ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 5, Issue 1, January 2017

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

A Review on Diabetes Mellitus diagnoses using classification on Pima Indian Diabetes Data Set

R. Sivanesan[1]
Assistant Professor, Dept. of BCA & MSc SS,
Sri Krishna Arts and Science College, Coimbatore,
Tamil Nadu – India

K. Devika Rani Dhivya[2]
Assistant Professor, Dept. of BCA & MSc SS, Sri Krishna
Arts and Science College, Coimbatore,
Tamil Nadu – India

Abstract: Diabetes Mellitus is a major health issue originates in all over the world and is a varied group of diseases categorized by chronic promotion of glucose in the blood. It rises because the body is incapable to yield ample insulin for its individual requirements. Diabetes distresses some 300 million people through world-wide. This paper comprises with the diagnoses of Diabetes Mellitus using Classification, a technique generally used in medical data mining. It is a function or model finding process that is used for dividing the data into different classes of an object based on its characteristics. Unlike classification models and approaches are being used for diagnoses and treatment for Diabetes Mellitus. This paper focus on the performance of J48 Decision Tree, it is a prophetic machine-learning model that adopts the target value (dependent variable) of a new sample based on various attribute values of the accessible data. The J48 Decision tree classifier follows the simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. The data set for diabetic patients taken from machine learning repository consists of 768 instances with 9 attributes. The data set is evaluated using training set, 10 fold cross validation and percentage split method and the results to which algorithm builds best models in an effectual manner.

Keywords: Diabetes Mellitus, Data Mining, Classification, J48 Decision Tree.

I. Introduction

Data mining refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form which is easily comprehensible to us. The richness and fast evolution of the data mining discipline comes from its large variety of research areas of interest. Data mining applications can use different kind of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects). Classification is one of the most frequently studied problems by Data Mining and Machine Learning (ML) researchers. Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called training set, which contains the same set of attributes except the class label, not yet known. The algorithm analyses the input and produces a prediction. This paper aims to study the behaviours of different classification algorithm on PIMA Indian diabetes data set.

Diabetes Mellitus (DM), also known as simply diabetes, is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period. This high blood sugar produces the symptoms of frequent urination, increased thirst, and increased hunger. Untreated diabetes can cause many complications. Acute complications include diabetic ketoacidosis and nonketotic hyperosmolar coma. Serious long-term complications include heart disease, stroke, kidney failure, foot ulcers and damage to the eyes. Diabetes is due to either the pancreas not producing enough insulin, or the cells of the body not responding properly to the insulin produced.

II. Background Study

P.Yasodha and N.R.Ananthanarayanan [3], in their research paper they have compared classification algorithm using WEKA tool on clinical available dataset. In that they have employed J48, LAD Tree, REP Tree and estimated the accuracy of algorithms when the attributes have changed.

Shelly Gupta, Dharminder Kumar and Anand Sharma [4], they have been employed various classifiers using WEKA tool on PIMA Indian Diabets, Wisconsin Breast cancer, StatLog Heart Disease and BUPA Liver Disorder Health care datasets and discovered all the classification tools were produce same accuracy and times to build a model or training set on same dataset. When the dataset and attributes changes it behaves differently on every datasets and produce various accuracy rates and times.

Raj Kumar and Dr. Rajesh Verma[5], presented a paper on classificaction algorithms for datamining as a survey to understand how the classification algorithm behaves differently on various scenarios.

Jianchao han [6], in his research work the decision tree using WEKA has been used to build the prediction model of the type 2 diabetes data set. These models consider the Plasma Insulin attribute as the main attribute for predicting the disease.

III. Methodology

A. J48 Algorithm

J48 is an open source Java implementation of the C4.5 algorithm in WEKA data mining too. C4.5 is a program that creates a decision tree based on a set of labeled input data. It is a simple C4.5 decision tree for classification, it creates a binary tree. The decision tree approach is the most useful in the classification problem. With this technique, a tree is built to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple while the tree is built. It ignores the missing values. The basic idea of J48 is to divide the data into range based on the attribute values for that item that are found in the training sample. It allows classification either in the form of decision tree or rules generated based on the test set provided.[9]

B. Decision Tree

Decision tree [7], is a tree structure, which is in the form of a flowchart. It is used as a method for classification and prediction with representation using nodes and internodes. The root and internal nodes are the test cases that are used to separate the instances with different features. Internal nodes themselves are the result of attribute test cases. Leaf nodes denote the class variable.

Decision tree provides a powerful technique for classification and prediction in Diabetes diagnosis problem. Various decision tree algorithm are available to classify the data, including, ID3,C4.5,C5,J48 and CART. In this paper, J48 decision tree algorithm has been chosen to establish the model. Each node for the decision tree is found by calculating the highest information gain for all attributes and if a specific attribute gives an unambiguous end product (explicit classification of class attribute), the branch of this attribute is terminated and target value is assigned to it.

C. Evaluation Metrics

- ✓ **Time**: This referred to as the time required to complete training or modeling of a dataset. It is represented in seconds.
- ✓ **Kappa Statistics:** A measure of the degree of non-random agreement between observers or measurements of the same categorical variable.

- ✓ **Mean Absolute Error:** Mean absolute error is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error.

- ✓ **Mean Squared Error:** Mean-Squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value. The mean-squared-error is simply the square root of the mean-squared error. The mean-squared error gives the error value the same dimensionally as the actual and predicted values.

- ✓ **Root relative squared error:** relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value. As with the root mean-squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicated value.

D. Confusion Matrix

The confusion matrix [7] is a useful tool for analyzing how well the classifier can recognize tuples of different classes. Table 4 shows the General form of the confusion matrix.

		Predicted Class		
		Yes	**No**	**Total**
Actual Class	**Yes**	**TP**	**FN**	**P**
	No	**FP**	**TN**	**N**
	Total	**P**	**N**	**P + N**

Table 4. Confusion matrix

Terminologies of confusion matrix as follows:

1. **True Positives [TP]:** These refer to the positive instances that were correctly classified by the classifier.
2. **True Negatives [TN]:** These are the negative instances that were correctly classified by the classifier.
3. **False Positives [FP]:** These are the negative instances that were incorrectly classified as positive.
4. **False Negatives [FN]:** These are the positive instances that were misclassified as negative.

E. Dataset

For our experiment we take PIMA Indian diabetes dataset (PID) **,** and this data set is used by many approached for the sake of classification. The data set is taken from UCI machine learning repository [2][8].

S.No	Attribute	Description	Type
1	Pregnant	Number of times pregnant	Real
2	Plasma	Plasma glucose concentration in an oral glucose tolerance test	Real
3	Diastolic	Diastolic blood pressure (mm/Hg)	Real
4	Triceps	Triceps skin fold thickness (mm)	Real
5	SerumInsulin	2-hour serum insulin (µU/ml)	Real
6	BMI	Body Mass Index (kg/m^2)	Real
7	PedigreeFun	Diabetes Pedigree function	Real
8	Age	Age (years)	Real
9	Diabetes	Status (0-Healthy, 1-Diabetes)	Discrete

Table 1. Brief Description of the Data Set

The dataset, originally donated by Vincent Sigillito from the Applied Physics Laboratory at the Johns Hopkins University, is one of the most well-known datasets for testing classification algorithms. This dataset consists of records describing 768 female patients of Pima Indian heritage which are at least 21 years old living near Phoenix, Arizona, USA, Form the 768 patients in the PID dataset, classification algorithms used a training set with 500 patients and a testing dataset with 268 patients.

F. J48 Classifier

J48 classifier produces the decision tree, Based on the tree classifier it produces some other results like error rate, correctly classified instances, kappa statistics RMSE, RAE, RRSE for training set methods, percentage split method and 10 fold cross validation method. The results are discussed in the following tables.

- **Based on training set**

	No. of Instances	Percentage
Correctly Classified	646	84.11%
Incorrectly Classified	122	15.89%

Table 2. Performance results from J48 classification algorithm – Training set

Other results of J48 classification algorithm-Training set are followed as in the below table.

Metrics	%
Kappa Statistics	0.63
Mean Absolute Error	0.23
Root Mean Squared Error	0.34
Relative Absolute Error	52.43
Root Relative Squared Error	72.42
Total Number of Instances	768

Table 3. Other results of J48 classification algorithm – Training set

Confusion matrix generated by J48- Training set

	A_Tested_Postitive	B_Tested_Negative
A_Tested_Positive	468	32
B_Tested_Negative	90	178

Table 5. Confusion Matrix of J48 – Training set

- **Based on 10 Fold Cross Validation**

	No. of Instances	Percentage
Correctly Classified	567	73.82%
Incorrectly Classified	201	26.17%

Table 6. Performance results from J48 classification algorithm – 10 Fold Cross Validation

Other results of J48 classification algorithm-10 Fold Cross validation are followed as in the below table.

Metrics	%
Kappa Statistics	0.41
Mean Absolute Error	0.32
Root Mean Squared Error	0.44
Relative Absolute Error	69.48
Root Relative Squared Error	93.62
Total Number of Instances	768

Table 7. Other results of J48 classification algorithm – 10 Fold Cross Validation

Confusion matrix generated by J48- 10 Fold Cross Validation

	A_Tested_Postitive	B_Tested_Negative
A_Tested_Positive	407	93
B_Tested_Negative	108	160

Table 8. Confusion Matrix of J48 – 10 Fold Cross Validation

- **Based on Percentage Split Model**

Since a 65:35 percentage split was applied on the dataset 269 of the instances were used as the test dataset while the rest of were using for training the model. The J48 algorithm gives the following correctness results for the given dataset.

	No. of Instances	Percentage
Correctly Classified	206	76.58%
Incorrectly Classified	63	23.42%

Table 9. Performance results from J48 classification algorithm – Percentage Split

Other results of J48 classification algorithm – Percentage Split are followed as in the below table.

Metrics	%
Kappa Statistics	0.49
Mean Absolute Error	0.29
Root Mean Squared Error	0.41
Relative Absolute Error	65.05
Root Relative Squared Error	87.68
Total Number of Instances	269

Table 10. Other results of J48 classification algorithm – Percentage Split

Confusion matrix generated by J48- Percentage Split

	A_Tested_Postive	B_Tested_Negative
A_Tested_Positive	140	41
B_Tested_Negative	22	66

Table 11. Confusion Matrix of J48 – Percentage Split

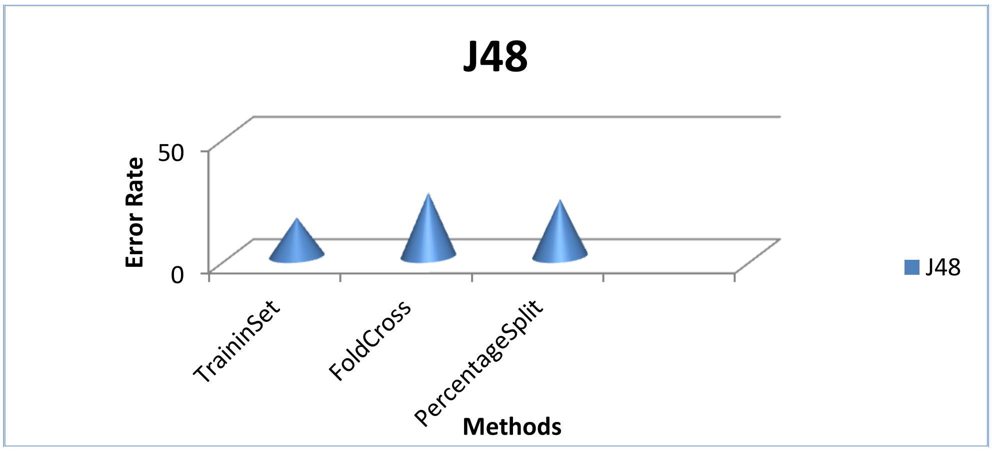

Fig. 2 Comparison of Error Rate Generated by the Classifier

IV. Conclusion

The performance of J48 Decision Tree, are analyzed with different metrics. Decision are used to model actual diagnoses of diabetes for local and systematic treatment, along with presenting related work in the field. An experimental result shows the effectiveness of the comparative study of various data set. The performance was investigated for the diabetes diagnosis problem. In future it is planned to propose a model that give best accuracy in Naïve Bayes classifiers using Gini Index based fuzzy classification for diagnosing the Diabetes Mellitus. The work can be extended and improved for the automation of diabetes analysis.

References

1. National Diabetes Information clearinghouse (NDIC) http://diabetes.niddk.nih.gov
2. UCI Machine Learning Repository-Center for Machine Learning and Intelligent System, http://archive.ics.uci.edu.
3. P.Yasodha and N.R.Ananthanarayanan, Comparative Study of Diabetic Patient Data's Using Classification Algorithm in WEKA Tool, IJCAT, Vol.3 Issue 3, 2014.
4. Shelly Gupta, Dharminder Kumar and Anand Sharma, Performance Analysis of Various Data Mining Classification Techniques on Healthcare Data, IJCSIT, Vol. 3, No.4 August 2011.
5. Raj Kumar, Dr.Rajesh Verma, Classification Algorithms for Data Mining: A Survey, IJIET, Vol. 1 Issue 2 August 2012.

6. Top 10 algorithms in datamining, x.Wu et al.
7. Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, Morgan Kauffman Publishers, 2001.
8. U.M.Ashwinkumar and Dr.K.R.Anandakumar, Predicting early detection of cardiac and diabetes symptoms using data mining techniques, IPCSIT vol.49 2012, IACSIT press, Sinngapore
9. Maninder Singh, A review on Data mining algorithms, IJCSIT, vol. 2 issue 2, April- June, 2014.
10. SushilKumar Rameshpant Kalmegh, comparative analysis of WEKA data mining algorithm RandomForest, Randomtree, and LADTree for classification of Indigenous News Data, IJETAE, Vol. 5 Issue 1, Janunary 2015.