Bulletin of Environment, Pharmacology and Life Sciences
Bull. Env. Pharmacol. Life Sci., Vol 4 [9] August 2015: 109-114

Online ISSN 2277-1808
Journal's URL:http://www.bepls.com
CODEN: BEPLAD
Global Impact Factor 0.533
Universal Impact Factor 0.9804

ORIGINAL ARTICLE **OPEN ACCESS**

Using Bayesian Network for the Prediction and Diagnosis of Diabetes

Mohtaram Mohammadi[1],Mitra Hosseini[2],Hamid Tabatabaee[3,*]
[1]Depatment of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran
[2]Depatment of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran
[3,*]Young Researchers and Elite Club, Quchan Branch, Islamic Azad University, Quchan, Iran

ABSTRACT
There are many feedback loops in the human body which keep the biotic balance. Disability or malfunction of any of these loops causes severe diseases with short- or long-term complications. Diabetes is one such disease which is caused due to the imprecise operation of these natural loops in the body. Diabetes is ascribed to the acute conditions under which the production and consumption of insulin is disturbed in the body which consequently leads to the increase of glucose level in the blood. Bayesian networks are considered as helpful methods for the diagnosis of many diseases. They, in fact, are probable models which have been proved useful in displaying complex systems and showing the relationships between variables in a graphic way. The advantage of this model is that it can take into account the uncertainty and can get the scenarios of the system change for the evaluation of diagnosis procedures. In this study, decision tree and Bayesian models have been compared. The results indicated that the Bayesian model is much more accurate in diabetes diagnosis.
***Key Words**: Prediction; Bayesian Model; Diabetes*

Received 11.05.2014 Revised 13.05.2014 Accepted 23.05.2015

INTRODUCTION

Diabetes, known as the silent murderer, is one of the most dangerous diseases in the present century. This disease is considered as a substantial threat for the public health of society in both developed and undeveloped countries. At the present moment, diabetes is the fourth reason of death in the majority of the developed countries. Diabetes is a chronic disease which occurs when the body cannot produce the adequate amount of insulin the cells require or when the body cannot make an efficient use of the insulin produced. Both of these two cases lead to the increase of glucose level in the blood. Diabetes is of two major types: Type 1 and Type 2. The body of those people who are affected by the first type produces little or no insulin so they need to inject it to their bodies. This type of diabetes mostly occurs in children and teenagers. Type two diabetic people are those whose bodies cannot use the required insulin in efficient ways. This type which is more dangerous and more widespread occurs in people who are over 30 years old [1].

Millions of people are stricken by this disease all over the world. Unfortunately, many of these people are not informed of their problem and while time is a vital factor for the prediction and treatment of diabetes, it may take a long time for them to know about it. As important complications of diabetes heart attack, blindness, nephrogenic problems, blood pressure and neural hurts can be named among many others. Diagnosing diabetes and gaining information about the probability of being stricken by it is a challenging job because it displays various emblems some of which exist in other diseases as well. A physician, therefore, should check the results of the patient's medical experiments and work on the decisions he has made for his previous patients with similar situations. The physician, in other words, needs both knowledge and experience. But because of the large number of patients and various medical experiments, the need for an automatic tool is felt to do a search among patients stricken by heart diseases [2]. An important way used for data inference is the Bayesian network.

In this article, the aim is to implement the Bayesian network, decision tree and the MATLAB software, to compare these models by the use of measuring scales on a data set collected from diabetic women and to

determine the most accurate and optimum model. The comparison of the two methods of decision tree and Bayesian model indicates that the Bayesian model is superior with regard to accuracy and precision. The data set in this paper is comprised of the information gained from diabetic women who have different situations.

DATA PREPROCESSING

Much of the data used for data-analysis were not collected with a specific purpose. Some of this data may include errors or missing values. In order to use this data set in data-analysis procedure, data normalization and data discretization is needed. The data set used in this study is made of information about diabetic women.

TABLE I: THE ATTRIBUTES OF PIMA DATASET

Attribute No.	Attribute	Description	Type
X1	PREGANAT	Numbers of time pregnant	Numeric
X2	GTT	Plasma glucose concentration in an oral glucose tolerance test	Numeric
X3	BP	Diastolic blood pressure(mmHg)	Numeric
X4	SKIN	Triceps skin fold thickness(mm)	Numeric
X5	INSULIN	Serum insulin(µU/ml)	Numeric
X6	BMI	Body mass Index(kg/m)	Numeric
X7	DPF	Diabetes pedigree function	Numeric
X8	AGE	Age of patient(years)	Numeric
Y	DIABETE	Diabetes diagnose results("tested_ positive", "tested_ negative")	Nominal

Data Normalization

We first do a brief statistical analysis on the data. Table 2 indicates the mean and the standard deviation of each of these attributes. As seen in this table, the value range between the attributes is high.

Table2: The mean ± SD initial data problem

Attribute Number	Mean	Standard Deviation
X1	3.4	3.8
X2	32.0	120.9
X3	19.4	69.1
X4	16.0	20.5
X5	115.2	79.8
X6	7.9	32.0
X7	0.3	0.5
X8	11/8	33/2
Y	32/0	12/09

In this study, we use the MIN-MAX model to change the attributes to a new range. The used formula is:

$$(1)\ \mathbf{X} = \frac{MAX(X)-X}{X-MIN(X)}$$

The data is normalized in order to get better results.

Numerical Data Discretization

All attributes in Table 1 are numeric. In order to have the standard conditional probability of Bayesian networks, discretization of data is required. At first each Binary attribute is examined according to high and low values and then a numerical probability distribution is fit for each node. MATLAB software is a

set of machine learning algorithms used for data mining tasks [3] and includes tools for preprocessing, categorization, regression, clustering, association rules and visualization [4].

The Suggested Method

Many methods have been designed for the efficient deduction in Bayesian networks which aim to deduct accurately. The model we presented here compares Bayesian method with the decision tree.

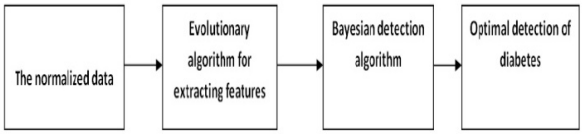

Figure1: Structure of the proposed system for the diagnosis of diabetes

Implementation

The Bayesian networks are good displays for the control of uncertainty. A Bayesian network is a directed no circle graph the vertexes of which are accidental variables and each vertex is a conditional distribution based on parent.

Bayesian Foundation

According to Table 1, from a training set of patient data, marginal probabilities of symptoms P (si) and disease P(dj) and conditional probabilities of symptoms on all diseases P(si/dj) are calculated by counting frequencies in the data. Given a set of symptoms (S={si}) for a patient, the posterior probability for each diagnosis is calculated as:

(2) $p(d_j|s) = p(d_j)\prod_{s_i \in S} \frac{P(s_i|d_j)}{p(S_i)}$

Using Formula number 2, we can think of an idea for designing the structure of Bayesian network. This formula determines the meaning of a Bayesian network. The conditional probability of Bayesian model is calculated by Formula number 3:

$$p(w|D,\alpha,\beta,p(w|D,\alpha,\beta,M) = \frac{p(D|W,\beta,M)p(w|\alpha,M)}{p(D|\alpha,\beta,M)}$$

Where W is the vector of network weights, α and β are function parameters, D represents the data vector, and M is the neural network used. Using Formula number 4 and 5, we can have the difference between the maximum and minimum amounts.

(4) $\frac{H_n - C_t}{H_n - l_n} \times 100$

(5) $\frac{C_t - L_{t-n}}{HH_{t-n} - U_{t-n}} \times 100$

The performance of the Bayesian networks is measured by computing the mean absolute percentage error (MAPE) which was first used for disease diagnosis and prediction by finding the absolute value of the derivation. This value is then divided by the diagnosis and multiplied by 100 to determine the percentage error.

(6) $\text{MAPE} = \sum_{i=1}^{r} \frac{(\text{abs}(y_i - p_i)/y_i)}{r} \times 100$

Where r is the total number of disease (i) is the disease diagnosis and P(i) is the prediction.

Bayesian Network Structure

A Bayesian structure can be made based on diabetes data. Pregnancy, age, DPF (Diabetes Pedigree Function) can be some of the effective factors on the appearance of diabetes. A considerable part of data set is related to two measures of obesity: SKIN (triceps skin fold thickness) and BMI (Body Mass Index) which we assume as a hidden variable in the network. Regarding the fact that skin fold thickness is not a good evidence of diabetes, BMI is considered as obesity value. Both GTT and insulin measurements are used for testing diabetes and cause diabetes.

Whether blood pressure is a reason for diabetes or not is a question. Following the experiments, it has been found that blood pressure is not a cause of diabetes. Pregnancy, age and obesity are all reasons for blood pressure. According to the presented analysis, Figure 1 indicates the main Bayesian structure in diabetes diagnosis.

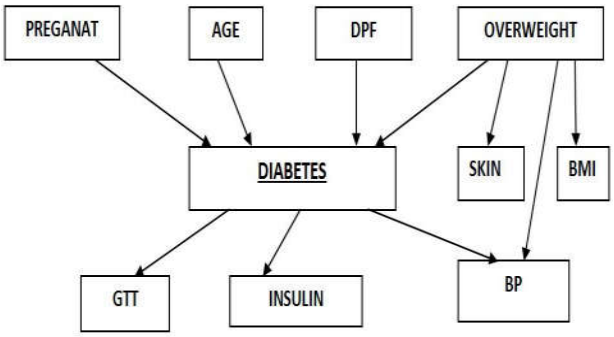

Fig. 2 Bayes Network [3]

RESULTS AND DISCUSSION

As Figure 1 displays, we used MATLAB software for the implementation of data and filled the conditional probability table with 768 samples which have been collected completely in compatibility with the standards of sanitary organization. In this data set, we have examined the background of 9 diabetic women who are at least 21 years old. In Table 3, the data related to these women who have different conditions is shown.

Table3: Attributes9diabeticfemale patients with different conditions

[1] RR	[2] GTT	[3] BP	[4] SKIN	**[5] INSULIN**	[6] BMI	[7] AGE	[8] result
[9] 6	[10] 148	[11] 72	[12] 35	[13] 0	[14] 33.6	[15] 50	[16] 1
[17] 1	[18] 85	[19] 66	[20] 29	[21] 0	[22] 26.6	[23] 31	[24] 0
[25] 8	[26] 183	[27] 64	[28] 0	[29] 0	[30] 23.3	[31] 32	[32] 1
[33] 1	[34] 89	[35] 66	[36] 23	[37] 94	[38] 28.1	[39] 21	[40] 0
[41] 0	[42] 137	[43] 40	[44] 35	[45] 168	[46] 43.1	[47] 33	[48] 1
[49] 5	[50] 116	[51] 74	[52] 0	[53] 0	[54] 25.6	[55] 30	[56] 0
[57] 3	[58] 78	[59] 50	[60] 32	[61] 88	[62] 31.0	[63] 26	[64] 1
[65] 10	[66] 115	[67] 0	[68] 0	[69] 0	[70] 35.3	[71] 29	[72] 0
[73] 2	[74] 197	[75] 70	[76] 45	[77] 543	[78] 30.5	[79] 53	[80] 1

For the first part of implementation, we have used the decision tree the graph of which is as the following:

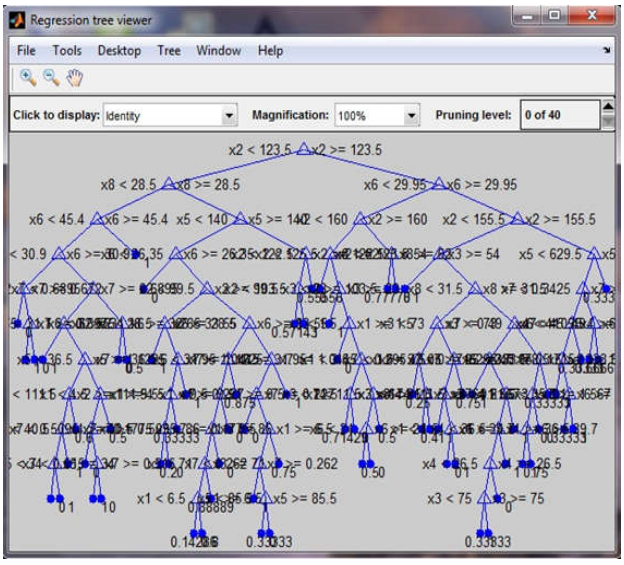

Figure2: The data in Table3, the MATLAB implementation of the method of decision tree

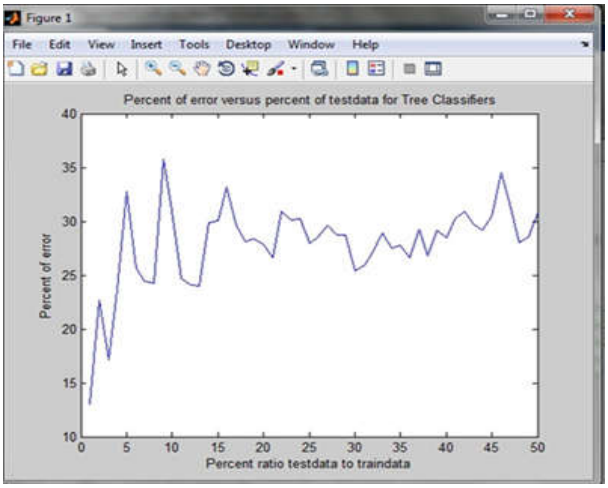

Figure3: Comparison of the percentage error percentage rate compared to experimental data using decision tree

In the second part of implementation, we have used the Bayesian network. The results related to this section are shown in Figure 4.

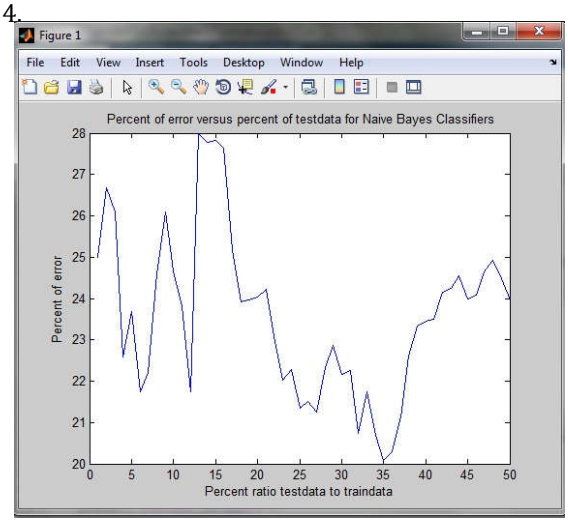

Figure4: Comparison of the percentage error percentage rate compared to experimental data using Bayesian

As seen in this figure, Bayesian graph has the fewest number of errors when the train data is about 35 percent while the training data have the best results in decision sentence method. The findings indicate that the precision and accuracy of results are higher when Bayesian method is used.

CONCLUDING REMARKS AND SUGGESTIONS FOR FURTHER RESEARCH

Based on the statistics presented by the international sanitary organization, diabetes is the world future disease. The complications of this disease can be extremely severe in the long run and may lead to heart attack, brain attack and blindness. Before discovering insulin, the quantity of diabetics' life was important but after discovering this hormone, the quality of their life became the focus of attention. There are too many complexities and uncertainties which compel us to use advanced controlling techniques. In this study, the aim was to design an algorithm which could help to diagnose diabetes having the fewest number of errors. Bayesian networks and decision tree are more effective in the diagnosis of many diseases because of their structures.

Bayesian network uses statistical calculations which are based on equations, Lagrangian functions and sample frequencies which lead to more precise and reliable results.

Future Studies: The use of more parameters may lead to gaining higher accuracy in diabetes diagnosis.

REFERENCES

1. Leung, Ross KK, et al. (2013).Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case control cohort analysis. BMC Nephrology 14.1.162.
2. Khajehei, Marjan, and Faried Etemady. (2010). Data Mining and Medical Research Studies" Computational Intelligence, Modelling and Simulation (CIMSiM), Second International Conference on. IEEE.
3. Patil, B.M.; Joshi, R.C.; Toshniwal, D.; (2010). Association Rule for Classification of Type-2 Diabetic Patients," Machine Learning and Computing (ICMLC), 2010 Second International Conference on , vol., no., pp.330-334, 9-11.
4. Guo, Yang, GuohuaBai, and Yan Hu (2012)."Using Bayes Network for Prediction of Type-2 Diabetes" Internet Technology and Secured Transactions, 2012 International Conference for. IEEE.

CITATION OF THIS ARTICLE

Mohtaram M, Mitra H, Hamid T. Using Bayesian Network for the Prediction and Diagnosis of Diabetes. Bull. Env. Pharmacol. Life Sci., Vol 4 [9] August 2015: 109-114