International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 10, 2020
ISSN: 1475-7192

Decisions and Machine Learning for Diabetes Diagnosis System

Dr. Zaki .S. Tawfik[1], Omar F. Youssif[2], Mohammed A. Subhi[3]

Abstract

The Medical Diagnosis System of Diabetes aiming to identify the correct diagnosis of Patient's diabetes as quickly as possible and at lower cost. The Diabetes Diagnosis System (DDS) has three subsequent stages; the first stage is the construction of medical dataset (MD) with eight features taken for 1000 patients and covering three classes (Diabetic, Non-Diabetic, and Predicted- Diabetic). .

The second stage is the data mining- based machine learning, which introduces Interactive Dichotomizer 3 (ID3) classifier; is employed to diagnose the condition of a patient from his/her medical history. The outcome of implementing the proposal system showed that the accuracy of the ID3 model has been found approximately (98.25).

Keywords: *Diabetes Diagnosis System (DDS), medical dataset (MD), Interactive Dichotomizer 3 (ID3)*

I. Introduction

Data Mining provides several benefits such as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. Thathelps the healthcare researchers make efficient policies, constructing drug recommendation systems, developing profilesof individuals *etc*. [1]. Diabetes is an opportune disease which has large wealth of data available and has with ithuge complications. There is a need for a better and a more accurate approach in the diagnosis of the disease. Data mining is the better approach, since it is a process of extracting useful, hidden information from larger databases. There are many data mining techniques available to do this process like association mining, clustering, classification, predictive analysis… etc.[2].

Theparticular decision tree, ID3 algorithm and how it can be used with data mining for medical research. The evaluation of the different types of decision trees along with clustering algorithms to determine if there is a better approach for the medical industry specifically for determination of the risk of heart disease. It using these algorithms is an iterative process where processes are always being improved. Shukr (2013)[5], the Iterative

[1] *Computers Engineering Techniques Department, Al-Hikma University College*
[2] *Computers Engineering Techniques Department, Al-Hikma University College*
[3] *Computers Engineering Techniques Department, Al-Hikma University College*

Dichotomiser3(ID3) algorithm was detect the classification of the cardiac arrhythmia from a normal ECG signal based on wavelet decomposition, produce a set of rules which five types of ECG arrhythmias including the normal case. Angeline and Sivaprakasam(2013) [6].

Construct Diabetes Dataset

The DM-Based Diabetes Diagnosis System (DM-DDS) will be explained in details. The proposed system is a supervised learning system. This system aims to advance the prediction of patient's diabetes class (Diabetic, Non-Diabetic, and Predicted- Diabetic) using data mining techniques. The advantages of prediction arise from reducing cost and time of diagnosis, and increasing the accuracy of prediction. The proposal was applied on a constructed dataset of 1000 individuals', from Baghdad society, covering the three classes. The algorithm of data mining was applied in this proposal ID3 classifier. The data are collected from Iraqi society. The data were acquired from the laboratory of Medical City Hospital and (the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital). Patients' files are taken and data extracted from them and entered in to the database to construct the diabetes dataset see figure (2). The data consist of medical information, laboratory analysis… etc. The data that have been entered initially into the system are: No. of Patient, Sugar Level Blood, Age, Gender, Creatinine ratio(Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile, including total, LDL, VLDL, Triglycerides(TG) and HDL Cholesterol , HBA1C, Class (the patient's diabetes disease class may be Diabetic, Non-Diabetic, orPredict-Diabetic).

ID	No_Patio	Gender	Name	AGE	Sugar	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
1	34325	M	######	58	6.7	20.8	800	9.1	6.6	2.9	1.1	4.3	1.3	33	Y
2	44835	M	######	60	10.8	2.1	56	7.6	3.3	1.7	0.9	1.7	0.8	36.6	Y
3	23972	F	######	56	7	4	45	9.2	4.1	0.6	1.3	1.4	0.9	30	Y
4	34301	F	######	43	5	2.1	55	5.7	4.7	5.3	0.9	1.7	2.4	25	P
5	35150	M	######	63	9.7	7	84	8.1	6	2.2	1.1	4	1	28	Y
6	23973	F	######	61	7.2	5.1	72	11.5	4.4	2.1	1.1	2.5	0.9	26	Y
7	34278	F	######	46	5.6	3	59	5.1	5.7	3.8	1.3	2.8	1.7	24	Y
8	45703	M	######	51	12.9	3.9	53	10.9	3.6	1.1	0.8	2.3	1	29	Y
9	23974	F	######	60	6.7	6	72	10.7	4.4	2.1	1.1	2.5	0.9	26	Y
10	34286	F	######	45	5.7	3.1	54	4	5.9	1.8	1.6	3.5	0.8	24	Y
11	45702	F	######	54	13	7	72	7.6	4.9	2.8	0.8	3	1.2	31	Y
12	23975	M	######	31	4.3	3	60	12.3	4.1	2.2	0.7	2.4	15.4	37.2	Y
13	34395	F	######	60	9.3	5.4	47	6.8	5.1	2.1	1.1	3	1	28	Y
14	43261	F	######	64	10.2	2	35	7.9	5.3	3.1	0.9	3.1	1.4	33	Y
15	23976	F	######	61	6.9	4.3	56	12.1	4.4	2	1	2.5	0.9	29	Y
16	20321	M	######	50	5.3	4.8	62	5.9	5.3	1.3	1	3.7	0.6	19	P
17	34396	M	######	61	9.3	3.8	62	8.9	6.8	4.3	1.2	3.9	1.9	29	Y
18	23977	M	######	30	4.7	7.1	81	6.7	4.1	1.1	1.2	2.4	8.1	27.4	Y
19	26925	F	######	50	5.4	5.7	53	6	5.4	1.7	1.4	3.3	0.7	25	P
20	34420	F	######	58	10.5	5.8	41	9.7	5	4.5	1.1	2.2	2	37	Y

Figure (1): the dataset of the DDS

The main data preprocessing tasks done in the proposal are: Remove Redundancy, Noisy Data, and Feature Selection. By feature selection process only eight out of the total sixteen features were taken into account (Age, Gender, HBA1C, TG, Urea, Chol, HDL and BMI), the less important features were ignored as their information gain

is of no crucial significance except for the Sugar Level Blood which was ruled out because it is the decisive factor in diabetes diagnose. Missing Values, Also it's noteworthy to mention that the Hba1c values were found missing in some of the patients physical examinations. Accordingly figures for the missing values were taken as estimated by the laboratory experts.

ID3 algorithm

Procedure subsets produced until "pure" or homogenous nodes, contains elements of only one class. Starts with complete dataset of training examples, are given in attribute-value representation with eight categorical attributes (Age, Gender, HBA1C, TG, Urea, Chol, HDL, BMI and class attribute) see figure (3), which will help determine the class, the diabetes class can be one of three classes of diabetes. Those three samples can be (diabetic(Y), non-diabetic (N) and predicted- diabetic (P) attributes). The ID3 algorithm have, following steps are listed, see algorithm(1):

Algorithm (1): ID3(Examples, Target_attribute, Attributes)

Input:Examples *// are the trainingexamples//*

*,**Target_attribute** // is the attribute whose value is to be predicted*

by the tree //

,Attributes *// is a list of other attributes that may be tested by thelearned*

decision tree//

Output: *decision tree that correctly classifies the given* ***Examples***

Begin

Create a ***Root*** node for the tree

If all ***Examples*** are positive, **Return** the single node tree ***Root***, with label = *Pos* **If** all ***Examples*** are negative, **Return** the single node tree ***Root,*** with label = *Neg*. **If** all ***Examples*** are predictive, **Return** the single node tree ***Root,*** with label =*pre*.

If *Attributes* is empty, then **Return** the single node tree **Root**, with label = most common value of the ***Target _attribute*** in the ***Examples***.

Else Begin

A* =** the attribute from ***Attributes that best classifies ***Examples***

The decision attribute for ***Root = A***

For each possible value, ***vi***, of ***A* do**

Add a new tree branch below ***Root,*** corresponding to the test ***A*** = ***vi* Let** be the subset of ***Examples*** that have value ***vi*** for ***A* Add** a sub tree ***ID3(Examplesvi , Target attribute, Attributes – {A}))***

End For End

Return *Root*

End

Decision Tree Classifier

Can described in may steps as.

- classifying data using attributes.
- has two or more branches, with leaf nodes and values for the attribute.

A. Made Root node equal to null

Witch are Label, Attribute and Children..

B. Values of the class attribute list isempty.

C. And select the best to start as the rootnode

C. 1.Entropy:

Can be calculated by using Eq.(2). :

$$\text{Entropy}(S) = -P_{pos} \log_2 P_{pos} - P_{neg} \log_2 P_{neg} \quad \text{Eq.(2)}$$

Best attribute: The attributeHBA1C is the best because it has the maximum gain, and as a result it will be the root node of a decision tree, which had the maximum gainvalue

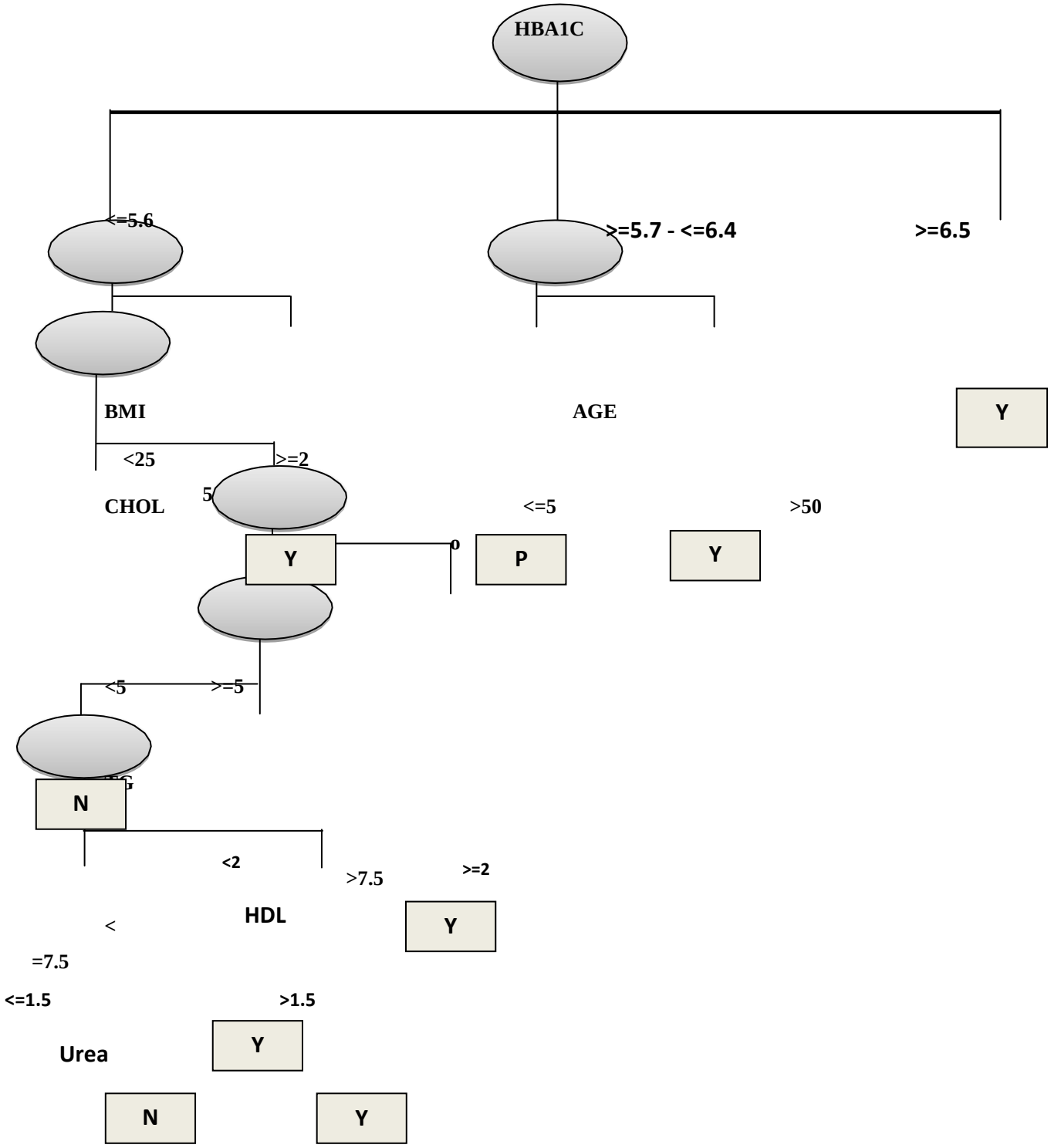

Figure (5): The finalclassify three types of Diabetes

Rules

1.IF HBAIC <=5.6 THEN BMI ELSE IF HBAI >=5.7 - <=6.4 THEAGE ELSE Y

2. IF BMI <25 THEN CHOL Y

3. IF CHOL >=5 THEN N ELSE TG

4. IF TG =< 2 THEN HDL ELSE Y

5. IF HDL < = 1.5 THEN Urea ELSE Y

6. IF Urea > 7.5 THEN Y ELSE N

7. IF AGE <= 50 P ELSE Y

Implementation and experimentalresults

The system deals with diabetes disease system for experiment on the model, It used training of the classifier (ID3) on TrainingDDS has been done with the sets of features, and then the obtained classifiers are tested on Testing DDS. The model has been experimented for many times to assess the accuracy of the classifiers, then light which of them rate higher accurate

II. Results

This section will explain results according the standards evaluation measures of classifications. The classification model has been constructed in each of these experiments. Next, this model has been applied to the same Testing DDS.

The classification results areeither;

1. TP -true positive denoting correct classifications of positive cases (one ofDDS).
2. TN -true negative denoting correct classifications of negative cases (normal).
3. FP -false positive answers denoting incorrect classifications of negative cases into class positive (misclassified records as one ofDDS).
4. FN - false negative answers denoting incorrect classifications of positive cases into class negative (misclassified records asnormal).
5. Unknown1 (Predicted- Diabetic true positivesPr-p).
6. Unknown2 (Predicted- Diabetic true negativesPr-n).

Table (1) which presents the performance of the classifier ID3 with the set of eight features along 200 patients (testing dataset), see figure (8):

Table (1): Classification Results of ID3 Classifier (testing of 200 patients)

Value	Classifier	TP	TN	FP	FN	Unknown1 (Pr-p)	Unknown2 (Pr-n)	Accuracy
Without Missing	**ID3**	150	38	0	1	9	2	98.25%
With Missing	**ID3**	114	36	27	12	1	10	75.5%

III. Conclusions

7. Constructing diabetes dataset manually; is by selecting random samples of patients from Iraq Health Ministry. This system of work can be used as a reliable indicator for diabetes diseases diagnosis and (ID3) classifier is presented as the diagnostic tool to aid the physician in the analysis ofdiabetic.

8. Preprocessing dataset such as noise and redundancy removal, normalization, and optimization of the patient's attributes to eight only depending on physician experience in the analysis of diabetic. All this was used successfully to solve the problems of diabetes disease diagnosis such as reduction the time of classification process.

9. The system uses ID3 algorithms as a classification tool for diabetes disease. It has been proved efficient, since ID3 gives optimized rules to classify and diagnose the individual state as diabetic, non-diabetic or predicted with high accuracy, see figure (7) and table(1).

10. Prediction of ID3 algorithm show when all the independent variables are statistically independent of each other.

11. Diabetes is complicated to diagnose as many and diverse symptoms can be important. The rule base for the prototype system was organized in frames and templates (forms) with basic attributes. It provides a guideline which combined with some additional rules would be very useful to predicate thediabetic.

References

[1] *Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth* "**From Data Mining to Knowledge Discovery in Databases**" *American Association for Artificial Intelligence* , 37-54-1996.

[2] K.NALINI KUMARI &G.SUBBALAKSHMI "**A SHORTEST PATH IDENTIFICATION FOR**

FEATURE GENERATION AND EXTRACTION FROM MEDLINE" - **International Journal of Computers Electrical and Advanced Communications Engineering Vol.1 (3), ISSN:2250**-2012.

[3] Asma Shaheen& Waqas Ahmad khan "**Intelligent Decision Support System in Diabetic eHealth Care**"From the perspective of Elders Blekinge Institute of Technology-Sweden ThesisMCS-2009

[4] Mary Slocum "**DECISION MAKING USING ID3 ALGORITHM** " RIVIER ACADEMIC JOURNAL, VOLUME 8, NUMBER 2, FALL2012.

[5] Nidhal Hameed Shukr "Classification of Cardiac Arrhythmias using ID3 Classifier" Thesis MCS.2013.

[6] Y. Angeline Christobel&P.Sivaprakasam "**A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset**" International Journal of Engineering and Advanced Technology Volume-2, Issue-3, February 2013.

[7] Tom M. Mitchell, "**Machine Learning**", McGraw-Hill, March ,1997.

[8] Sin-Min Lee, SJSU, **"ID3 Algorithm**",http://cs.sjsu.edu/~lee/cs157b/cs157b.html.

[9] N.Suneetha, CH.V.M.K.Hari, V.Sunil Kumar, "**Modified Gini Index Classification: A Case Study Of Heart Disease Dataset**", (IJCSE) International Journal onComputer Science and Engineering, Vol. 02, No. 06,2010.