Prediction of Type II MODY3 Diabetes Using Backpercolation

Nawaz Khan
Computing Science
Middlesex University, UK
n.x.khan@mdx.ac.uk

Chukwuemeka A Ikejiaku
Computing Science
Middlesex University, UK
c.afam@mdx.ac.uk

Shahedur Rahman
Computing Science
Middlesex University, UK
s.rahman@mdx.ac.uk

Abstract

In this study, a neural network based approach is used to predict the presence of Maturity Onset Diabetes type 3, referred as MODY3 Type II diabetes mellitus. The study has used backpercolation neural network algorithm to predict the specific genetic mutation that causes the MODY3 type II diabetes mellitus. A set of coded numeric values are assigned for numeric representation of genetic data that are available in public domain repositories. A point mutation is introduced in a portion of the nucleotide for the mutation prediction to train the data set. The study has demonstrated that backpercolation neural network algorithm is useful to train and to predict gene point mutation that leads to MODY3 type II diabetes.

1. Introduction

In a recent study it is shown that diagnostic and predictive genetic testing is possible for the diabetes mellitus, especially in patients with MODY [3]. They have analysed a total of 116 families and have identified sixty five different mutations in *TCF 1* gene that cause MODY3. This study has revealed that by analysing genetic data and its changes it is possible to predict or diagnose diseases, but it requires application of intelligent neural network agent. Since then many approaches are emerged to implement intelligent neural network agent based algorithms to predict diseases associated to gene. For example, [2]. They have revealed that using neural networks it is possible to predict the length of survival period among patients with follicular lymphoma. The use of feed forward neural networks trained with backpropagaion are reported in several research, for instance, [4] and [6]. [1] has used neural network-based medical diagnosis system for the diagnosis of acute myocardial infarction which is difficult to diagnose. In this study, we have demonstrated how a supervised feed forward neural network model that utilises the BackPercolation algorithm ("BackPerc") as its learning algorithm can effectively predict the presence of genetic traits of diabetes mellitus and consequently its onset.

2. Approach

This research has attempted to implement the backpercoline algorithm for a number of reasons. Backpercolation technique helps to reduce the size of training cycles needed effectively. Moreover it demonstrates the following flexibilities: (i) training stability does not degrade when there are many hidden layers, (ii) training does not require non-local calculations (such as matrix inversion), (iii) training does not automatically increase the number of elements in the architecture and (iv) weights converge quickly toward attaining arbitrarily accurate output.

A number of steps are required to implement the algorithm: (i) a pattern is propagated through the network and the global error is computed, (ii) The gradient is computed and propagated back through the hidden layers, (iii) The error in the activation of each hidden neuron is computed. This error specifies the value by which the output of this

neuron has to change in order to minimize the global error, (iv) All weight parameters are changed according to internal errors and (v) If necessary, an adaptation of the error magnifying parameter is performed once every learning epoch.

The third step is divided into two phases: First each neuron receives a message, specifying the proposed change in the activation of the neuron. Then each neuron combines the incoming messages to an optimal compromise, the internal error of the neuron. The first phase is performed in forward direction (from input to output), the second phase goes backwards. The internal error of the output units is defined according to the global error magnification parameter. Backpercolation algorithm does not have a learning parameter, instead, it has an error magnification parameter. A threshold value can be set for the training and the error magnification error parameter may be adapted after each epoch when the total mean error of the network falls below the threshold value.

The following sections describe different parameters that have been implemented to apply the backpercolation algorithm.

Construction of network topology: Unlike backpropagation network topology, where the optimum topology are determined at the outset and backpropagation is proceeded to fine tune the network, this research has attempted to alter the topology dynamically in conjunction with the normal gradient descent.

Training the neural network: To predict the incidence of MODY3 Type 2 diabetes mellitus, we have collected a set of genetic data from online repository [5]. This data set is then compared with mutated genes that cause the disease. A set of expectation is constructed out of this data which is to be carried in training within the range of certain assumptions.

Benchmark comparison: In this study, we have chosen to supervised feedforward approach and training is stopped when it reaches to minimum absolute error.

3. Experiment and Results

This research has attempted to demonstrate the possibility of finding the presence of any genetic mutation that causes changes from CCG to CTG which indicates a change from amino acid *proline* to *leucine* [7]. A selected set of bases are represented in alphabets of the bases of the nucleotide and then converted to numbers as shown in Table 1.

Table 1. Alphabetic and numeric values of nucleotide bases

Nucleotide Bases	Alphabetic Representation	Numeric Representation
Adenine	A or a	1
Cytosine	C or c	2
Guanine	G or g	3
Thymine	T or t	4
Uracil or any other mutation		0

To optimise the generalisation the experiment is limited to single input, output and hidden layer. For the training size, 30 rows of training sets and 10 rows of testing set have been applied. The training of the neural network has been carried out with 10 epochs of pairs of inputs and outputs. The 10 inputs are received from the conversion of the bases from alphabets to numbers. For the experiment, the first 10 rows of the INPUT 3 are used as the testing set where the mutation is introduced in the 9th row. The idea is to train the network with the normal non-mutated genes and then test if it can detect the mutated gene. The network prediction which is presented in Table 2 was empty at the beginning of the experiment, since there were no predictions to be made at the beginning of the experiment.

Once the network is tested, the neural net supplied the results to the blank column. The experiment continued until a minimal average absolute error of -0.14 and error percentage of -8.38% was reached. The total training time calculated was 10mins which demonstrates the speed of the Backperc algorithm.

Table 2: Testing set and its corresponding network prediction

No.	Base sequence	Test Data	Net Prediction	Abs Error	% Error
1	gc**c**atggttt	2	1.90	0.10	5.24%
2	gggctgagca	3	3.09	-0.09	-3.06%
3	ggagaaggcc	1	1.26	-0.26	-25.69%
4	ctgcccaatg	3	3.09	-0.09	-3.06%
5	gacttcacgc	3	3.09	-0.09	-3.06%
6	cagaaagccg	3	3.09	-0.09	-3.06%
7	aagtcctacc	3	3.09	-0.09	-3.06%
8	aaccagtccc	2	1.90	0.10	5.24%
9	gc**t**gccctgt	**4**	3.53	0.47	11.86%
10	gcagggcagg	1	1.36	-0.36	-35.71%
				-0.14	-8.38%

Net prediction for mutation at number 10 = 3.53 gives rise to Abs error of 0.47 and a percentage error of 11.86%. From the results as shown in Table 2, it is clear that the neural net has learnt the pattern of the numbers, for example *c* which is represented with the test data number 2 and suddenly when it is changed to *t* in test data number 4 the absolute error was significantly high and in fact it was the highest. The value for the absolute error was 0.47 and the percentage error was 11.86% which are significantly higher compared to the rest. This suggests that a possible mutation has occurred here.

4. Conclusion

It is evident from the experiment and result that a possible point mutation for diabetes mellitus can be identified by using intelligent neural network based agent which utilises feed forward backpercolation algorithm. However, the degree of predictability depends largely on data that are fed to the network. Therefore, different types of mutations information and data regarding patients' history can increase the prediction rate by several folds.

5. References

[1] Baxt, W.G. (1990). "Use of artificial neural network for data analysis in clinical decision–making: The diagnosis of acute coronary occlusion," Neural Computation, 2 (4), 480-489.

[2] Dave S. et al. (2004). Prediction of Survival in Follicular Lymphoma based on Molecular Features of Tumor-infiltrating immune cells. The New England Journal of Medicine. Vol 351:2159-2169. no. 21.

[3] Ellard, S. (200), Hepatocyte nuclear factor 1 alpha (HNF-1-alpha) mutations in maturity-onset diabetes of the young. Hum. Mutat. 16: 377-385, PubMed ID: 11058894.

[4] El-Sloh AA, et al. (2004). Predicting active pulmonary tuberculosis using artificial neural network. URL: http://www.ncbi.nlm.nih.gov/entrez

[5] Entrez Nucleotide. (2004) Available at: http://www.ncbi.nlm.nih.gov/entrez/

[6] Shiau Y.C, Chou H.H. and et al. (2004). "Using Back Propagation Neural Network for Cardiac Wall Motion Detection on Angiocardiography," IEEE symposium on computer-based medical systems, p210- 215.

[7] Yamagata, K.et al. "Mutations in the hepatocyte nuclear factor-1-alpha gene in maturity-onset diabetes of the young (MODY3)." *Nature* 384:455-457,1996. PubMed ID : 8945470