SURVIVABILITY ANALYSIS OF PEDIATRIC LEUKAEMIC PATIENTS USING NEURAL NETWORK APPROACH

T. M. D.Saumya[1], T. Rupasinghe[2] and P. Abeysinghe [3]

[1] Department of Industrial Management, University of Kelaniya, Sri Lanka, Email: t.m.d.saumya@gmail.com
[2] Department of Industrial Management, University of Kelaniya, Sri Lanka, Email: thashika@kln.ac.lk
[3] National Cancer Institute, Maharagama, Sri Lanka, Email: prasadabeysinghe@hotmail.com

ABSTRACT

Classification technique plays an important role among other data mining techniques. Many real world problems in various fields have been solved using classification approach. Artificial Neural Networks have emerged as an important tool for classification which enables users in efficient classification of given data. This study will present how Artificial Neural Network approach is used in analyzing Leukaemic cancer patients' details for classifying them into two classes as survivors and non survivors depending on the identified prognosis factors. The study will be continued based on 120 pediatric Leukemia patients' data collected from Sri Lankan National Cancer Institute with regards to the identified prognosis factors. The study resulted in a multilayer artificial neural network which is capable of classifying Leukaemic patients' data with an accuracy of 88.23%.

***Key words*:** Data mining, Classification, Artificial Neural Networks, Leukemia, Cancer

1. INTRODUCTION

Advancement of information technology has also resulted in low cost hardware and software, with this low cost hardware and software the amount of data being collected and stored in databases (both in medical and in other fields) has increased dramatically in the last decade. As a result, traditional data analysis techniques have become inadequate for processing such volumes of data, and new techniques have been developed. The most prominent area of development is called knowledge discovery in databases (KDD). Knowledge Discovery in Databases (KDD) is the process of extracting high-level knowledge from low-level data and interpreting the discovered knowledge using visualization techniques.

KDD is well equipped with variety of statistical analysis, pattern recognition and machine learning techniques. In general, KDD can be defined as a formal process which contains the steps of understanding the domain, understanding the data, data preparation, gathering and formulating knowledge from pattern extraction, and visualization and demonstration of knowledge discovered which are employed to exploit the knowledge from large amount of recorded data .The step of gathering and formulating knowledge from data using pattern extraction methods is commonly referred to as data mining.

As the above text is explaining data mining is the process of automating the information (knowledge) extraction which is an important sub task in knowledge discovery in database process. As previously mentioned, data mining plays an important role in the KDD due to its nature of interdisciplinary. Its main aim is to uncover relationships in data and to predict outcomes. Also data mining helps to extract patterns in the process of knowledge discovery in databases in which intelligent methods are applied. The emerging field of data mining promises to provide new techniques and intelligent tools which help the human to analyze and understand large bodies of data remains on difficult and unsolved problem. The common functions in current data mining practice include Classification, Regression, Clustering, Rule generation, Discovering association rules, summarization, dependency modeling, and sequence analysis.

Classification is one of the important techniques of data mining which can be is used to classify each item in a set of data item into one of predefined set of classes or groups. Classification mainly deals with two types of variables.

Classification will require two data sets called training data set and testing data set. The input to the classification problem is a data-set called the training set having a number of attributes. The attributes are either continuous or categorical. One of the categorical attributes set is known as

classifying attribute which will that define the class of an instance or record. And the other set is known as predicting attributes which will determine the class of an instance in the data set. The objective is to use the training set to build a model of the classifying attribute based on the predicting attributes such that the model can be used to classify new data not from the testing data-set. Various classification problems can be handled effectively by multiple soft computing data mining techniques. These techniques are fuzzy logic, neural networks, genetic algorithms, decision trees and rough sets, which will lead to an intelligent, interpretable, low cost solution than traditional techniques.

Artificial Neural Network (ANN) is one of the most used data mining method to extract patterns in an intelligent and reliable way and has been greatly used to find models that describe data relationship in classification problems. [12, 13] In view of the above said significant characteristics of ANN, this technique is adopted in this study for cancer data classification.

Data mining techniques have been widely used in diagnostic and health care applications because of their predictive power. Data mining algorithms can learn from past examples in clinical data and model the oftentimes non-linear relationships between the independent and dependent variables. The resulting model represents formalized knowledge, which can often provide a good diagnostic opinion. In this study the neural network approach to generate efficient classification rules is proposed. To perform classification task of medical data, the neural network is trained using Back propagation algorithm. As the structure of neural network is convenient for parallel processing, the output at each neuron in different layers is calculated in parallel. The performance of the network is analyzed with various types of test data. . Since this paper mainly focus on cancer data analysis using artificial neural network, it is better to get an understanding about applications of ANN in medical domain, especially in cancer field.

2. LITRETURE REVIEW

2.1. Artificial Neural Networks in Medical Field

Neural networks are known to produce highly accurate results in practical applications. Neural networks have been successfully applied to a variety of real world classification tasks in industry, business and science.[14] Also they have been applied to various areas of medicine, such as diagnostic aides, medicine, biochemical analysis, image analysis, and drug development. They are used in the analysis of medical images from a variety of imaging modalities. Applications in this area include tumor detection in ultra-sonograms, detection and classification of micro calcifications in mammograms, classification of chest x-rays, and tissue and vessel classification in Magnetic Resonance Images. Artificial neural networks provide a powerful tool to help doctors analyze, model, and make sense of complex clinical data across a broad range of medical applications. [1, 2,3,5, 8, 9]

As the volume of stored data increases, data mining techniques assume an important role in finding patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities. Neural networks can be used to extract rules from a disease classification. From the rules system so discovered, we can predict if someone will have a particular stage of a particular disease.

2.2. Artificial Neural Network

A Neural Network (NN) consists of many Processing Elements (PEs), loosely called “neurons” and weighted interconnections among the PEs. Each PE performs a very simple computation, such as calculating a weighted sum of its input connections, and computes an output signal that is sent to other PEs. The training (mining) phase of a NN consists of adjusting the weights (real valued numbers) of the interconnections, in order to produce the desired output. [11] The Artificial Neural Network (ANN) is a technique that is commonly applied to solve data mining applications. Neural Network is a set of processing units when assembled in a closely interconnected network, offers rich structure exhibiting some features of the biological neural network. The structure of neural network provides an opportunity to the user to implement parallel concept at each layer level. Another significant characteristic of ANN is fault tolerance. ANNs are well suited in situations where information is noisy and uncertain. ANN are an information processing methodology that differs drastically from conventional methodologies in that it employ training by examples to solve problem rather than a fixed algorithm.[4,6] They can be divided into two types based on the training method: supervised training and unsupervised training. Networks that are supervised require the actual

desired output for each input whereas an unsupervised network does not require the desired output for each input. A key feature of neural networks is an iterative learning process in which data cases are presented to the network one at a time, and the weights associated with the input values are adjusted each time. [11] After all cases are presented, the process often starts over again. During this learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of input samples. Once a network has been structured for a particular application, that network is ready to be trained. To start this process, the initial weights are chosen randomly. Then the training or learning, begins. The most popular neural network algorithm is back-propagation algorithm. Although many types of neural networks can be used for classification purposes. [10] The focus is on the feedforward multilayer networks or multilayer perceptrons which are the most widely studied and used neural network classifiers. The feedforward, back-propagation architecture was developed in the early 1970's. This back-propagation architecture is the most popular, effective, and easy-to-learn model for complex, multi-layered networks. Its greatest strength is in non-linear solutions to ill-defined problems. The typical back-propagation network has an input layer, an output layer, and at least one hidden layer. There is no theoretical limit on the number of hidden layers but typically there are just one or two. Some work has been done which indicates that a maximum of five layers (one input layer, three hidden layers and an output layer) are required to solve problems of any complexity. Each layer is fully connected to the succeeding layer. Training inputs are applied to the input layer of the network, and desired outputs are compared at the output layer. During the learning process, a forward sweep is made through the network, and the output of each element is computed layer by layer. The difference between the output of the final layer and the desired output is back-propagated to the previous layers, usually modified by the derivative of the transfer function, and the connection weights are normally adjusted. This process proceeds for the previous layers until the input layer is reached. [7]

3. EXPERIMENT

In this experiment the medical data related to Leukaemic cancer is considered. Data source was identified as the National Cancer Institute (NCI) of Sri Lanka and the data will be collected from the individual Leukaemic patients' files from their file repository. Mainly data will be collected from the patients' who are below 16 years, from them also the information will be collected from the patients' who have survived for 5 years (diagnosed in 2009 and still alive patients) and from the patients who have not survived for 5 years' time. This study mainly concerns classification of person into five year survivor and five year non-survivor of Leukaemia cancer in order to predict the five year survivability of that particular instance.

The information needed to collect from each individuals for classification purpose are the factors which are affecting the survivability which are known as prognosis factors and these were mainly identified from the literature review and based on discussion had with the oncologists at the NCI.These factors are validated by experts and also these are specially focused to the Sri Lankan context and these factors are the predicting attributes used for classification in ANN technique. Summary of the target data selected for ANN application can be represented as follows (table1):

•Number of instances: 120

•Number of classes : 2 [class 0 –Non Survivors, class 1 - Survivors]

Table 1: Data source summary

Predicting Attribute Name	**Description**
Gender	Female(0), Male(1)
Response to treatment	Low Risk (0), Standard Risk(1), High Risk (2), Very High Risk(3)
Residual Disease	RD-(0), RD+(1), RD++(2)
Type of Leukaemia	B cell (0), T cell(1)
Chromosome Translocations	Not Detected (0), Detected(1)
LP Test Results	Negative (0), Positive(1)
Testicular Test Results	Negative (0), Positive(1)
White Blood Count at Diagnosis(mm^3)	Lower than 2000 mm^3(0), Between 2000-5000 mm^3(1), Between 5000-20000mm^3(2), Between 20000-

	50000mm^3(3)
Distance to Residence(km)	Distance bellow than 20km(0), Distance between 20 and 100km(1), Between 100km and 200 km(2), Between 200km and 300 km(3), Above 300km 2 9.8- 69.6 km(4)
Age at Diagnosis	Lower than 2 years(0), Between 2 and 9 years(1), Higher than 9 years(2)

This Leukaemic data set is a scenario where its' two types of classes, survivors and non-survivors are linearly inseparable, so this study will use multilayer perceptron in order to classify the five year survivability data of Leukaemic patients using ANN.

3.1. Training the Neural Network

In this experiment the neural network is trained with Leukaemic data set by using back propagation learning algorithm with momentum and variable learning rate. The input layer of the network consists of 30 neurons to represent each attribute, as the dataset consists of 30 categories of all 10 predicting attributes. The numbers of classes are two: 0 – non survivors and 1- survivors. The output layer consists of two neurons to represent these two classes. The description of the back propagation algorithm is specified in the above is used to train the neural network during the training process. Several neural networks are constructed with hidden layers and trained with Leukaemic dataset. Finally the most accurately classifying multilayer perceptron is used for classifying the testing data set. The adjusted weightages, bias terms and the activation functions between the layers can be tabulated as bellow tables (table 2 and 3):

Table 2: Activation functions summary

Between	Activation function
Input Layer and Hidden Layer	Hyperbolic tangent
Hidden Layer and Output Layer	Softmax

Table 3: Weight matrix of ANN multi perceptron model

Predictor		Predicted		
		Hidden Layer1	Output Layer	
		H(1:1)	[Survive=0]	[Survive=1]
Input Layer	(Bias)	-.214		
	[Age at Diagnosis(0)]	-.626		
	[Age at Diagnosis(1)]	.778		
	[Age at Diagnosis(2)]	.227		
	[White Blood Count(0)]	.299		
	[White Blood Count(1)]	-.108		
	[White Blood Count(2)]	-.433		
	[White Blood Count(3)]	-.543		
	[White Blood Count(4)]	-.023		
	[Type of Leukaemia(0)]	.583		
	[Type of Leukaemia(1)]	-.203		
	[Chromosome Translocations(0)]	-.093		
	[Chromosome Translocations(1)]	-.660		
	[LP Test Results(0)]	-.338		
	[LP Test Results(1)]	-.749		
	[Testicular Test(0)]	.064		
	[Testicular Test(1)]	.120		
	[Response to Treatment(0)]	-.103		
	[Response to Treatment(1)]	.370		
	[Response to Treatment(2)]	.301		
	[Response to Treatment(3)]	-.710		
	[Residual Disease(0)]	.967		
	[Residual Disease(1)]	-.930		
	[Residual Disease(2)]	-.675		
	[Residence(0)]	.396		
	[Residence(1)]	.119		
	[Residence(2)]	-.228		
	[Residence(3)]	-.161		
	[Residence(4)]	-.459		
	[Gender(0)]	-.097		
	[Gender(1)]	.667		
Hidden Layer1	(Bias)		-.395	.097
	H(1:1)		-.796	1.571

3.2. Performance of the Network

For testing the performance of the net various samples are collected as test data. The test data is given as the input to the trained network and the output of the net is calculated with the adjusted weights. The output of the net is compared with the target output to study the learning ability of the network for classifying the Leukaemic data set. The results are tabulated in bellow table 4

Table 4: Output summary

Classification				
Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	21	4	84.0%
	1	5	51	91.1%
	Overall Percent	32.1%	67.9%	88.9%
Testing	0	8	2	80.0%
	1	3	26	89.7%
	Overall Percent	28.2%	71.8%	87.2%
Dependent Variable: Survive				

4. CONCLUSION

Classification is an important problem in the rapidly emerging field of data mining. Many problems in business, science, industry, and medicine can be treated as classification problems. Owing to the wide range of applicability of ANN and their ability to learn complex and nonlinear relationships including noisy or less precise information, neural networks are well suited to solve problems in biomedical engineering. In this study neural network technique is adopted for classification of medical dataset. The experiment is conducted with Leukaemic dataset by considering the multilayer neural network modes. Back propagation algorithm with momentum and variable learning rate is used to train the networks. To analyze performance of the network various test data are given as input to the network. Parallelism is implemented at each neuron in all hidden and output layers to speed up the learning process. The result of this study are defining the methodology which should be followed for analyzing Sri Lankan Leukaemic patients' data depending on set of predicting attributes. Also the experimental results proved that neural networks technique provides satisfactory results for the classification task on medical data sets such as the Leukaemic data set used in the above study.

5. REFERENCES

[1] W. G. Baxt, "*Use of an artificial neural network for data analysis in clinical decision making*", Neural Comput., vol. 2, pp. 480–489, 1990.

[2] H. B. Burke, "*Artificial neural networks for cancer research*", Sem.Surg. Oncol., vol. 10, pp. 73–79, 1994.

[3] J. Cruz, and S. Wishart, "*Applications of Machine Learning in Cancer Prediction and Prognosis*", Cancer Informatics 2006:2 59–77Ullah, I, 2006.

[4] G. Cybenk, "*Neural Networks in Computational Science and Engineering*", IEEE Computational Science and Engineering, pp.36-42, 1996.

[5] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, Scatter/gather: a cluster based approach to browsing large document collection, 1992.

[6] K. A. Jain, Mao, and K. M. Mohiuddi, "*Artificial Neural Networks: A Tutorial*", IEEE Computers, pp.31-44, 1996.

[7] S. Haykin,S. "*Neural Networks – A Comprehensive Foundation*", Pearson Education, 2001.

[8] A. Kandaswamy, "*Applications of Artificial Neural Networks*", Proceedings of the Zonal Seminar on Neural Networks, Nov 20-21, 1997.

[9] A. Kusiak, K. H. Kernstine, J. A. Kern, K. A. McLaughlin, and T. L. Tseng, "*Data mining: Medical and Engineering Case Studies*", 2000.

[10] R. P. Lippmann, "*Pattern classification using neural networks*", IEEE Commun. Mag., pp.47–64, 1989.

[11] R. Rojas, "*Neural Networks: a systematic introduction*", Springer-Verlag, 1996.

[12] J. Shafer, R. Agarwal, and M. Manish, "*SPRINT: A scalable parallel classifier for data mining*", In Proc. Of the VLDB Conference, Bombay, India, 1996.

[13] S. Sohn, and H. D. Cihan, "*Ensemble of Evolving Neural Networks in classification*", Neural Processing Letters,Kulwer Publishers, 2004.

[14] B. Widrow, D. E. Rumelhard, and M. A. Lehr, "*Neural networks: Applications in industry, business and science*," Commun. ACM, vol. 37, pp.93–105, 1994.