International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Learning Diabetes Data Using Bayesian Network and Examining the Factors that Contribute to Development of Diabetes Types I and II

S. M. Maibulangu[1], A. Bishir[2], R. M. Madaki[3]

[1, 2, 3]Department of Mathematical Sciences, Faculty of Sciences, Abubakar Tafawa balewa university, Bauchi, Nigeria

Abstract: ***Bayesian networks (BN) are an excellent tool for classification of a complex inter-correlated data such as medical data. This study is aimed at learning diabetes data using CI Search Algorithm and to determine the factors that cause Type 1 and Type 2 Diabetes. Diabetes dataset was obtained from medical records unit of the Abubakar Tafawa Balewa University Teaching Hospital (ATBUTH) Bauchi, Bauchi State consisting of 569 cases with 8 different variables. To achieved the stated objectives, Bayesian Network (BN) were used i.e. CI Search Algorithm. CI Search Algorithm explored the nature of relationship between the attributes and diabetes type (T1D/T2D). Conditional Probability Tables (CPT) obtained by CI Search Algorithm reveals that a diabetes patient of age between 45.5-72 years is 86% likely to be T2D, a diabetes patient of BMI between 24.6-38.6 is 89% likely to be T2D, a diabetes patient with history of diabetes is 64% likely to be T1D, a diabetes patient without history of diabetes is 66% likely to be T2D.***

Keywords: Nodes, Directed Acyclic Graph, Edge, Descendent and Ancestor, Conditional (mutual) information

1. Introduction

Diabetes, also called diabetes mellitus (DM) is one of the most dangerous diseases in the present century [5]. This disease is considered as a substantial threat for the public health of society in both developed and undeveloped countries. Diabetes is a chronic disease in which the body does not produce insulin or use it properly [5]. This increases the risks of developing, kidney disease, blindness, nerve damage, blood vessel damage and contribute to heart disease [1]. High blood sugar produces the classical symptoms of polyuria (frequent urination), polydipsia (increased thirst) and polyphagia (increased hunger) [5].

Although the reasons for developing diabetes are still not known, several important factors considered to contribute to the development of this disease have been identified which include: obesity, poor diet, physical inactivity, increasing age, family history of diabetes, ethnicity, poor nutrition during pregnancy affecting the developing child, [2]. Symptoms of diabetes are gradual but typically extreme thirst, frequent passing of water and heavy weight loss over a short period are the main symptoms. Others include fatigue, frequent infections, itching and rashes as well as disturbed vision. However, some people show none of these symptoms. As a result, most people remain undiagnosed for a long time until when complications of the disease become evident [8].

Bayesian networks provide efficient and effective representation of the joint probability distribution over a set of random variables. A Bayesian network is a graphical representation of the probabilistic relationships among sets of variables and is used for doing probabilistic inference with those Variables [2]. A Bayesian network consists of two qualitative components: The structure of the model which represents the set of conditional independencies among the variables in the data. The second component, the parameter, describes a conditional distribution for each variable. Models of entire systems can be built, rather than modeling the outcome of a one variable. Numerous variables of interest and their inter-relationship can be studied simultaneously, which is needed for complex models of diabetes dynamics.

[6], in their study, designed an algorithm which could help to diagnose diabetes having the smallest error rate. They discovered that Bayesian networks and decision tree are more effective in the diagnosis of many diseases because of their structures. [7] used a decision tree model and Bayesian network for the diagnosis of diabetes. His research found out that Classification with Bayesian network shows the best accuracy of 99.51 percent and error in the classification is .48 percent when the results were compared to a decision tree model.

A major problem in medical science is to determine the factors that cause of certain important disease. As a result, the need for an automatic tool is felt to do a search among diabetic class (T1D/T2D) of the patient. The Bayesian Network (CI Search Algorithm) is built in this research to give useful guides for detecting factors that causes diabetes (T1D/T2D) among diabetes patients. The aim of this research is to study diabetes dataset using Bayesian Networks and to determine the factors that cause Type 1 and Type 2 Diabetes. The following are specific objectives:

1) To fit the Bayesian Networks i.e. CI Search Algorithm.
2) To describe the probabilistic relationship between the attributes and diabetes type (T1D/T2D).

2. Material and Method

2.1 Representing distributions with Bayesian networks

Bayesian network is a representation of a joint probability distribution. This representation consists of two components. The first component, G, is a directed acyclic graph (DAG) whose vertices correspond to the random variables $X_1, X_2, \ldots, X_N$. The second component, θ, describes a conditional distribution for each variable, given

Licensed Under Creative Commons Attribution CC BY

its parents in G. Together, these two components specify a unique distribution on $X_1, X_2, \dots, X_N$. [4]

The graph G represents conditional independence assumptions that allow the joint distribution to be decomposed, economizing on the number of parameters. The graph G encodes the Markov Assumption: Each variable X_i is independent of its non descendants, given its parents in G.

By applying the chain rule of probabilities and properties of conditional independencies, any joint distribution that satisfies Markov Assumption can be decomposed into the product form

$P(X_1, X_2, \dots, X_N) = \prod_{i=1}^{n} (P(X_i \mid P_a(X_i))$ (1)

Where P_a is the set of parents of X_i in G. [4]

2.2 Conditional- and Mutual-Information

Conditional and mutual information is the information shared between a set of nodes. In a Bayesian network, if two nodes are dependent on each other, they share information. This information is called mutual information. This mutual information can be used to measure how close two nodes are related. The mutual information of two nodes X_i and X_j is defined as:

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \quad (2)$$

Where X_i, X_j returns the current values of the nodes, hence $P(x_i)$ returns the probability of the set of nodes X_i having values X_j.

The conditional (mutual) information is the mutual information between two nodes X_i and X_j, assuming the value for another node C is known. It is defined as:

$$I(X_i, X_j \mid C) = \sum_{x_i, x_j} P(x_i, x_j \mid C) log \frac{P(x_i, x_j \mid C)}{P(x_i \mid C)P(x_j \mid C)} \quad (3)$$

When $I(X_i, X_j \mid C)$ is smaller than a certain threshold ϵ, we say the nodes X_i, X_j are marginally independent given C. [3]

2.3 Continuous Data Discretization

Bayesian networks deal with nominal attributes only and are not suitable for analysis with continuous data. However, To use Bayesian Networks on general datasets, continuous attributes must first be "discretized" into a small number of distinct ranges using Equal-Interval Binning techniques. In this method of discretization, the algorithm divides a continuous attribute into

In this method of discretization, the algorithm divides a continuous attribute into K intervals of equal size. The width of the interval is:

$$w = \frac{Max - Min}{K} \quad (4)$$

Where Max denote the maximum value and Min denote the minimum value in a dataset.And the interval boundaries are:

$Min + w, Min + 2w, \dots, Min + (K-1)w$, Max

Where $Max = Min + KN$

2.5 Conditional Independence Test

The conditional dependencies present in a Bayesian Network are represented by directed arcs between nodes (variables).The lack of directed arcs among attributes represent a conditional independence relationship. This property makes graphical models a useful tool for identifying structure within data. This is advantageous if results are to be interpreted by individuals without a statistical knowledge.

This study investigates conditional independence among variable in diabetes data using Conditional Information-Based Search Algorithm (CI algorithm).

The CI-algorithm calculates the optimal network structure for a given set of nodes (variables) by making use of the mutual and information formulae. The algorithm works in 2 steps:

In the first phase, the algorithm computes mutual information $I(X_i, X_j)$ for each pair of nodes (X_i, X_j) using equation 2 described above. In the second phase, the algorithm adds edge for all the pairs of nodes that have mutual information greater than a certain small value ε. [3]

3. Data Analysis

The dataset used for this research work is a secondary data, which was obtained from medical records unit of the Abubakar Tafawa Balewa University Teaching Hospital (ATBUTH), Bauchi, Bauchi State. They were records of patients from March 2009 to April 2013. Basic information on the individual patients was extracted directly from the patients‟ respective files. Appendix I shows the diabetes dataset. In this research work, the total of 569 records were collected on diabetic patients for the periods of 2009 to 2013 understudy. Table 1 shows the summary of parameters extracted from patients‟ files.

Table 1: Data Set Description

Variable	Name	Description
X_1	Age	Age of patient(years)
X_2	BMI	Body mass Index(kg/m)
X_3	Sex	Sex of a patient
X_4	Marital	Marital status of a patient
X_5	History	Family history of diabetes
X_6	Exercise	Exercise status
X_7	Hypertensive	Hypertensive status.
C	Class	Diabetic class of the patient.

3.1 Conditional Independence Test

After data discretization this study investigated conditional independence among variable in the data using CI-Algorithm described in Chapter 3. We performed this independence test in WEKA package. The result of this algorithm is shown in the following diagram:

Licensed Under Creative Commons Attribution CC BY

Figure 1: CI Search Network for Diabetes Data

From Fig 1,we conclude that the variable C depends on $\boldsymbol{X_1}$(Age), $\boldsymbol{X_2}$(BMI) and $\boldsymbol{X_5}$(History). Also the variables $\boldsymbol{X_1}$(Age), $\boldsymbol{X_2}$(BMI) and $\boldsymbol{X_5}$(History) are conditionally independent given C (Diabetes).

The variable X_2(BMI) depends on X_3(Sex),X_4(Marital)and X_7(Hypertensive). AlsoX_2(BMI) depends on X_3(Sex), X_4(Marital) and X_7(Hypertensive) conditionally independent given X_2. Finally dependency exists between X_5(History) and X_6(Exercise). So inter-dependency exists in the variables.

Appendices I to VII show Conditional Probability Tables (CPT) of figure1which are summarized in Table 2 to Table 8.

Table 2: CPT for C

T1D	*T2D*
0.21	0.79

From Table 2, P(T1D)=0.21 and P(T2D)=0.79

Table 3: CPT for X1(Age)

C	11-38.5	38.5-41.5	41.5-45.5	45.5-72
T1D	0.55	0.22	0.13	0.10
T2D	0.01	0.04	0.09	0.86

From Table 3, the highest probability is 0.86 which indicates the chance of a diabetes patient of the age between 45.5-72 years to be T2D is 86%.

Table 4: CPT for X2(BMI)

C	8.7-24.6	24.6-38.6	38.6-40.4
T1D	0.49	0.43	0.10
T2D	0.11	0.89	0.00

From Table 4, the highest probability is 0.89 which indicates the chance of a diabetes patient of BMI between 24.6-38.6 to be T2D is 89%.

Table 5: CPT for X3(SEX)

BMI	Female	Male
8.7-24.6	0.43	0.57
24.6-38.6	0.48	0.52
38.6-40.4	0.85	0.15

From Table 5, the highest probability is 0.85 which indicates the chance of a diabetes patient of BMI between 38.6-40.4 to be a female is 85%.

Table 6: CPT for X4(Marital)

BMI	Singled	Married	Divorce
8.7-24.6	0.24	0.76	0.01
24.6-38.6	0.86	0.12	0.02
38.6-40.4	0.91	0.05	0.04

From the Table 6, the highest probability is 0.91 which indicates the chance of a diabetes patient of BMI between 38.6-40.4to be a singled is 91%.

Table 7: CPT for X5(History)

C	No	Yes
T1D	0.36	0.64
T2D	0.66	0.34

From Table 7, the probability 0.64 indicates the chance of a diabetes patient with family history of diabetes to be T1D is 64%. The probability 0.66 indicates the chance of a diabetes patient without family history of diabetes to be T2D is 66%.

Table 8: CPT for X6 (Exercising)

History	Exercising	Not Exercising
No	0.97	0.03
Yes	0.78	0.22

From Table 8, the probability 0.97 indicates the chance of a diabetes patient with family history of diabetes to be Exercising is 97%. The probability 0.78 indicates the chance of a diabetes patient without family history of diabetes to be Exercising is 78%.

4. Result and Discussion

4.1 Summary

The aim of this research work is to study diabetes dataset using Bayesian Networks and to determine their accuracy in detecting Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D).

The data used for this research work is a secondary data, which was obtained from medical records unit of the Abubakar Tafawa Balewa University Teaching Hospital (ATBUTH) Bauchi, Bauchi State. Total of 569 records were collected on diabetic patients for the periods of 2009 to 2013 understudy. Basic information (Age, Body Mass Index (BMI), Sex, Marital status, Family history, Exercise status, Hypotensive status and Diabetes type) on the individual patients have been extracted directly from the patients respective files.

Among the objectives is finding the relationship between variables which is achieved by Bayesian Network called CI Search Algorithm. The algorithm approach allows us to find relationships in data.

The result of CI Search Algorithm shows that diabetes depends on Age, BMI and History. The variable BMI depends on Sex, Marital status and Hypertensive status.

Finally the variable History depends on exercising status.

Also CI Search Algorithm was used in finding the Conditional Probability Distribution of the attributes. This distribution reveals the following findings;

Licensed Under Creative Commons Attribution CC BY

1) The chance of a diabetes patient of the age between 45.5-72 years to be T2D is 86%.
2) The chance of a diabetes patient of BMI between 24.6-38.6 to be T2D diabetes is 89%.
3) The chance of a diabetes patient with history of diabetes to be T1D is 64%.
4) The chance of a diabetes patient without history of diabetes to be T2D is 66%.

5. Conclusion

From the result obtained by CI Search Algorithm, one can draw the conclusion that diabetes type depend on Age, BMI and Family History of diabetes. From The Conditional Probability Table obtained by CI Search Algorithm, we conclude that a diabetes patient of age between 45.5-72 years is 86% likely to be T2D, a diabetes patient of BMI between 24.6-38.6 is 89% likely to be T2D, a diabetes patient with history of diabetes is 64% likely to be T1D, a diabetes patient without history of diabetes is 64% likely to be T2D is 66%.

6. Recommendations

Considering the diabetes dataset, there might be other risk factors that the data collections did not consider. It is therefore, suggested some other factors like gestational diabetes, metabolic syndrome, and smoking, should be include in further study to improve prediction accuracy.

References

[1] [1]Bellazzi, R. (2008). Telemedicine and diabetes management: Current challenges and future research directions.*Journal of Diabetes Science and Technology*, **2**(1):98-104.
[2] [2]Doctor, N.J. Strylewicz, G. (2010). *Artificial Intelligence in Medicine*, **50**:75-82.
[3] Hussein, I. (2009). An individual-based evolutionary dynamics model for networked social behaviors.In Proceedings of the American Control Conference, St. Louis 2009.
[4] Friedman, N., Geiger, D. and Goldzmidt, M. (1997). Bayesian Network Classifiers. *Journal of Machine Learning*, **29**:131–163.
[5] IDF (2012): *Diabetes Atlas*, Retrieved: April 4, 2014. Available at: **http://www.idf.org/fact-sheets.**
[6] Mohtaram, M., Mitra, H. and Hamid T. (2014). Using Bayesian Network for the Prediction and Diagnosis of Diabetes . *MAGNT Research Report*,**2** (5): 892-902.
[7] Mukesh, K., Rojan V., and Anshul A. (2014). Prediction of diabetes using Bayesian Network. *International Journal of Computer Science and Information Technologies*, **5**(4):5174-5178
[8] WHO (2013). Health topics: diabetes. Available at:http://www.who.int/topics/diabetes_mellitus/en/ Accessed on November 2,2014.

Appendices

Appendix I: Conditional Probability Table For Age

Probability Distribution Table For X1(Age)

C(Diabetes)	'(-inf-38.5]'	'(38.5-41.5]'	'(41.5-45.5]'	'(45.5-inf)'
0	0.55	0.219	0.128	0.103
1	0.012	0.043	0.087	0.857

Appendix II: Conditional Probability Table For BMI

Probability Distribution Table For X2(BMI)

C(Diabetes)	'(-inf-24.5496]'	'(24.5496-38.5726]'	'(38.5726-inf)'
0	0.494	0.436	0.071
1	0.112	0.885	0.003

Appendix III: Conditional Probability Table for Family History

Probability Distribution Table For X5(History)

C(Diabetes)	0	1
0	0.362	0.637
1	0.662	0.338

Appendix IV: Conditional Probability Table for Sex

Probability Distribution Table For X3(Sex)

X2(BMI)	0	1
'(-inf-24.5496]'	0.432	0.568
'(24.5496-38.5726]'	0.477	0.523
'(38.5726-inf)'	0.85	0.15

Appendix V: Conditional Probability Table For Marital Status

Probability Distribution Table For X4(Marital)

X2(BMI)	1	2	3
'(-inf-24.5496]'	0.24	0.756	0.005
'(24.5496-38.5726]'	0.859	0.123	0.019
'(38.5726-inf)'	0.905	0.048	0.048

Appendix VI: Conditional Probability Table For Exercising Status

Probability Distribution Table For X6(Exercise)

X5(History)	0	1
0	0.969	0.031
1	0.779	0.221

Appendix VII: Conditional Probability Table For Hypertensive Status

Probability Distribution Table For X7(Hypertensive)

X2(BMI)	0	1
'(-inf-24.5496]'	0.868	0.132
'(24.5496-38.5726]'	0.094	0.906
'(38.5726-inf)'	0.05	0.95

Licensed Under Creative Commons Attribution CC BY