SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
ISSN 2054-7390

Detection of the Onset of Diabetes Mellitus by Bayesian Classifier Based Medical Expert System

[1]Md. Mozaharul Mottalib, [2]Md. Mokhlesur Rahman, [3]Md. Tarek Habib and [4]Farruk Ahmed
[1]Department of Computer Science and Engineering, Green University, Bangladesh;
[2]Department of Computer Science and Engineering, Prime University, Bangladesh;
[3]Department of Computer Science and Engineering, Daffodil International University, Bangladesh;
[4]Department of Computer Science and Engineering, Independent University, Bangladesh
mm.mottalib@gmail.com; mmarks_cse@yahoo.com; md.tarekhabib@yahoo.com; farruk60@gmail.com

ABSTRACT

Expert systems play an important role in medical diagnosis research. Researches are still being conducted for building expert systems capable of diagnosing different diseases. Diabetes mellitus is one of the diseases that have gained attention in the past years. Patients are usually unaware of having this disease and are finally diagnosed with diabetes after several years from onset. Since diabetes can be controlled, it is much desirable to harness it at the onset. Therefore, the prediction of onset of diseases like diabetes has been the point of interest for the researchers. Researchers are continuously trying to formulate an inference engine, a part of an expert system, in order to predict the disease at the beginning. In this paper, we present a Bayesian classification approach to identify the onset of diabetes mellitus in patients using a well-known data set as the sample. We have found an intriguing result with more than 87% accuracy.

Keywords: Expert system, diagnosis of disease, pattern recognition, classification, Bayesian classifier.

1 Introduction

Expert systems are the special type of systems, which offer the solution of different problems through providing suggestions equivalent to human experts in those particular fields [1]. Real-life problems are solved using expert systems, specifically the problems that do not have predefined solution in general. Expert systems are used in the fields where sufficient amount of human expertise is required. Examples are the medical diagnosis of disease, financial advice, designing of products, etc. In the sector like the medical diagnosis of diseases, the inference engine of an expert system is built following several techniques. Pattern recognition is one of the well-known techniques.

Pattern recognition and data mining are used in different fields of our life. Especially, these techniques are more frequent in military, medical and industrial areas. With the passage of time, data collection and analysis have been drastically improved. New technologies contributed in these aspects. Possibilities of new researches are dramatically increased.

Since data analysis has traversed a lot of advancements and is continuously experiencing more, analysis of diseases in medical sectors got growth as well. Researchers have been doing research on how to predict the onset of diseases before any harm occurs or how to minimize the adversities.

DOI: 10.14738/tmlai.44.1962
Publication Date: 19th July, 2016
URL: http://dx.doi.org/10.14738/tmlai.44.1962

However, diabetes is one of the diseases those are under-diagnosed [2]. About one-thirds of the patients with diabetes are not aware of their having it. An average of 7years is the period between onset and diagnosis. Diabetes is a disease that needs constant monitoring and it could substantially decrease the life quality. As other diseases, early diagnosis of diabetes is crucial and can reduce the harm inflicted on the body.

We selected the Pima Indians Diabetes Dataset, available in [3], for building our model to predict the onset of diabetes mellitus. The description of the data set is described later.

We arranged the rest of the paper as follows. Section II discusses the related works in this field. Section III is used for describing the methodology of our classification model. Section IV contains the implementations. Experimental result and comparative analysis are given in the section V and finally, we conclude our work by giving future direction in section VI.

2 Literature Review

As it has been noted that researchers are continuously working on predicting diseases at onset, there have been several works in this field. Some works are related to the diagnosis of diseases while many are done on diabetes itself.

A method of biomedical signal classification using complex-valued pseudo autoregressive (CAR) modelling approach was proposed in [4]. An improvement on traditional multilayer perceptron (MLP) has been proposed in [5]. Solely on diabetes, work has been done using principal component analysis (PCA) and adaptive neuro-fuzzy adaptive systems in [6]. Generalized discriminant analysis (GDA) and least square support vector machine (LS-SVM) was used and a new cascade learning system based on that GDA and LS-SVM was proposed in [7].H. Kahramanli et al. presented a hybrid neural network that includes artificial neural network (ANN) and fuzzy neural network (FNN) [8].

Jack W. Smith et al. used ADAP learning algorithm to discern the onset of diabetes mellitus [9]. Neural network is also used for diabetes mellitus prediction in [10]. Again, Bayesian network was implemented for prediction of type-2 diabetes in [11]. Besides these, there are more works on disease classification and also on diabetes prediction. We only discuss the works that closely resembles our work.

3 Methodology

Our approach starts with building a model. After building the model, we train the model with the help of training dataset. Then applying the model in test samples we retrieve some predictions. These predictions tell us whether a patient or sample is in risk of diabetes mellitus foreknowing the onset. Figure 1 depicts the total overview.

3.1 Expert Systems

Expert systems consist two essential components: knowledge base and inference engine. A knowledge base is a repository for the domain-relevant knowledge. Algorithms for handling the knowledge base are the main attributes of an inference engine. A specific expert system is developed using resources from various knowledge banks, such as human experts, textbooks, and databases. In our case, the source is Pima Indian Database.

The knowledge base is subject to changes. Since our concern is not the knowledge base but the inference engine, we built the knowledge data set from the above mentioned source.

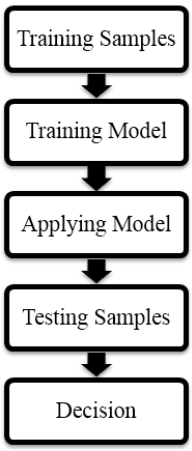

Figure 1. The approach for building the model for predicting the onset of diabetes mellitus.

Figure 2 demonstrates the simple architecture of a global expert system. We are focusing on the inference part. The inference engine we built follows the Bayesian theorem for predicting disease, in this case, diabetes mellitus.

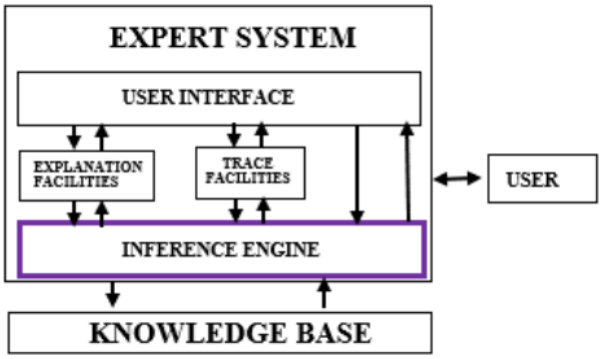

Figure 2. Global architecture of an expert system.

3.2 Dataset Description

The Pima Indian population who resides near Phoenix, Arizona was the research population. The National Institute of Diabetes, Digestive, and Kidney Diseases has been constantly studying the population since 1965 because of its high incidence rate of diabetes [12]. Community residents over the age of 21 years were asked to undertake a standardized test every two years. The test involved an oral glucose tolerance test and some other bodily measurements. Diabetes was diagnosed as per World Health Organization (WHO) Criteria [13]. That defined, if the 2-hour post-load plasma glucose was at least 200 mg/dl (11.1 mmol/l) at any survey test or if the Indian Health Service Hospital serving the community found a glucose concentration of at least 200 mg/dl during the time period of routine medical care. The database is being used for study by the researchers since its formation. Moreover, this data set provides a well-validated data resource to delve into the prediction of the date of onset of diabetes with various techniques.

Set of Features

There are eight features with samples 768 in total. The binary target values 0 and 1 represented 'tested negative' and 'tested positive'

i. *Number of times of pregnancy*
ii. *Plasma glucose concentration at 2hours in an oral glucose tolerance test (GTT)*
iii. *Diastolic blood pressure (mmHg)*
iv. *Triceps skin fold thickness (mm)*
v. *2-hour serum insulin (mu U/ml)*
vi. *Body mass index, i.e. weight/height^2 (Kg/ m^2)*
vii. *Diabetes pedigree function*
viii. *Age (years)*
ix. *Class variable (0 or 1)*

The data set contained all the features including a class variable to defining the class of each sample. There are 500 positive cases and the rest 268 are negative.

3.3 Bayesian Classifier

Using traditional Bayes theorem, Bayesian classifier classifies unknown samples based on maximum likelihood. This is a parametric computational model for solving classification problems. The Bayes theorem states the relationship of events as in (1).

$$P(event\ E|event\ G) = \frac{P(event\ E|event\ G)P(event\ G)}{P(event\ E)} \quad (1)$$

- $P(E)$ and $P(G)$ are the probabilities of event *E* and *G* without regard to each other.
- $P(E \mid G)$, a conditional probability, is the probability of event *E* given that event *G* is true.
- $P(G \mid E)$, a conditional probability, is the probability of event *G* given that event *E* is true.

In the case of naïve Bayesian classifier, the attributes are assumed independent. Not only it is fast and easy to compute but also, it is not sensitive to irrelevant data. From Bayes theorem, it can be deduced that as in (2).

$$P(G|E) \approx P(E|G)P(G) \quad (2)$$

Naïve Bayesian classifier is reliable for classification and has been used for many years. Since the classifier assumes the attributes independent, they do not affect the probability of each other. On the other hand, there are some adversities like, bias, variance and training data noise. Training data noise is sometimes the result of feature extraction and can be reduced by selecting distinguishing features.

The formal definition of Naïve Bayesian classifier concisely stands as in (3):

$$C = arg \max_{k \in \{1,2,\ldots,K\}} P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \quad (3)$$

where,

- $P(C_k)$ = probability of class k (prior probabilty)

- $P(x_i|C_k)$ = probability of feature x_i given class k

(class-conditional probability)

Now, since the attributes are continuous in this case, the likelihood of each class was calculated using the probability density estimations of the attributes. Assuming the distribution as normal, the density estimation function was defined as in (4).

$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4)$$

Density function expresses the relative probability of a point where μ is the mean, σ is the standard deviation, $\varphi_{\mu,\sigma}(x)$ is used for calculating $P(x|C_k)$.

The model training was accomplished by estimating the prior $P(C_k)$ and for every attribute A_i, for every attribute value v of A_i estimating $P(A_i = v|C_k)$.

Applying the probabilistic model for a given sample with $(v_1, v_2, v_3, \ldots, v_n)$, the class picked that maximized the value of (5).

$$P(C_k)\prod_{i=1}^{n} P(A_i = v_i|C_k) \quad (5)$$

Algorithm 1 describes our Bayesian classification algorithm.

Algorithm 1: Bayesian classification

i. Input training dataset where each feature vector X_i consists of the feature set $\{x_{i1}, x_{i2}, \ldots, x_{i8}\}$.

ii. Calculate the prior probability $P(C_k)$ for each class k, where k = 0, 1.

iii. Calculate the class-conditional probability $P(X|C_k)$ for each class k, where k = 0, 1.

iv. Input validation dataset.

v. Calculate the generalization error using the validation dataset.

vi. Calculate the posterior probability for each class k considering feature vectors X_i consisting feature set $\{x_{i1}, x_{i2}, \ldots, x_{i8}\}$.

vii. Predict the class of each test sample based on the posterior probability of each class k, greater probability of positive refers the risk of diabetes mellitus whereas that of negative detects the opposite.

4 Implementation

Implementation of the Naïve Bayes classifier is done using the programming language Java running on NetBeans IDE version 8.0.2. The computer system deployed is of the configuration: the processor is Intel core i3 CPU 1.90 GHz, size of installed memory (RAM) is 4GB and operating system (OS) is Windows 8.1 (64-bit OS).

Since it only requires two parameters from the training sample namely mean and standard deviation, we formulated a three-dimensional table to contain the values. The table is referred as “Analysis table” which

contains classes along the rows, features along the columns and levels are preserved for the mean and standard deviation of the specific feature in that specific class. It is to be mentioned here that the last column only contained the prior probability of that class.

The sample data set was inserted into the classifier and trained with different ratios of training and testing samples. As mentioned earlier, the sample data set consists of 768 samples with feature values in the form of numerical values. The sample dataset was divided into training data set and test data set. The means and standard deviations were calculated using standard formulae and stored in the "Analysis table". Once the classifier is initialized and the mean, standard deviations of features are stored, the classifier is ready to classify.

After building the classifier, test samples were input into the classifier one at a time for classifying. We used validation sets of different ratios for validating the training dataset. For classification, the posterior probability of each class for every test sample is calculated. The test sample was assigned to the class with maximum posterior probability.

5 Description of Results Found

We have applied our method to the data set several times. The result we achieved is promising. For calculating accuracy, we followed the simple technique of correctness ratio.

$$accuracy = \frac{no.of\ samples\ correctly\ classified}{no.of samples\ tested} \times 100\%$$

We have tested the samples with different test-train ratios and found a positive result in increasing the training samples. Figure 4 demonstrates the accuracy curve rising along with the increase of training samples. When the test-train ratio was 30%-70%, the result we achieved barely met the expectation. Increasing the training sample percentage to 80%, the accuracy increased a bit. Nevertheless, increasing the training sample to 90% the accuracy had a gradual pickup. It quickly ascends above 85% that is clearly more than previous works in this field. Table I summarizes all the results found. The best result we obtained is 87.28% for the case of 10-fold cross validation.

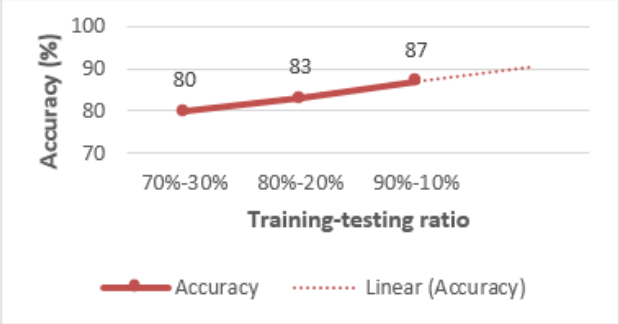

Figure 4. Performance curve

Table 1. Disease Classification Performance

Test-Train Ratio	Accuracy	Average Accuracy
30% - 70%	83.12%	80.74%
	80.95%	
	80.09%	
	78.79%	
20% - 80%	86.36%	82.78%
	82.47%	
	77.92%	
	84.36%	
10% - 90%	88.31%	87.28%
	87.01%	
	86.12%	
	87.69%	

6 Comparative Discussion of Results

Due to the limitation of data, i.e. deficiency of samples, disease prediction does not seem to be much robust. Still, we achieved a promising result by following our approach. In comparison to other works, it would not be an overstatement that our work gives satisfactory results. Yang Guo et al. proposed Naïve Bayes network yielded an accuracy of 72.3% [11]. Jack W. Smith et al. achieved an accuracy of 76% [10]. Again, K. Polat et al. considerably tried to gain more accuracy through SVM. They could achieve an accuracy of 82.05% [7]. S. Karatsiolis achieved 82.2% accuracy using modified support vector machine [14]. Mohammad Amine Chikh et al. raised the mark to 82.69% in their work [15]. H. Kahramanli achieved an accuracy of 84.24% [8] using ANN and FNN. Table II compares different results obtained over the dataset using various methods.

Table 2. Comparative Analysis

Reference	Classifier	Accuracy
[11]	Bayes Belief Network	72.3%
[10]	Neural Network	76%
[7]	Support Vector Machine	82.05%
[14]	Modified Support Vector Machine	82.2%
[15]	Fuzzy K-Nearest Neighbor	82.69%
[8]	ANN and FNN	84.24%
This paper	Proposed Bayesian Classifier	87.28%

7 Conclusion and Future Works

In this paper, we have demonstrated a comprehensive approach to predict the onset of diabetes mellitus based on a well-known data set. Obtained results are very intriguing. Although our findings show a good promise compared to previous works reported in the literature, we still believe there is a chance to improve it further with new techniques like ensemble learning, balancing classes, etc. considering.

As future work, we plan to work with an expert system capable of diagnosing diabetes mellitus robustly. For example, the expansion of the sample data set and reduction of the number of features are required for the improvement of classification. Work is in progress to integrate feature ranking to reduce the feature set and optimize the prediction based on an expanded data set.

REFERENCES

[1] Peter J.F. Lucas & Linda C. van der Gaag(1991)"The Principles of Expert Systems", Amsterdam, Addison-WesleyPublishing Company.

[2] Christopher D. Saudek et al.(2013) "A New Look at Screening and Diagnosing Diabetes Mellitus", *The Journal of Clinical Endocrinology & Metabolism*, vol 93, issue 7.

[3] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

[4] Aibinu, A.M, Salami, MJE, & Shafie, A.A. (2011). "A novel signal diagnosis technique using pseudo complex-valued autoregressive technique". *Expert Systems with Applications, 38*(8), 9063-9069.

[5] Isa NAM, & Mamat WMFW. (2011). "Clustered-Hybrid Multilayer Perceptron network for pattern recognition application". *Applied Soft Computing*, 11(1), 1457-1466.

[6] Polat,K.,& Gunes, S. (2007). "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease". *Digital Signal Processing*, 17(4), 702-710.

[7] Polat, K., Gunes, S.,& Arslan, A.(2008). "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine". *Expert Systems with Applications*, 34(1), 482-487.

[8] Kahramanli, H.,& Allahverdi, N. (2008). "Design of a hybrid system for the diabetes and heart diseases". *Expert Systems with Applications*, 35(1-2), 82-89.

[9] Jack W. Smith, JE Everhart, WC Dicksont, WC Knowler, RS Johannes, (1988), "Using the ADAP Learning Algorithm to Forecastthe Onset of Diabetes Mellitus".

[10] Murali S. Shankar (1996), "Using Neural Network to Predict the Onset of Diabetes Mellitus", *Journal of Chemical information and Computer Science,* vol. 36.

[11] Yang Guo, Guohua Bai, Yan Hu (2012), "Using Bayes Network for Prediction of Type-2 Diabetes", *7th International Conference for Internet Technology and Secured Transactions (ICITST)*, London.

[12] Knowler, W.C., DJ. Pettitt, PJ. Savage, and P.H. Bennett 1981. "Diabetes incidence in Pima Indians: contnbutions of obesity and parental diabetes". *Am J Epidemiol* 113:144-156.

[13] World Health Organization, "Report of a Study Group: Diabetes Mellitus. World Health Organization Technical Report Series". Geneva, 727, 1985.

[14] S. Karatsiolis, C. N. Schizas, (2012), "Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset",*IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*, pages: 139-144.

[15] Mohamed Amine Chikh, Meryem Saidi, Nesma Settouti (2012), "Diagnosis of Diabetes Diseases Using an Artificial Immune Recognition System2 (AIRS2) with Fuzzy K-nearest Neighbor", *Journal of Medical Systems,* vol 36, issue 5.