a validation ROC area of 0.5819. The initial threshold gives 145 false negatives and 123 false positives, so the threshold is shifted to 0.51, resulting in 201 false negatives and 156 false positives on training.

Results and Discussion

Maximum likelihood approach

The performance comparison is based on classification accuracy and training times. Although the RBF trains much quicker in 1.975744s than the MLP in 12.197473s, the accuracy on the RBF was significantly lower than was achieved for the best MLP network.

A summary of results for the simple networks is given in below Table 1.

Table 1: Comparison of MLP and RBF simple classifiers

Comparison of MLP and RBF simple classifiers				
	MLP		**RBF**	
Confusion Matrix	**Predicted Positive**	**Predicted Negative**	**Predicted Positive**	**Predicted Negative**
Actual Positive	205	151	211	145
Actual Negative	89	130	96	123

Accuracy is calculated in two ways, the simple accuracy is the total correct results expressed as a percentage of total entries, the results are shown in Table 2.

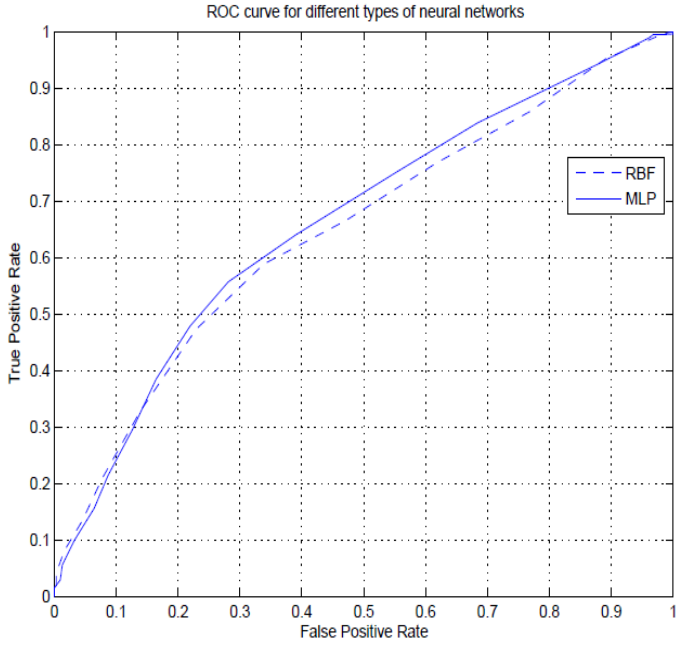
$$Simpleaccuracy = \frac{TP + TN}{TP + FN + FP + TN} \text{ and}$$

$$Geometricaccuracy = \sqrt{\frac{TP + TN}{(TP + FN)X(FP + TN)}}$$

Table 2: Accuracy comparison Multilayer Perceptron and Radial Basis Function classifiers

Type of Accuracy	**Multilayer Perceptron**	**Radial Basis Function**
Simple Accuracy	0.5826087	0.580869565
Geometric Accuracy	0.58465753	0.576961358

The Receiver Operating Characteristic (ROC) illustrates this binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of True Positives rate with the fraction of False Positive rates, at various threshold settings. The following ROC curves shows the difference between MLP and RBF Artificial Neural Network trained with maximum likelihood techniques.



Conclusion

Artificial intelligence methods can be used for classification. In this paper, supervised learning was used to train multilayer perceptrons and radial basis function networks to classify the HIV status of an individual, given certain demographic factors. Dempster-Shafer theory of evidence is used for data extraction and suitable attributes are selected for input nodes. Using the maximum likelihood method of training, the problems encountered are that different weights are obtained from different training sets, and the weights are also dependant on the order of the data entries in the training set. The network architectures Multilayer perceptrons and Radial Basis Function networks are chosen according to the lowest standard error between targets and predicted outputs, but the design process is time consuming since there are many

permutations of parameters. Finally using Maximum likelihood approach the classifiers MLP and RBF are compared with accuracies of simple and geometric methods. The result of average accuracy MLP is better than RBF and also proved that demographic data is not sufficient to accurately predict but the other factors like Geographic, Diagnostic, Treatment, Pharmacy, Laboratory variables may be included to predict the 100% accuracy classification of HIV Status.

References

[1] "National Behavioural Surveillance Survey 2006 " Questions and Answers; Facts about the HIV/AIDS epidemic and its impact." URL http http://naco.gov.in".

[2] S. Mitra, S. K. Pal, and P. Mitra. "Data Mining in Soft Computing Framework: A Survey." IEEE Transactions on Neural Networks, vol. 13, no. 1, pp. 3–14, 2002.

[3] D. Hand, H. Mannila, and P. Smith. Principles of Data Mining. Cambridge, Massachusetts: The MIT Press, 2001.

[4] O. Shisana. Nelson Mandela HSRC study of HIV/AIDS: Full Report. South African national HIV prevalence, behavioural risks and mass media. Household Survey 2002. South Africa: Human Sciences Research Council (HSRC), 2002.

[5] S. Armstrong, C. Fontaine, and A. Wilson. "2004 Report on the global AIDS epididemic.", 2004. URL http://www.unaids.org/bangkok2004/ report.html.

[6] S. Allen. "Human immunodeficiency virus infection in urban Rwanda. Demographic and behavioral correlates in a representative sample of childbearing women." The Journal of the American Medical Association, vol. 266, no. 12, September 1991.

[7] W. Siriwasin. "HIV Prevalence, Risk, and Partner Serodiscordance Among Pregnant Women in Bangkok." The Journal of the American Medical Associ-ation, vol. 280, no. 1, pp. 49–54, July 1998.

[8] C. W. Lee and J.-A. Park. "Assessment of HIV/AIDS-related health performance using an artificial neural network." Information & Management , vol. 38, no. 4, pp. 231–238, February 2001.

[9] N. K. Kwak and C. Lee. "A Neural Network Application to Classification of Health Status of HIV/AIDS Patient." Journal of Medical Systems, vol. 21, no. 2, 1997.

[10] Dempster, A. P., A generalization of Bayesian inference, J. Roy. Stat. Soc. Ser.B 30(2), 205-247, 1968.

[11] Shafer, G., A Mathematical Theory of Evidence, Princeton University Press,Princeton, N.J., 1976.

[12] Shafer, G., Perspectives on the theory and practice of belief Functions, Int. J.Approx. Reasoning 4(5 / 6), 323-362, 1990.

[13] Shafer, G., Rejoinders to comments on "Perspectives on the theory and practice of belief functions," Int. J. Approx. Reasoning 6(3), 445-480, 1992

[14] M. Abdella and T. Marwala. "The Use of Genetic Algorithms and Neural Networks to Approximate Missing Data in Databases." In Proceedings of the 73 IEEE International Conference on Computational Cybernetics, pp. 1001–1013. April 2005.

[15] M. S. K. Hornik and H.White. "Multilayer Feedforward Networks are Universal Approximators." Neural Networks, vol. 2, no. 5, pp. 359–366, 1989.

[16] C. M. Bishop. Neural Networks for Pattern Recognition. Oxford: Oxford University Press, 1995.

[17] P. Lisboa, A. Vellido, and H. Wong. "Bias reduction in skewed binary classification with Bayesian Neural Networks." Neural Networks, vol. 13, pp. 407–410, 2000.

[18] A.M. Saravanan, R.Vijaya, Dr. C.Jothi Venkateswaran "Feature Selection for Prediction of HIV/AIDS using Data Mining Technique by applying the concept of Theory of Evidence " IJCSNS International Journal of Computer Science and Network Security, VOL 11 No.5 May 2011, Pages 285-288