Australian National University

Research School of Engineering
ANU College of Engineering and Computer Science

ASSIGNMENT COVER SHEET

THIS COVER SHEET SHOULD BE ATTACHED TO THE FRONT OF YOUR ASSIGNMENT WHEN IT IS SUBMITTED

STUDENT NAME: **Kiarie Ndegwa**

STUDENT ID NUMBER: **u4742829**

COURSE NAME: Bio-inspired computing.

COURSE CODE: COMP8420

DUE DATE: 5pm 24th April 2016

Submission of this assignment constitutes a declaration that:

- No part of this work has been copied from any other person's work except where due acknowledgement is made in the text; and
- No part of this work has been written by any other person except where such collaboration has been authorised by the course lecturer concerned, and
- No part of this work has been submitted for assessment in another course in this or another part of the University except where authorised by the lecturer(s) concerned.

Student Name: Kiarie Ndegwa
Date: 25th April 2016

Assignment 1: COMP8420

Kiarie Ndegwa

April 25, 2016

Contents

List of Tables

List of Figures

Abstract

The aim of this project is to test different neural net topologies and data preprocessing techniques that result in the most accurate binary classifier. The given data with which this network is trained and tested on is taken from the UCI machine learning Repository. It is labelled ”Pima Indians Diabetes Data set”. It contains 768 patient details each of which is comprised of 8 indicators and a binary prognosis indicating whether or not they had diabetes mellitus. Two network topologies namely a mulit-layered network and a cascade network are cross compared with the paper labelled ”ADAP learning algorithm to forcast the onset of diabetes mellitus.” The lowest error rate attained by this project is 0.21 and achieved by a cascade neural network with 2 hidden layers, comprised of 4 and 2 hidden nodes respectively. A result better than that achieved by the ADAP paper which recorded an error of 0.23.

1 Introduction.

This project makes use of the data set labelled ”Pima Indians Data set”, acquired from the UCI machine repository webpage. This is contains the records of 768 Pima female Indians of 21 years and older, recorded in 1988 by the National Institue of Diabetes and Digestive and Kidney Diseases.

The data set was picked as it contained no missing data attributes and allowed several neural net topogies to be tested and troubleshooted in full. The paper accompanying the data set [1] also contained an accuracy metric - specificity which was easy to benchmark against the results acquired by this report.

The 2 main neural network topologies tested in detail throughout this report consist of a multi-layer neural network and a cascade neural network - the other architecture tested was a sparse auto-encoder, however its specificity results were sub par in comparison, accordingly its implementation is not detailed throughout this report. Secondly the two topologies detailed were picked as they closely resembled the modified single layer neural network detailed in the paper accompanying the Pima Indians data set.

The paper [1] implemented a type of single layered neural network called the ADAP - Adaptive learning routine analog perceptron. The key difference between this and a traditional neural net, is the organization of data before it is fed into the net. Rather than feed the network with ”raw” data, the dataset is broken down into vectors representing different ranges within each variable.

This report is divided into 4 sections each of which is described briefly below:

1. Methodology: Details the experiment specifics such as implementation, data acquisation and code.

2. Results and discussion: The acquired data is then analyzed and cross compared with the ADAP paper.

3. Conclusion and future work: Ideas relating to accuracy classification improvement through various other topologies are explored in brief.

4. References: Containing the cited papers in this report.

2 Method

For the purposes of this report the data is validated through 10 fold k validation, and the error percentage measured on a test set containing 192 entries i.e. 10 % of the data set. Various topologies, i.e. numbers of hidden layers and their respective nodes are explored in an incremental fashion. A rule of thumb [3] [2] when exploring these architectures is to keep the number of nodes in each layer somewhere between the input layer and the output layer. Therefore the number of nodes used in each layer ranges in value between 8 and 2. So for example, should out of a choice of 8 nodes, 7 hidden nodes in one layer result in the best specificity, then the next layer tested will build on this - this process is continued until a no improvement given the previous layers is achieved.

2.1 Benchmark paper.

The paper used to benchmark this project is labelled, ”Using ADAP learning algorithm to Forecast the onset of Diabetes Mellitus”. The accuracy metric they use throughout this paper is based on their accuracy prediction on a test subset of 192 patient records [1]. They recorded a maximum accuracy of 76%. This paper used a slightly modified single layer neural network. The key difference between this architecture was the binary vector encoding input data before being fed into the input layer of the network.The topologies designed for this report are detailed below and accordingly try and replicate or outmatch these results of this ADAP network.

2.2 Data set description.

The data set used is comprised of 786 patients each of which has the following 9 attributes as listed below. It is mentioned at the UCI repository that this data set is complete and therefore contains no missing attribute values.

1. Number of times pregnant.

2. Plasma glucose concentration - taken from a 2 hours oral glucose tolerance test.

3. Diastolic blood pressure (mm Hg).

4. Triceps skin fold thickness (mm).

5. 2-hour serum insulin (mu U/ml).

6. Body mass index (weight in kg/ (height in m)^2)

7. Diabetes pedigree function.

8. Age(years).

9. Class variable (0 or 1)

This classification task the neural nets are taught to learn is to find -given 8 inputs-patients susceptible to Diabetes Mellitus.

2.3 Process of acquiring data.

The following steps are taken to acquire data from the neural networks tested throughout this report.

2.3.1 Data pre-processing

This involves finding the most suitable means of cleaning off out liers. A couple of approaches were initially taken, including performing Principle Component Analysis (PCA) analysis to compress the inputs, to finding the mean variance of the data set to normalizing the input data set. PCA analysis was however dropped as all data points corresponding to the 8 input data sets were all uncorrelated. Mean variance and normalization were therefore used to pre-process the data. 2 sets of error measurement containing these unprocessed data were therefore recorded.

2.3.2 Building up neural network topology.

10 fold cross validation was used throughtout this experiment. This was achieved via the use of an inbuilt matlab neural network function called, "**cvpartition**". The topology ranged in node number and hidden layers size from 8 to 2 nodes and 2 to 3 layers. The topologies were designed in an incremental manner. For instance given a theoretical 3 layer network. Should 2 nodes in the first network result in the lowest error accuracy, therefore the next layer is built atop this layer. Given that the next layer contains 4 hidden nodes with the lowest error accuracy, the final hidden layer is built atop this. This results in a small data set throughwhich various topologies can be explored.

2.3.3 Acquiring network accuracy.

For the purposes of this report the metric used in the ADAP paper is employed to cross check the acquired data set. The error accuracy is thusly measured throughout this experiment. This is achieved by taking the error ranges carried out on the test set during 10 fold cross validation. The average error accuracy is then recorded into the accompanying results table.

3 Results and Discussion

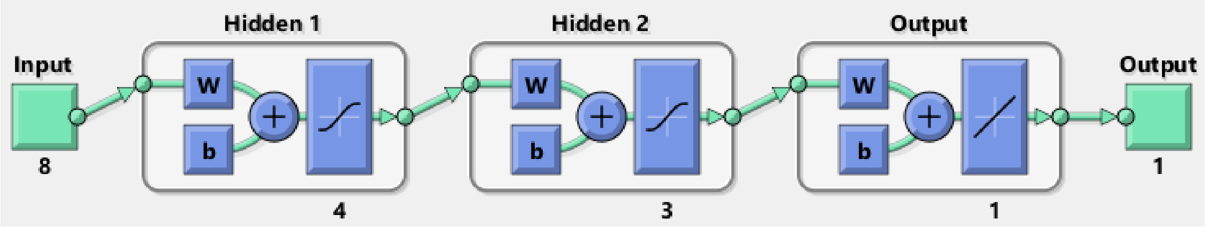
3.1 Deep Neural network

The following data set was acquired through the aforementioned testing regime.

No of hidden layers	No of Nodes in Layers	% Error **without** pre-processing	% Error **with** Pre-processing = variance and normalization
1	3	0.245	0.32
	4	0.230	0.32
	5	0.234	0.33
	6	0.250	0.30
	7	0.241	0.33
2, previous layer has 4 nodes for input and 6 nodes for processed input	3	0.224	0.31
	4	0.241	0.30
	5	0.253	0.32
	6	0.244	0.32
3, previous layer has 3 nodes for and 3 nodes for processed input	2	0.247	0.32
	3	0.237	0.32
	4	0.238	0.33
	5	0.247	0.31
	6	0.236	0.34

Deep neural network percentage error table.

As can be seen from the above results the best architecture given the data set is a network with 4 and 3 nodes in its first and second hidden layers respectively. This results in an error percentage of 22.4% This topology is demonstrated in the figure below:



2 layer neural network with highest tested accuracy.

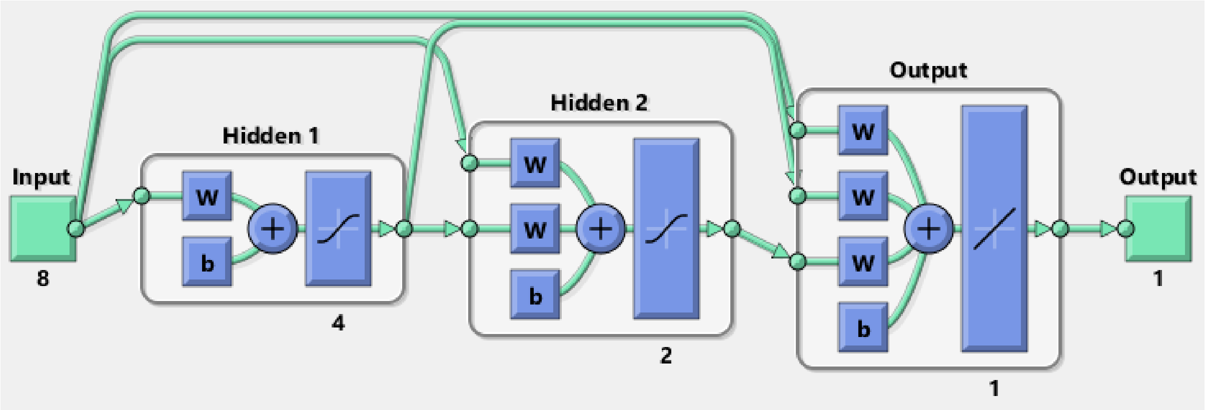
3.2 Cascade neural network.

The following data set was acquired through the aforementioned testing regime.

No of hidden layers	No of Nodes in Layers	% Error **without** pre-processing	% Error **with** Pre-processing = variance and normalization
1	3	0.23	0.29
	4	0.22	0.32
	5	0.23	0.30
	6	0.25	0.30
	7	0.241	0.31
2, previous layer has 4 nodes for input and 3 nodes for processed input	3	0.21	0.29
	4	0.230	0.33
	5	0.240	0.31
	6	0.240	0.32
3, previous layer has 3 nodes for both processed and unprocessed input	2	0.24	0.31
	3	0.25	0.31
	4	0.24	0.30
	5	0.24	0.31
	6	0.25	0.31

Cascade neural network percentage error table.

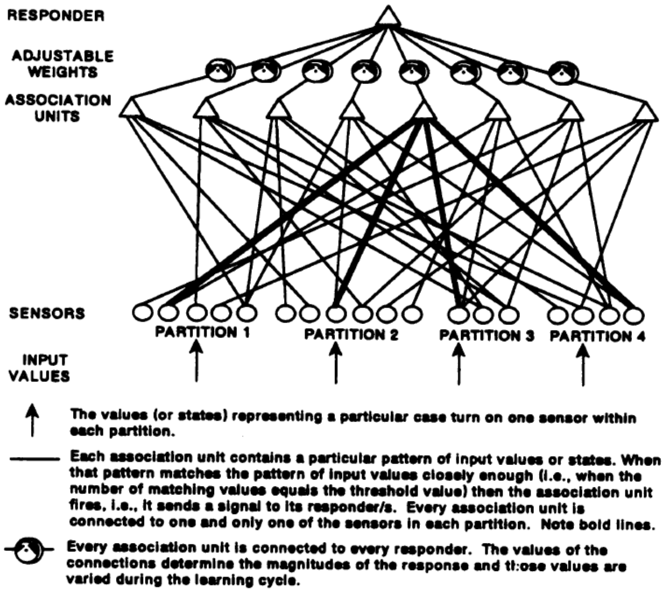
From the above it can be seen that the best cascade architecture is a 2 layered topology with 4 nodes in its first hidden layer and 2 nodes in its second hidden layer without pre-processed input data. This results in an percentage error of 21%. The cascade network described above is depicted in the diagram below:



2 layer cascade neural network with highest tested accuracy.

3.3 Comparison with ADAP paper.

The architecure employed in the ADAP paper is shown in the diagram below:



↑ The values (or states) representing a particular case turn on one sensor within each partition.

— Each association unit contains a particular pattern of input values or states. When that pattern matches the pattern of input values closely enough (i.e., when the number of matching values equals the threshold value) then the association unit fires, i.e., it sends a signal to its responder/s. Every association unit is connected to one and only one of the sensors in each partition. Note bold lines.

Every association unit is connected to every responder. The values of the connections determine the magnitudes of the response and tl:ose values are varied during the learning cycle.

ADAP single layer perceptron.

As can be seen the ADAP topology is similar to a single layered neural network with 4 nodes in its hidden layer. Key difference being in its input pre-processing whereby each of the 8 data input attributes are represented as different sized vectors representing different data ranges. The authors don't specifically mention why they used a single layer net or decided to partition their input data the way they did. However given that the paper was published in 1988, it can be inferred that they lacked the computational power to implement multi-layered networks.

This paper resulted in a an accuracy of 76%. Whereas from the above data set explored throughout this report and the corresponding network topologies, accuracy percentages of 77.6 % and 79% were achieved for the deep network and the cascade network, resulting in a classification improvement of 1.6% and 3% respectively.

4 Conclusion and Future Work

The highest classification accuracy achieved throughout this report was 77.6% and 79% using 2 layered deep networks and cascade networks respectively. This achieved classification improvements of Diabetes Mellitus by 1.6% and 3% respectively.

Given the limited data set size - 768 records and limited input variables of 8 medical measures, other architectures could not be fully explored such as stacked sparse autoencoders. A larger data set with a larger set of attribute variables would allow for the exploration of more types of deeper topologies and most likely result in a higher classification accuracy.

Another area that could have been explored had there been more time would've involved using an SVM classifier to classifiy the results from the cascade and multi-layered nets described above. However due to time limitations this area was not actively tested for the purposes of this report.

References

[1] Smith W. J., Everhart J.E., Dickson W.C., Knowler W.C., and DrPH R.S Johannes M.D. Using the adap learning algorithm to forecast the onset of diabetes mellitus. *The National Institute of Diabetes Digestive and Kidney Diseases: Epidemology and Data Systems programe and the Diabetes and Arthritis section and the John Hopkins University School of Medicine.*, 1988.

[2] Maciej A. Mazurowskia, Piotr A. Habasa, Jacek M. Zuradaa, Joseph Y. Lob, Jay A. Bakerb, and Georgia D. Tourassib. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Advances in Neural Networks Research: IJCNN 07*, 21(2-3):427–436, 2007.

[3] G. Baxt William. Use of an artificial neural network for data analysis in clinical decision-making: The diagnosis of acute coronary occlusion. *Neural computation.*, 2(4):480–489, 1990.