Prediction of Diabetes Using Probability Approach

T.monika Singh, Rajashekar shastry

T. monika Singh M.Tech

Dept. of Computer Science and Engineering , Stanley College of Engineering and Technology for Women, Telangana- Hyderabad, India.

Dr.B.srinivasu Associate Professor

Dept. of Computer Science and Engineering , Stanley College of Engineering and Technology for Women, Telangana- Hyderabad, India.

Rajashker Shastry Assistant Professor

Dept. of Computer Science and Engineering , Stanley College of Engineering and Technology for Women, Telangana- Hyderabad, India.

---***---

Abstract – *The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. With the emerging increase of diabetes, that recently affects around 346 million people, of which more than one-third go undetected in early stage, a strong need for supporting the medical decision-making process is generated. Diabetes mellitus is a chronic disease and a major public health challenge worldwide. 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. Using data mining methods to aid people to predict diabetes has gain major popularity.*

In this project, Bayesian Network classifier was proposed to predict the persons whether diabetic or not. 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. The dataset used is Pima Indian Diabetes dataset, which collects the information of persons with and without diabetes.

***Key Words*: classification, Bayesian network, attributes, prediction, probability.**

1.INTRODUCTION

Data mining is the process of discovering correlations, patterns or relationships through large amount of data stored in repositories, databases and data warehouse. Thus, new tools and techniques are being developed to solve this problem through automation [1]. Many techniques or solutions for data mining and knowledge discovery in databases are very widely provided for classification, association, clustering and regression, search, optimization, etc.

Health Informatics is a rapidly growing field that is concerned with applying computer science and information technology to medical and health data.

1.1 Diabetes

Diabetes is a chronic disease that occurs when the human pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces, which leads to an increase in blood glucose levels [2]. Generally a person is considered to be suffering from diabetes, when blood sugar levels are above normal.

1.1.1 Types of Diabetes

The three main types of diabetes are described below:

Type 1 – In this type of diabetes, the pancreatic cells that produce insulin have been destroyed by the defense system of the body.

Type 2- In this case the various organs of the body become insulin resistant, and this increases the demand for insulin.

Gestational diabetes – It is a type of diabetes that tends to occur in pregnant women due to the high sugar levels as the pancreas don't produce sufficient amount of insulin.

Controlling the blood glucose level of diabetic patients and keeping it within the normal range (70 mg/dL -120 mg/dL) is therefore the focal goal of physicians [3].

1.3 Problem Statement and Description

Prediction of diabetes using Bayesian Network : To identify whether a given person in dataset will be diabetic, non diabetic or pre-diabetic will be done on basis of attribute values. Dataset contains all the details of person like fast gtt value, casual gtt value**,** number of time pregnant, diastolic blood pressure (mmhg), triceps skin fold thickness(mm), serum insulin(μU/ml), body mass index (kg/m), diabetes pedigree function and age of person. Attributes like fast gtt, casual gtt, diastolic blood pressure values exceeding a specific value may contribute to identify whether a person is diabetic, non diabetic or prediabetic.

The aim of prediction of diabetes is to make aware people about diabetes and what it takes to treat it and gives the power to control. It makes necessary chances to improve lifestyle. classification evaluation for the prediction performance of Bayesian algorithms to predict diabetes. The proposed Bayesian Network classifier will predict the persons having diabetic or not.

1.5 Background

In the background we have defined the basic definitions and different strategies that can be used for diabetes detection.

1.5.1 Classification Techniques in Healthcare

The objective of the classification is to assign a class to find previously unseen records as accurately as possible. Classification process consists of training set that are analyzed by a classification algorithms and the classifier or learner model is represented in the form of classification rules. Test data are used in the classification rules to estimate the accuracy. The learner model is represented in the form of classification rules, decision trees or mathematical formulae [4].

1.6 Machine Learning

Machine learning is a branch of computer science that consists of algorithms that can learn from data, it provides set of methods that can detect patterns in the data and use the patterns to generate future predictions [5].

Machine learning is divided into two main types supervised and unsupervised learning. Supervised learning is the machine learning technique in which the learning algorithms make use of labelled data. In unsupervised learning, the model is trained on unlabelled data.

1.6.1 Naive Bayes

it is a classification technique based on Bayes theorem. a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes classifiers are based on Bayesian Theorem; it simplifies the learning method by assuming that features are independent of each other on the class context . This strong assumption is known as Naïve Bayes Assumption . Let us consider x ϵ X, the input feature vectors; y ϵ {1,..., c}, the class labels; then the Naïve Bayes Assumption is given by,

$$p(x/y = c) = \prod_{i=1}^{D} p(xi/y = c) \quad \text{....(Eq1.1)}$$

1.6.2 Bayesian Network

Bayesian Networks (BN) is a directed, acyclic, graphical representation of the probabilistic relationships among a group of variables. It is represented in the form of a directed graph whose nodes represents the attributes and edges represent the relationship between them.

The main building block of BN theory is Bayes' Theorem. This theorem is stated as follows:

$$P(x/y) = \frac{P(y/x)*P(y)}{P(x)} \quad \text{....(Eq1.2)}$$

where:

- $p(X|Y)$ is the posterior probability of the hypothesis X, given the data Y,
- $p(Y|X)$ is the probability of the data Y, given the hypothesis X, or the likelihood of the data,
- $p(X)$ is the prior probability of the hypothesis X, and
- $p(Y)$ is the prior probability of the data Y, or the evidence.

2. LITERATURE SURVEY

Diabetes prediction using Data Mining has been explored by various researchers from time to time and developed encouraging solution for medical expertise and researchers. As a result of all these research, diagnostic and prognostic models have been developed and influenced the existing clinical practices.

Subham Khanna and **Sonali Agarwal** [6] proposed classification method considering the impacts of different attributes present in dataset for the severity of the diabetics. The method intended to find out the total score for a patient indicates various categories such as low, medium and high risk patients. This research work is based on SGPGI.

Sudajai Lowanichchai, Saisunee Jabjone and **Tidanut Puthasimma** [7] proposed the application Information technology of knowledge-based DSS for an analysis diabetes of elder using decision tree. The decision tree is a decision modeling tool that graphically displays the classification process of a given input for given output class labels. The result showed that the Random Tree model has the highest accuracy in the classification is 99.60 percent when compared with the medical diagnosis that the error MAE is

0.004 and RMSE is 0.0447. The NBTree model has lowest accuracy in the classification is 70.60 percent.

Ashwin kumar U.M and **Dr. Anand kumar K.R** [8] proposed a novel learning algorithm to find the symptoms of cardiac and diabetes using data mining technology by using the method of Decision Tree and Incremental Learning at the early stage. The learner model is represented in the form of classification rules, decision trees.
These datasets were gathered from the patient files which were recorded in the medical record section of the BGS Hospital Bangalore. The ID3 algorithm is used to build a decision tree, given a set of non-categorical attributes Decision tree are commonly used for gaining information for the purpose of decision making.

Literature Review on Diabetes, by National Public health :Women tend to be hardest hit by diabetes with 9.6 million women having diabetes. This represents 8.8% of the adult population of women 18 years of age and older in 2003 and a two fold increase from 1995 (4.7%). By 2050, the projected number of all persons with diabetes will have increased from 17 million to 29 million. [9].

3. ARCHITECTURE

3.1 proposed system

we propose a system it identifies whether a person in the dataset will be diabetic or non-diabetic on the basis of pre-processed attribute values. dataset contains all the details of person. pre processing is used to improve the quality of data. we apply classifier to the modifies dataset to construct the bayesian model. classification with bayesian network shows the best accuracy. the proposed architecture is shown in fig. 3.1

3.2 System architecture

In the training phase each record of the patient is supplied with a class label. Every record is associated with a class label. In our work we have used two class labels, so each record will fall under one of the specified class category. Here in the training process we segregate the data into two classes i.e., class 0 and class 1. and we convert the data into discrete values. next we find the term frequency which means here we are finding the most frequent attribute values which are highly repeated.

Bayesian Probability Computation: we find the probabilities for dependent attributes, independent attributes and conditionally independent attributes.

Testing Phase : Test data is used to test the performance of the classifier on unseen 'test set'. In the testing phase, patient's record without class label is provided it means that the record is unseen before. With the previous examples or experiences provided to the Bayesian algorithm in training phase the machine is categorizing the record into one of the class category provided. The input in the test phase is a new record and the output is class.

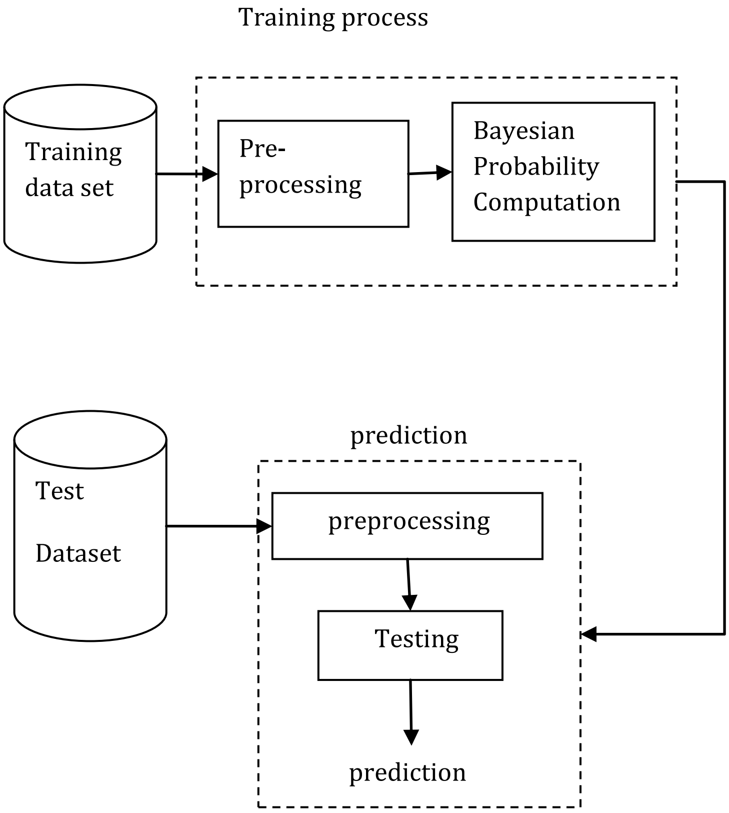

Figure 3.1 Architecture

4. IMPLEMENTATION

4.1 Pre-processing

The Pima Indian diabetes database is a collection of medical diagnostic reports of 768 examples from a population living near Phoenix, Arizona, USA. The Pima Indian Diabetes dataset available from the UCI Server and can be downloaded at www.ics.uci.edu/~mlearn/MLRepository.html. The data source uses 768 samples with two class problems to test whether the patient would test positive or negative for diabetes. All the patients in this database are Pima Indian women at least 21 years old. Class variable (0 or 1) where '1' means a positive test for diabetes and '0' is a negative test for diabetes. There are 268 cases in class '1' and 500 cases in class '0' . The aim is to use the first 8 variables to predict 9. 500 of the females are non-diabetic (65.1%) and the rest (34.9%) are diabetic. We are splitting the data as 80% of training data and 20% as testing data i.e., 615 records for training data and 153 records for testing. There are eight attributes that describe each female within this dataset, as well as, the class attribute. The defined Attributes are described in following Table-1.

Table-4.1:defined attributes

Attribute id	Attribute Description
1	Number of times Pregnant
2	Plasma glucose conc a 2 hours in an Oral glucose tolerance test
3	Diastolic blood pressure(mmHg)
4	Triceps skin fold thickness(mm)
5	2-hour serum insulin(mm U/ml)
6	Body mass index(wg in kg/(height in m))
7	Diabetes pedigree function
8	Age(years)
9	Class variable(0=no or 1=yes)

Having '0' in the class variable (attribute ID = 9) indicates a healthy female, while '1' indicates a diabetic one.

4.2 Training Process

The Figure 4.3 will give the description about the Training process required for finding the probabilities. It takes the input as training datasets and defined diabetes attributes. Using the defined attributes we find the probabilities in relevance to each attributes defined for the diabetes prediction.

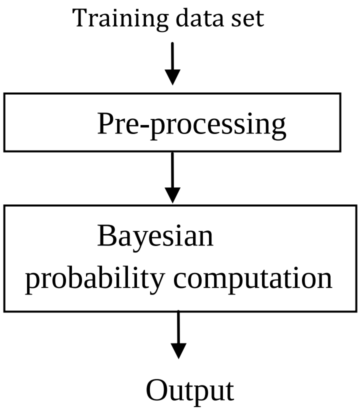

Figure 4.1: flow chart training process

4.3 preprocessing

In pre-processing we are transforming the data into discrete values depending upon the ranges of each attribute.

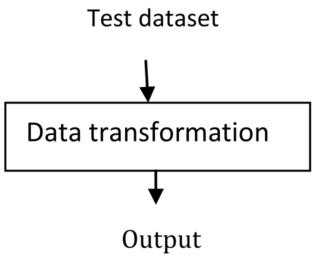

Figure 4.2: flow chart for pre-processing

4.4 Bayesian Probability Computation

Here the input is training dataset. first read the data set line by line then divide them into sub sets based attributes. Then it finds the probabilities of each attribute.

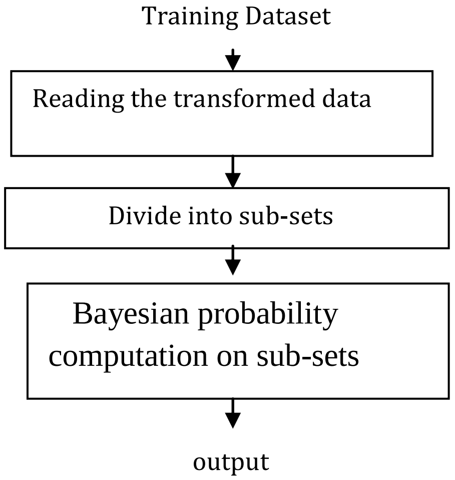
Training Dataset
Reading the transformed data
Divide into sub-sets
Bayesian probability computation on sub-sets
output

Figure 4.3 Generation of cluster

Procedure :
1. Open the input file
2. Read the file line by line
3. Divide the data set into sub set based on attributes
4 Calculate the probability of each class.
5. Calculate ΣP(ai|cj) i.e. the probability of attributes for the given class. Take attribute by attribute compute the each attribute probability based on class. Individual attributes probability compute to record level.
6. Repeat the process

To evaluate the modified Bayesian approach, we calculate the probabilities of each class and probabilities of each attributes given that class. First we will find Pr(Diabetes) and Pr(Non-diabetic). Next we will find probability of each attribute given the class.

$$P(ai|cj) = \sum_{i=1,j=1}^{n,m} \frac{P(cj|ai)*P(ai)}{P(cj)}$$

(Eq4.1)
where, n and m = End of file.

4.5 Class Prediction

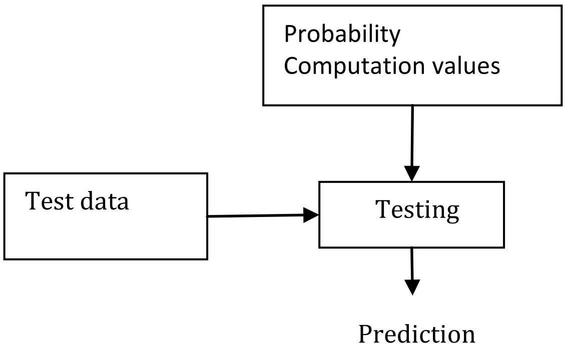

Figure 4.3: flow chart for class prediction

The process takes the probability computed values as input and performs an comparison check against the probability value and the test data value then it performs calculations then, the data record is predicted as Class-1, else it will predicted as Class-0. Data record classified as Class-1 is considered as Diabetes predicted and Class-0 as a Non-Diabetes.

5. EXPERIMENT RESULTS

5.1 Datasets Description

The dataset that is taken for this work is collected from "*Pima Indians Diabetes Database*" obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. It consists of 768 records of patient data. Here 80% of the data is taken for training and remaining 20% is taken for testing. The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care).

5.2 Results

T2he problem of work is about predicting whether a person is diabetic or non diabetic in a dataset by applying bayesian network . This problem is solved using the primary attribute . The dataset variables which are used for prediction of diabetes are fast plasma glucose concentration in an oral glucose tolerance test ,casual plasma glucose tolerance test and diastolic blood pressure (mmHg) is decision variable.

The quality of the prediction models is assessed by Accuracy. The Accuracy is the proportion of testing set examples that is correctly classified by the model.

Cross-validation:

In our system we classify 154 records, it gives the accuracy 65%. Here we are using non-exhaustive cross-validation called k-fold cross validation with 10-folds.

Confusion Matrix:

Confusion Matrix is useful when we are building a classification model to know how accurate the model is. It gives the count of number of matches and number of mismatches.

Precision: Precision is also referred to as positive predictive value.

Precision = TP/(TP+FP) (Eq5.1)

Where, TP and FP are the number of true positive and false positive predictions for the considered class FP is the sum of values in the corresponding column.

Recall: Recall is commonly referred to as sensitivity, corresponds to the true positive rate of the considered class.

Recall = TP/(TP+FN) (Eq5.2)

Where, FN is the sum of values in the corresponding row.

5.3 Performance Evaluation

F-Measure

The F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score: *p* is the number of correct positive results divided by the number of all positive results, and *r* is the number of correct positive results divided by the number of positive results that should have been returned. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst at 0.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad \text{(Eq5.3)}$$

Accuracy is measured as,

$$accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad \text{(Eq5.4)}$$

Table 5.1:Number of records

Number of training records	614
Number of Test records	154

Table 5.2: confusion matix

	Yes	No
Yes	80	29
No	25	20

On validating the test records we obtain accuracy 65%. Where the TP = 80, TN = 20, FP = 25 and FN = 29. From the above values we obtained precision and recall as 0.761% and 0.73%. Finally we estimate the F measure as 0.73%

6. CONCLUSIONS

Lately, medical machine learning has gained in interest by the scientific and research communities. Diabetes is considered as the world's fastest-growing chronic disease. It needs continuous self-management and control to maintain blood glucose level within the normal range, in order to prevent complications and prevent diabetic events. Diabetic is a condition that occurs when blood glucose is too low. The occurrence of diabetic may result in seizures, unconsciousness, and possibly permanent brain damage or death.

We proposed a model in predicting diabetes by applying data mining technique. Diabetes mellitus is a chronic disease and a major public health challenge worldwide. Using data mining methods to aid people to predict diabetes has gain major popularity. In this Bayesian Network classifier is proposed to predict the persons whether diabetic or not. Results have been obtained.

For future work, more input features can be used, e.g. exercise, heart rate, and metabolism rate. In addition, drawn blood samples of plasma insulin are needed to compare with the simulated values. Moreover, we recommend the proposed models to be tested on a larger dataset.

ACKNOWLEDGEMENT

I express my deepest gratitude to my parents for all their support and love. *It is with a sense of gratitude and appreciation that I feel to acknowledge any well wishers for their king support and encouragement during the completion of the project.*

I would like to express my heartfelt gratitude to my Project guide ***Dr. Srinivasu Badugu,*** *Coordinator of M.Tech in the Computer Science Engineering department, for encouraging and guiding me through the project.*

REFERENCES

1. Mukesh kumari, Dr. Rajan Vohra and Anshul arora, "Prediction of Diabetes Using Bayesian Network", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 5174-5178, 2014.
2. WHO. "Diabetes Programme", Internet: http://www.who.int/diabetes/en/> [Sep. 27, 2014].
3. Ronald Aubert, William Herman, Janice Waters, William Moore, David Sutton, Bercedis Peterson, Cathy Bailey, and Jeffrey P. Koplan, "Nurse Case Management To Improve Glycemic Control in Diabetic Patients in a Health Maintenance Organization," in Annals of Internal Medicine, vol. 129, no. 8, pp. 605-612, 1998.
4. Mukesh kumari, Dr. Rajan Vohra and Anshul arora, "Prediction of Diabetes Using Bayesian Network", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 5174-5178, 2014.
5. WHO. "Diabetes Programme", Internet: http://www.who.int/diabetes/en/> [Sep. 27, 2014].
6. Ronald Aubert, William Herman, Janice Waters, William Moore, David Sutton, Bercedis Peterson, Cathy Bailey, and Jeffrey P. Koplan, "Nurse Case Management To Improve Glycemic Control in Diabetic Patients in a Health Maintenance Organization," in Annals of Internal Medicine, vol. 129, no. 8, pp. 605-612, 1998.
7. Z.H. Zhou, and Z.Q. Chen, "Hybrid Decision Tree", Knowledge-Base
8. Systems, Vol. 15, pp. 515-528, 2002.
9. U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In Advances in Knowledge discovery and data mining, pages 1{34. American Association for Artificial Intelligence Menlo Park, CA, USA, 1996

BIOGRAPHIES

T.monika singh received her B.E and M.tech degee in computer science from Osmania university. Her research interest are data mining.