A CLASSIFICATION AND PREDICTION MODEL FOR DIABETIC DATASET BY USING DIFFERENT TRANSFORMATION TECHNIQUES

R.Brindha,
Research Scholar,
M.Phil. Computer Science,
Vellalar College for Women (Autonomous),
Erode -638 012.

P.Anitha,
Assistant Professor,
Department of Computer Applications,
Vellalar College for Women (Autonomous),
Erode -638 012.

Abstract: Diabetes is a chronic disease that contributes to a significant portion of the healthcare expenditure for a nation as individuals with diabetes need continuous medical care. Currently in the healthcare industry different data mining techniques are used to mine the interesting pattern of disease using the statistical medical data with the help of different machine learning techniques. The proposed system assists doctor to predict disease correctly and the prediction makes patient and medical insurance are also get benefited. This research focuses on to diagnosis diabetes disease as it is a great threat to human life worldwide. The system uses the K-Nearest Neighbor (KNN) and ID3 Algorithms as supervised classification models. Finally, the proposed system calculates and compares the accuracy of ID3 and KNN and the experimental result demonstrates that the ID3 provides better accuracy for diagnosis diabetes. For the clinical database, the Pima Indians Dataset is used in this research.

Keywords: Data Mining, Health Informatics, Classification, KNN, ID3, Dataset.

I.INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amount of data. It is a step in the KDD process of applying data analysis and discovery algorithm. Data mining software is one of a number analytical tool for analyzing data.There are two primary goals of data mining tend to be prediction and description. Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest. On the other hand Description focuses on finding patterns describing the data that can be interpreted by humans. The Disease Prediction plays an important role in data mining. Classification is the most popular data mining technique for medical diagnosis. There are number of techniques available such as K-Nearest Neighbor, ID3, Naïve Bayes etc,. Data mining algorithm is used for testing the accuracy in predicting diabetic status.

DIABETES

Diabetes is the most common disease nowadays in all population and in all age groups. It is a disease in which the body does not produce or properly use insulin. The cells in our body require glucose for growth for which insulin is quite essential. When someone has diabetes, little or no insulin is secreted. In this situation, plenty of glucose is available in the blood stream but the body is unable to use it. The types of diabetes are Type-1 Diabetes, Type-2 Diabetes and Gestational Diabetes. Diabetes can causes serious health complications including heart disease, blindness, kidney failure, and lower-extremity amputations.

Type-1:

Type-1 diabetes occurs when the body's immune system is attacked and the beta cells (these cells produce insulin) of pancreas are destroyed. This results in insulin deficiency. The only treatment to Type-1 diabetes is insulin.

Type-2:

Type-2 diabetes is caused by relative insulin deficiency. Pancreas in Type-2 diabetes still produces insulin produces insulin but it may not be effective or may not produces sufficient amount of insulin to control blood glucose Type-2 diabetes is the most common type of diabetes which usually develops at age 40 and older. Treatment focuses on diet and exercise.

Gestational diabetes:

Gestational diabetes during pregnancy when a woman has high blood sugar (glucose) level, who does not have diabetes before pregnancy is said to be gestational diabetes. According to the recently announced diabetes criteria, it is found that around 18% of pregnant women have gestational diabetes.

II.DATA MINING TECHNIQUES: CLASSIFICATION

In my research work there are two classification techniques are used to predict disease. K-Nearest Neighbor (KNN) and ID3 classifiers are used to train dataset in this system.

K-NEAREST NEIGHBOR (KNN)

KNN is a supervised learning algorithm which classifies new data based on minimum distance from the new data to the K-Nearest Neighbor. The proposed work has used Euclidean Distance to define the closeness.

$$\text{dist}(X,Y) = \sqrt{\sum\nolimits^{n}_{i=0}} (X_i\text{-}Y_i)^2$$

In other words, for each numeric attribute, we take the difference between the corresponding values of that attribute in tuple X and in tuple Y, square this difference, and accumulate it.

Min-max normalization, for example, can be used to transform a value X_{mm} of a numeric attribute A to X_{mm} in the range [0,1] by computing

$$X\text{-min}(X)$$

$$X_{mm} = \frac{}{\max(X)-\min(X)}$$

Rules, rather than a single rule with highest confidence, when predicting the class label of a new tuples.

Z-score Normalization transforms the data by converting values to a common scale with an average of zero and standard deviation of one.

$$v' = \frac{v - \dot{F}}{\sigma_F}$$

ID3 ALGORITHM

ID3 algorithm is used to generate a decision tree. It is a precursor to the C4.5 algorithm.

The process follows,

Step1:Take all unused attributes and calculates their entropies.

Step2: chooses attributes that the lowest entropy is minimum or when information gain is maximum

$$\text{Entropy (S)} = -p + \log^2 (p+) - p \log^2 (p-)$$

Step3: Makes a node containing that attribute.

DATASET

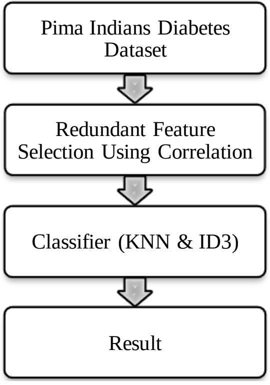
The Pima Indians Diabetes Dataset has been used for prediction the diabetes disease. This dataset consists of 768 samples with 8 numerical valued attribute where 500 are tested.

Attributes Description

S.NO	Attribute name
1	Number of times pregnant
2	Plasma glucose concentration in an oral glucose tolerance test
3	Diastolic blood pressure (mm/Hg)
4	Triceps skin fold thickness (mm)
5	2-hour serum insulin (μU/ml)
6	Body mass index (kg/m^2)
7	Diabetes Pedigree function
8	Age (years)
9	Status (0-Healthy, 1-Diabetes)

III. PROPOSED WORK

An expert system is created to predict heart disease. The workflow of this system is given below:


Pima Indians Diabetes Dataset
Redundant Feature Selection Using Correlation
Classifier (KNN & ID3)
Result

Step 1: Irrelevant feature selection method using correlation is applied to Pima Indian Diabetic Dataset. Out of eight attribute, age attribute is selected as removal attribute. The new dataset is constructed with seven attributes.

Step 2: KNN and ID3 algorithm have executed on preprocessed train dataset and which is the model to classify the test data.

Step 3: Analyze the results of two classification algorithm.

Performance Evaluation:

Performance evaluation is carried out by accuracy calculation as which is the ratio of the number of correctly classified instances to the total number of instances of the test data.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \times 100\%$$

Where,

TP, FP, TN and FN are the number of true positive, false positive, true negative and false negative respectively.

Result and Discussion:

In this research, an expert system isproposed for predicting the disease like diabetes using data mining classification technique. The system gives benefits to the doctors, physicians, medical students and patients to make decision regarding the diagnosis of the diseases. The system finds 95% accuracy for training phase using ID3 algorithm. ID3 algorithm has been found highest accuracy than the KNN algorithm respectively.

1) Performance evaluation of KNN algorithm:

In this table1, the prediction result shows thatKNN with Z-score classifier has correctly classified 327 instances and incorrectly classified 33 instances. The accuracy of correctlyclassified instance is 90.83% and incorrectly classifiedinstance is 9.16%.KNN with Min-Max classifier has correctly classified 338 instances and incorrectly classified 22 instances. The accuracy of correctly classified instance is 93.88% and incorrectly classified instance is 6.11%.Feature Selection KNN with Z-score classifier has correctly classified 328 instances and incorrectly classified 32 instances. The accuracy of correctly classified instance is 91.11% and incorrectly classified instance is 8.88%.

KNN	Prediction Result		Accuracy	
	Correctly Classified Instances	Incorrectly Classified Instances	Correctly Classified Instances	Incorrectly Classified Instances
Training Instances:400 Testing Instances:360 (Z-score normalization)	327	33	90.83%	9.16%
Training Instances:400 Testing Instances:360 (Min-Max normalization)	338	22	93.88%	6.11%
Training Instances:400 Testing Instances:360 (Feature Selection KNN Z-score)	328	32	91.11%	8.88%

Table 1: Classification Summary of KNN Classifier

2) Performance evaluation of ID3 algorithm:

In this table2, the prediction result shows that ID3 classifier has correctly classified 342 instances and incorrectly classified 18 instances. The accuracy of correctly classified instance is 95% and incorrectly classified instance is 5%.

ID3	Prediction Result		Accuracy	
	Correctly Classified Instances	Incorrectly Classified Instances	Correctly Classified Instances	Incorrectly Classified Instances
Training Instances:400 Testing Instances:360	342	18	95%	5%

Table 2: Classification Summary of ID3 Classifier

3) Comparison of classification accuracy:

The performance of proposed clinical expert system was analyzed with diabetesdisease dataset using KNN and ID3. According to experimental results, Table3 represents the performance comparison of KNN and ID3 classifier based on percentage split (360:40) technique.

Classifier	Number of instances		Accuracy
KNN with Min-Max	Correctly classified	338	93.88%
	Incorrectly classified	22	6.11%
KNN with Z-score	Correctly classified	327	90.83%
	Incorrectly classified	33	9.16%
Feature Selection KNN with Z-score	Correctly classified	328	91.11%
	Incorrectly classified	32	8.88%
ID3	Correctly classified	342	**95%**
	Incorrectly classified	18	5%

Table 3: Performance Comparison of KNN & ID3

IV.CONCLUSION

In experimental studies, the dataset have been partitioned into 52% training and 46% for testing of KNN and ID3 algorithm. It has been performed on PIDD and the results are compared. It may be seen that by applying the Feature Selection method, 7 attributes have been selected from 8 attributes and by performed classification of the selected attributes. The proposed method ID3 classifier maximizes the classification accuracy thsn the existing method KNN. For the future research work, we suggested to develop expert system with different Feature Selection and Classification methods which could significantly decrease healthcare cost via early prediction and diagnosis of diabetes.

V. REFERENCES

[1]. Han, J.,&Micheline, K. "Data mining: Concepts and Techniques", Morgan Kaufmann .Publisher, 2006.

[2]. V. Anuja Kumari, R.Chitra " Classification Of Diabetes Disease Using Support Vector Machine", 2013.

[3]. K.R Lakshmi, S.Premkumar, " Utilization of Data mining Techniques for prediction of Diabetes Disease survivability", 2013.

[4]. Veenavijayan, AswathyRavikumar, " Study of Data mining algorithms for prediction and diagnosis of Diabetes Mellitus", 2014.

[5]. Omar S.Soliman, Eman AboElhand "Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine", 2014.

[6]. Amit kumar Dewangan, Pragati Agrawal "Classification of Diabetes Mellitus Using Machine Learning Techniques", 2015.

[7]. Tahani Daghistani, Riyad Alshammari "Diagnosis of Diabetes by Applying Data Mining Classification Techniques", 2016.

[8]. Akshat Sharma, Anuj Srivastava "Understanding Decision Tree Algorithm by Using R Programming Language", 2016.

[9]. Dr. M. Renuka Devi, J. Maria Shyla "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", 2016.

[10]. Wei Peng, Juhua Chen and Haiping Zhou "An Implementation of ID3 - Decision Tree Learning Algorithm".