PREDICTION OF DIABETIC SYMPTOMS IN PATIENTS USING DATA MINING TECHNIQUES

Dr.D.Shanmuga Priyaa[1]

ABSTRACT

This paper focuses on the classification of algorithms for identifying the presence of diabetes in patients using data mining techniques. Data-mining is used to predict the pattern from large databases.Data-mining includes data acquisition, data integration, data exploration, model building, and model validation.In this paper different classification algorithms were used to predict the signs of diabetes according to the norms of World Health Organization. The dataset consists of 768 instances and 8 attributes along with class attribute. To predict the presence of diabetes classification algorithms such as BayesNet, Naivebayes, J48 andIBK were used. The result shows that Naïvebayes algorithm works well compared to other techniques.

Keywords: Datamining, classifiers, Bayesnet, Naivebayes.

I. INTRODUCTION

In recent days the size of the databases has moved to terabytes. Data mining concept is used to retrieve useful information from large databases. Data mining is the analysis of largedata sets inorder to find unknown relationships, summarize the data in such a way to make them understandable and useful to the data owner [1].

The data that are present in the databases may be un-preprocessed, incomplete, and noisy. For example, the databases may containrepeatedfields, missing values and outliers, data not suitable for data mining models, Values that are not consistent.

For the purpose of data mining, the databases need to undergo preprocessing, in the form of data cleaning and data transformation [1]

1.2 Data Mining in Knowledge Discovery Process

The term KDD or Knowledge Discovery in Databases, refers to the process of finding knowledge in data by using data mining algorithmsaccording to the specifications of measures and thresholds.Data mining refers to the process of finding interesting patterns in data that are not accessible by basic queries [2]

A data mining process generally includes the following steps [3]

- Removing noisy and irrelevant data is Data Cleaning
- Combining multiple data sources is Data integration.
- To retrieve the data that is relevant to the analysis is Data selection.
- Transformingthe selected data into forms for the mining procedure is Data transformation and it is also known as data consolidation.

[1] Associate Professor, Dept of CS, CA& IT, Karpagam Academy of Higher Education Coimbatore, 641021, India,

- ⏩ Useful patterns are extracted using Data mining techniques.
- ⏩ To identify interesting patternsbased on given measures is Pattern evaluation.
- ⏩ Knowledge representation is used to visually present the discovered knowledge. This step uses visualization techniques which helps the user to understand and interpret the data mining results.

II. RELATED WORK

Diabetes mellitus is classified as Type 1 (insulin resistance) and Type 2 (gestational). There are two major types of diabetes (ie)Type 1 and Type 2. Type 1 is immune relatedand is caused by damage to the islet cells of the pancreas. Type 2 is caused by the grouping of genetic factors related to environmental factors such as [4]:

Obesity

Overeating

Lack of exercise and stress

Ageing

Insulin secretion

Insulin resistance

Clinical Features of Type 1:

- Weight Loss by polyuria, nocturia and polydipsia.
- Usually appears in the age10-12 years.
- Increased appetite and fatigue.
- Thigh Muscular Atrophy.
- Acetone smell.

Clinical Features of Type 2 :

- Overweight persons are usually affected.
- The age factor is over 40 years.
- Genital candidiasis, urinary tract infections/skin infections.

In [5] the main purpose is to predict how likely people are to be affected by diabetes with different age groups, based on their life style and find out factors that are responsible for the person to be diabetic.

The design of prediction model design for diabetic diagnosis has been an active research area for the past decade. Most of the literature survey have mentioned that the clustering algorithms and artificial neural networks (ANNs) are the models used for diabetes prediction.

In [6] the authors have used the techniques such as EM algorithm, H-means+ clustering and Genetic Algorithm (GA), for the prediction of diabetesin patients. The performance for H-means+ proved to be better than the other techniques.

In [7]for diabetes diagnosis Fuzzy Ant Colony Optimization (ACO) has been used.

The paper [8]has approached the aim of diabetic prediction by using ANNs. In this work they have done preprocessing to preplace the missing values in the dataset.

In [9]based on 13 symptoms of the disease implementation has been done in MATLAB for the prediction of diabetes in patients using neural network.

III. PROPOSED WORK

3.1 Dataset Description

The proposed work uses the diabetes dataset from Pima Indians Diabetes Database National Institute of Diabetes and Digestive and Kidney Diseases.

The value that is in binary form indicates whether the patient has the diabetic signs according to World Health Organization criteria. In this dataset all patients are femaleage at least 21 years.

The dataset consists of 768 instances, and 8 attributes in addition to 1 class attribute. All the attributes are of numeric value. The attributes used in the dataset are as follows

1. Number of times Pregnancy Occurred in numerals
2. PGC (Plasma Glucose Concentration)
3. Level of Blood pressurein mm Hg
4. Thickness of Triceps skin fold in mm
5. Serum insulin(2-Hour) in mu U/ml
6. BMI (Body Mass Index)
7. DPF (Diabetes Pedigree Function)
8. Age(years)
9. Class variable (ie) 0 or 1

If the class value is 1 it indicates that the patient has diabetes and if the class value is 0 it indicates the patient doesn't have diabetes.

3.2 Classifiers

The classification algorithms used in tis work are BayesNet, Naivebayes, J48,and IBK.

3.2.1 BayesNet

It is based on probabilistic graphical model (GM). It is used to represent a set of random variables and their conditional dependencies through a directed acyclic graph (DAG). This graphical structure used to represent knowledge about an uncertain domain

3.2.2 NaiveBayes

It is based on the Bayesian theorem and its suitability for the dimensionality of the inputs is high. NaiveBayes uses the method of maximum likelihood for parameter estimation. In real world complex situation it performs well. It is highly scalable and requires a number of parameters in a linear manner in the form of variables (features/predictors) in a learning problem.

$P(C|F\_1,\dots.F\_n\ ) = P(C)P(F\_1,\dots\dots F\_n|C)P(F\_1,\dots..F\_n)$

3.3.3 J48

It is based on decision tree algorithm which decides the target valuebased on different attribute values available in the dataset. To classify a new item, it creates a decision tree based on the attribute valuesavailable in the training data.The internal nodes represents the different attributes, the branches between the nodes denotes the possible values that these attributes can have and the terminal nodes denotes the final value (classification).

3.3.4 IBK

It is based on k-nearest algorithm. It works on similarity distance calculation between instances. It predicts the majority class amoong the neighbors. It is also known as lazzy learning algorithm. Depending upon the data different distance metrics can be used. Euclidean distance metrics can be used for continous variables, whereas other metrics can be used for categorical data.

IV. ESTIMATION OF MODEL PERFORMANCE

Weka 3.6[11] data mining tool kit is used for analyzing the results. 10 fold cross validation method is used as the classification model. Confusion matrix, ROC curve and other statistical methods are also used to identify the diabetic symptoms in the patients.

4.1 Confusion Matrix

This matrix is one of the evaluation measures that is used to evaluate the performance of a classifier. This is a binary classification model which classifies each instance into two classes (1) true class (2) false class. Four possible classifications can be raised from this matrix such as true positive rate, true negative rate, false positive rate and false negative rate as shown in Fig 1.

Actual Class		Predicted Class	
		+	-
	+	TP	FN
	-	FP	TN

Fig. 1. Confusion Matrix

- The number of positive instances that is correctly predicted is True Positive (TP)
- The number of positive instances that is wrongly predicted as negative is False Negative (FN)
- The number of negative instances that is wrongly predicted as positive is False Positive (FP)
- The number of negative instances that is correctly predictedTrue Negative (TN)

V. EXPERIMENTAL RESULTS

The proposed work contains 768 instances and 8 attributes. 9th attribute is class value 1 and is interpreted as "tested positive for diabetes"and class value 0 is interpreted as "tested negative for diabetes". Classification algorithms will produce the correctly and incorrectly classified instances as shown in Table 1.

Table 1. Comparison of classifier models based on correctly and incorrectly classified instances

Classification Algorithms	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**Time taken (secs)**
BayesNet	571	197	0.02
Naivebayes,	586	182	0.02
J48	567	201	0.02
IBK	539	229	0

In Table 2 specificity, sensitivity and accuracy of the classifiers are shown. The chart also shows that the NaiveBayes classifies the prediction of diabetics as positive, when compared to the other classifiers. From the result it is observed that NaiveBayes produces high specificity, sensitivity and accuracy, when compared to the other classifiers.

Table 2. Comparison of classifier models based on Sensitivity, Specificity, Precision, Accuracy

Clsiifiers/ Performance Measures	**Sensitivity (%)**	**Specificity (%)**	**Precision (%)**	**Accu racy (%)**
BayesNet	61	82	64	74
Naivebayes	**61**	**84**	**68**	**76**
J48	60	81	63	74
IBK	53	80	60	70

CONCLUSION

From the above observations it is found thatNaiveBayes classifier performs well when compared to the other classifiers. The approaches used in this paper efficiently classify the diabetic patient correctly.

REFERENCES

[1] https://www/mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf

[2] Jiawei Han and MichelineKamber, Data Mining: Concepts and Techniques, August 2000.

[3] http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

[4] Vitamins Aarif Ali, Mashooq Ah Dar and AadilAyaz, "Diagnostic Approaches to Diabetes Mellitus and the Role of Vitamins", Journal of Nutrition & Food Sciences, 2017

[5] PardhaRepalli, "Prediction on Diabetes Using Data mining Approach"

[6] VeenaVijayan V, AswathyRavikumar, "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus", International Journal of Computer Applications (0975 - 8887) Volume 95- No.17, June 2014.

[7] MostafaFathiGanji and Mohammad SanieeAbadeh, "Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease", Proceedings of ICEE 2010, May 11-13, 2010

[8] T.Jayalakshmi and Dr.A.Santhakumaran, "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks", International Conference on Data Storage and Data Engineering, 2010, pp. 159-163

[9] SonuKumari and Archana Singh, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus", Proceedings of 71h lnternational Conference on Intelligent Systems and Control (ISCO 2013).