http://dx.doi.org/10.5455/sf.85443

Science Forum (Journal of Pure and Applied Sciences)

journal homepage: www.atbuscienceforum.com

Performance of Bayesian networks (naive Bayes and tree augmented naive Bayes) in detecting diabetes types 1 and 2

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

Department of Mathematical Sciences, Faculty of Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria

ABSTRACT

This study is aimed at studying diabetes data using Bayesian networks and evaluating their performance in detecting types 1 and 2 diabetes. The diabetes dataset was obtained from the medical records unit of the Abubakar Tafawa Balewa University Teaching Hospital Bauchi, Bauchi State, consisting of 569 cases with 8 different variables. Classification of diabetes patients was carried out by naive Bayes and tree augmented naive Bayes (TAN) networks; the networks were trained and tested using a 10-fold cross-validation and the quality of prediction of these networks in terms of sensitivity, specificity, area under the ROC (AUR), kappa statistic, mean absolute error (MAE), and root mean square error (RMSE) was evaluated. The results indicate that TAN network proved to be the better network because the method correctly classified 530 (93.15%) and misclassified only 39 (6.85%) patients; it also has higher AUR (0.949) and kappa (0.7869), as well as lower MAE and RMSE of 0.1032 and 0.2384, respectively.

ARTICLE INFO

Article history:
Received 03 June 2021
Received in revised form
01 August 2021
Accepted 01 August 2021
Published 12 October 2021
Available online 12 October 2021

KEYWORDS

Nodes
Directed acyclic graph
Edge
Markov blanket
ROC curve

1. Introduction

Diabetes is a significant public health problem in Nigeria and many other parts of the world. Although diabetes prevalence has been rising for several decades, governments have been slow to implement policies that may have an impact at a population level (Bellazzi and Zupan, 2008). So many factors have been linked with diabetes, and they are highly correlated, but identifying relevant factors or at-risk population groups is difficult (Abed and Zaoui, 2011).

Bayesian networks (BN) are excellent tools for classifications of a complex inter-correlated data. BNs have been an important tool for decision-making and statistical inference in many fields such as medicine (Doctor and Strylewicz, 2010).

BN models can be applied to classification tasks, for example, in medical diagnostics (Bart et al., 2002) and evaluating economic risks (Kononenko, 2001). A BN model can be constructed from expert opinion, learned from data, or a combination of both (Bart et al., 2002).

Mohtaram et al. (2014), in their study, designed an algorithm which could help to diagnose diabetes having the smallest error rate. They discovered that BN and decision tree are more effective in the diagnosis of many diseases because of their structures.

Mukesh et al. (2014) used a decision tree model and BN for the diagnosis of diabetes. His research found out that the classification with BN shows the best accuracy of 99.51% and error in the classification is 0.48% when the results were compared to a decision tree model.

* **Corresponding author** Shehu M. M ✉ shehumm@gmail.com ✉ Department of Mathematical Sciences, Faculty of Sciences, Abubakar Tafawa balewa university, Bauchi, Nigeria

1.1. Statement of the problem

A major problem of medical data is the fact that many factors related to medical dataset are highly correlated (Nicholas, 2011), making identification of relevant factors difficult using the ordinary classification models such as logistic regression and discriminant analysis. These ordinary classification models may cause instable and inaccurate estimates' tests (Hoerl and Kennard, 1988).

On the other hand, BNs are an excellent tool for classification of complex inter-correlated data, and they explore the nature of relationship between variables in a given dataset (Flaxman, 2012).

In this research work, the dataset is assumed to be highly correlated. As a result of that, the BN is utilized in learning and predicting diabetes types 1 (T1D) and 2 (T2D).

1.2. The aim and objectives of the study

The aim of this research is to study the diabetes dataset using BNs and to determine their accuracy in detecting T1D and T2D.

The following are specific objectives:

1. To fit the BN, i.e., naive Bayes (NB) and tree augmented naive Bayes (TAN).
2. To determine the accuracy of predicting NB and TAN networks using root mean square error (RMSE), mean absolute error (MAE), kappa statistic, etc.

1.3. Data description

The dataset used for this research work is a secondary data, which was obtained from the 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.

In this research work, a total of 569 records were collected on diabetic patients for the periods of 2009–2013 under study. Table 1 shows the summary of parameters extracted from patients' files.

Table 1. Dataset description.

Variable	Name	Description
X_1	Age	Age of patient (years)
X_2	Body mass index (BMI)	BMI (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.

2. Methodology

2.1. Specification of the data

Before any analysis on a given dataset, one must specify the dependent (response) variables and the independent (predictor) variables.

Let C denote an individual status in respect of T1D and T2D level. In this study, the dependent variable of interest is C, which defines as dummy variable where

$$C = \begin{cases} 0, & \text{For } T1D \\ 1, & \text{For } T2D \end{cases}$$

Let $X = (X_1, X_2, \ldots, X_k)/$ be a vector to represent the parameters extracted from the patients' files. Let y_n denote an n^{th} individual record in the diabetes dataset. Each data point y_n has an associated class label in C, i.e., $y_n \in (0,1)$.

2.2. Continuous data discretization

BNs work with discrete data only and are not suitable for analysis with continuous data. However, to use BNs on general datasets, continuous attributes must first be transformed into discrete data, which can be achieved using the equal-interval binning techniques.

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

$$w = \frac{Max - Min}{K} \tag{1}$$

where Max denotes the maximum value and Min denotes the minimum value in a dataset.

And the interval boundaries are:

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

where $Max = Min + KN$.

2.3. Classification

In this research work, NB and TAN networks are used in predicting the class of new diabetic patients whose attributes are known.

2.4. Naive Bayes (NB)

NB is a popular classification method and used to fit our dataset in form of (Yang and Guohua, 2012):

$$P(C \mid X = x_1, x_2, \ldots, x_7) = \frac{P(C)P(X \mid C)}{P(X)} \quad (2)$$

$$P(C \mid X = x_1, x_2, \ldots, x_7) \alpha P(C) \prod_{i=1}^{7} (X_i|C) \quad (3)$$

For each attribute of the two classes, we need to estimate the following probabilities:

$$P(X|C=1) = \frac{P(C=1)\prod_{i=1}^{7}(X_i|C=1)}{P(X)} \quad (4)$$

And $P(X|C = 1) = 1 - P(X|C = 2)$ (5)

where

$$P(X|C=1) = \frac{No\ of\ times\ state\ of\ x\ occur\ if\ C=1}{number\ of\ cases\ in\ class\ C=1} \quad (6)$$

$$P(C) = \frac{No\ of\ times\ class\ C\ occurs}{total\ number\ of cases} \quad (7)$$

2.4.1. Classification rule

The classification task consists of classifying a variable C called the class variable given a set of variables $X = X_1 \ldots X_n$ called attribute variables. In this research, we classify a new patient input y_{n+1} as T2D if

$P(C = 2|X) > P(C=1|X)$ (8)

Otherwise, the patient is classified as T1D.

2.4.2. The tree augmented naive Bayes (TAN)

To learn the TAN network, we followed the following algorithm proposed by Friedman et al. (1997) and based on a well-known method reported by Pearl (1998).

TAN Algorithm:

{Input: *A* set of n nodes, a class node, and a database *D* containing m cases.}

{Output: *A* list of edges between pairs of nodes.}

1. Compute the conditional mutual information, $I(X_i, X_j \mid C)$, between each pair of variables, given the class variable. We set $X_1 to X_k$ as k attributes, C as the class variable

$$I(X_i, X_j \mid C) = \sum_{x_i, x_j} P(x_i, x_j \mid C) log \frac{P(x_i, x_j \mid C)}{P(x_i \mid C)P(x_j \mid C)} \quad (9)$$

2. Build a complete undirected graph in which the vertices are the attributes $X_i, \ldots X_n$,. annotate the weight of an edge connecting $X_1 to X_k$ by $I(X_i, X_j \mid C)$.
3. Use Kruskal's algorithm to build a maximum weight spanning tree.
4. Transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all edges to be outward from it.
5. Add an edge from the class node to each of the k nodes (David, 2014).

Based on the structure above, the parameters for conditional probability distributions can be estimated using the training dataset. Then we can classify the testing dataset by computing the highest probability.

Then we classify a new patient input y_{n+1} as T2D if

$P(C = 2|X) > P(C = 1|X)$ (10)

Otherwise, the patient is classified as T1D.

2.5. Examining the performance of the methods

Classification and prediction methods can be compared and evaluated according to the following methods:

2.5.1. *K*-fold cross-validation

In k-fold cross-validation, the original data are randomly partitioned into M mutually exclusive groups each of approximately equal size. Training and testing are carried out M times. In iteration i, partition $\boldsymbol{D_i}$ is reserved as the test set, and the remaining partitions are collectively used to train the model (David, 2014).

2.5.2. Confusion matrix

A confusion matrix contains information about actual and predicted classifications done by a classifier. The performance of such classifiers is commonly presented in the matrix called confusion matrix. Table 2 shows the confusion matrix for a two-class classifier.

True positives (TP) refer to the positive cases that were correctly classified by the classifier, while true negatives (TN) are the negative cases that were correctly classified by the classifier. False positives (FP) are the negative casas that were incorrectly labeled and false negatives (FN) are the positive outcomes that were incorrectly classified.

Table 2. Classification table for two groups.

Actual class	Predicted class	Sample size
C_1	TP FN	$n1$
C_2	FP TN	$n2$

2.5.3. Sensitivity

Sensitivity is also referred to as the TP (recognition) rate, (i.e., the proportion of positive samples that are correctly identified) and is defined as:

$$\text{TP rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

2.5.4. Specificity

Specificity is the TN rate (i.e., the proportion of negative samples that are correctly identified) and is defined as:

$$\text{TN rate} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{12}$$

2.5.5. Precision

We may use precision to access the percentage of cases labeled as positive that actually are positive. These measures are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

2.5.6. Kappa statistic

The kappa statistic is commonly used for accuracy assessment. Kappa can be used as a measure of agreement between model predictions accuracy and expected accuracy (Congalton, 1991). Kappa is computed as:

$$k = \frac{P_O - P_E}{1 - P_E} \tag{14}$$

Where P_E denote the expected accuracy and denote the predicted accuracy.

2.5.6. Predictor error measures

Loss functions measure the error between the actual value $\boldsymbol{p_i}$ and the predicted value $\boldsymbol{p_i'}$. The common loss functions used are:

$$\text{Absolute error} = \left| p_i - p_i' \right| \tag{15}$$

$$\text{Square error} = \left(p_i - p_i' \right)^2 \tag{16}$$

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^{m} \left| p_i - p_i' \right| \tag{17}$$

$$\text{MEA} = \frac{1}{m} \sum_{i=1}^{m} \left(p_i - p_i' \right)^2 \tag{18}$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(p_i - p_i \right)^2} \tag{19}$$

4. Results and Discussion

4.1. Software used for the study

WEKA TOOL: Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand.

4.2. Fitting BNs (NB and TAN)

Figure 1 shows a network structure of the diabetes data produced by the NB algorithm described. Directed edges between vertices represent dependency between the variables and lack of directed edge between vertices represent independency between variables.

Figure 2 shows a network structure of the diabetes data produced by TAN algorithm. Directed edges between vertices represent dependency between the variables and lack of directed edge between vertices represent independency between variables.

4.3. Performance of the BNs (NB and TAN)

In this work, in order to estimate the classification rate of the networks (NB and TAN), we used a 10-fold cross-validation for both training and testing the diabetes dataset. We first performed the classification of diabetes with NB and TAN using WEKA. Appendices 1 and 2 show the outputs of WEKA which are summarized in Table 3.

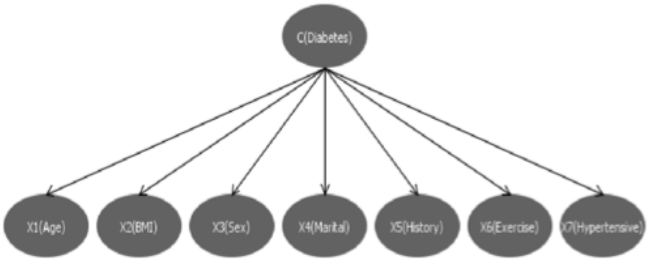

Figure 1. NB network structure for diabetes.

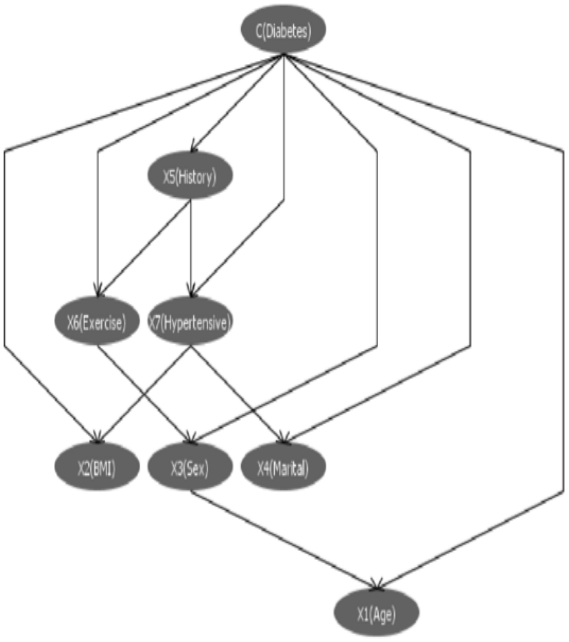

Figure 2. TAN network structure for diabetes.

Table 3. Performance of NB and TAN.

Performance measures	TAN	NB
Correctly classified	530 (93.15%)	505 (88.75%)
Misclassified	39 (6.85%)	64 (11.25%)
TP rate		
T1D	0.798	0.71
T2D	0.967	0.94
FP rate		
T1D	0.033	0.084
T2D	0.202	0.218
Precision		
T1D	0.864	0.782
T2D	0.948	0.916
ROC area	0.949	0.934
Kappa	0.7869	0.6721
MAE	0.1032	0.1258
RMSE	0.2384	0.2972

From Table 3, the following observations are made:

1. The highest accuracy of 93.15% belongs to TAN and the lowest accuracy of 88.75% belongs to NB.
2. The TP rate (0.967) and precision (0.948) of TAN is higher than that of NB.
3. The ROC area (0.949) and kappa (0.7869) of the TAN are higher than that of NB
4. TAN produced an algorithm which has a lower MAE (0.1032) and RMSE (0.2384) compared to NB.

5. Summary, Conclusion and Recommendations

5.1. Summary

The aim of this research work is to study the diabetes dataset using BNs and to determine their accuracy in detecting T1D and T2D.

The data used for this research work are a secondary data, which were obtained from the medical records unit of the ATBUTH Bauchi, Bauchi State. A total of 569 records were collected on diabetic patients for the periods of 2009–2013 under study. Basic information (age, BMI, sex, marital status, family history, exercise status, hypotensive status, and diabetes type) on the patients were extracted directly from the patients' respective files.

This research had studied two different methods of BN classification: NB and TAN. In terms of the classification accuracy, the model's performance was assessed from special cases of the 10-fold partitioning technique. And the research investigated the quality of prediction in terms of sensitivity and specificity, area under the ROC (AUR), and kappa statistic. The performance evaluation was carried out on the same diabetes dataset.

When studying the performance of NB and TAN on the diabetes dataset, the overall classification rate for both were good but TAN showed the highest accuracy of 93.15% and the lowest accuracy of 88.75% that belongs to NB. The TP rate (0.97) and precision (0.95) of TAN are higher than that of NB.

The ROC curves of the two models clearly indicate that the TAN is better with an AUR = 0.949. TAN produced an algorithm which has a lower MAE and RMSE of 0.1032 and 0.2384, respectively. According to kappa statistic, TAN classifier performs better with kappa = 0.7.

6. Conclusion

The classification of diabetes patients was carried out by NB and TAN networks. The networks were trained and tested using a 10-fold cross-validation and the quality of prediction of these networks in terms of sensitivity, specificity, AUR, kappa

statistic, MAE, and RMSE was evaluated. The results indicate that TAN network proved to be the better network because the method correctly classified 530 (93.15%) and misclassified only 39 (6.85%) patients; it also has a higher AUR (0.949) and kappa (0.7869), as well as lower MAE and RMSE of 0.1032 and 0.2384, respectively.

7. Recommendations

In this study, we only used few BNs to study diabetes. Considering the uncertain factors of some diabetes attributes, in the future work, more BNs and other machine learning methods should be investigated.

References

Abed H, Zaoui L. Partitioning an image database by K_means algorithm. J Appl Sci 2011; 11:16–25.

Bart B, Michael EP, Roberto C, Jan V. Learning Bayesian network classifiers for credit scoring using Markov. K. U. Leuven Dept. of Applied Economic Sciences, Leuven, Belgium, 2002.

Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008; 77:81–97.

Congalton RG. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens Environ 1991; 37:35–46.

David BH. Prepared lecture note (STAT J535). Springer Publication, New York, NY, 2014.

Doctor NJ, Strylewicz G. Detecting 'wrong blood in tube' errors: Evaluation of a Bayesian network approach. Artif Intell Med 2010;50:75–82.

Flaxman AD. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010. A systematic analysis for the global burden of disease study 2010. Lancet 2012; 21:63–96.

Friedman N, Geiger D, Goldzmidt M. Bayesian network classifiers. J Mach Learn, 1997; 29:131–63.

Hoerl A, Kennard R. Ridge regression. In: Encyclopedia of statistical sciences. Wiley, New York, NY, vol 8, pp 129–36, 1988.

Kononenko I. Machine learning for edical diagnosis: history, state of the art and perspective. Artif Intell Med 2001; 23(1):89–109.

Mohtaram M, Mitra H, Hamid T. Using Bayesian network for the prediction and diagnosis of diabetes. MAGNT Res Rep 2014; 2(5):892–902.

Mukesh K, Rojan V, Anshul A. Prediction of diabetes using Bayesian network. Int J Comput Sci Inf Tech 2014; 5(4):5174–8.

Nicholas JH. Application of Bayesian networks to problems within obesity epidemiology. Ph.D. Thesis, University of Manchester, Manchester, UK, 2011.

Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publisher Inc, San Francisco, CA, 1998.

Yang G, Guohua B. Using Bayes network for prediction of type- 2 diabetes. Yan Hu School of Computing Blekinge Institute of Technology Karlskrona, Karlskrona, Sweden, 2012.