DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4774

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A NEW STAD MODEL TO PREDICT THE DIABETES MELLITUS

N.Aswin Vignesh
Research scholar
Department of CSE,
Jamal Mohamed College (Autonomous)
Tiruchirappalli.

Dr.D.I.George Amalarethinam
Ph.D,Associate Professor, Bursar & Director MCA,
Department of CSE,
Jamal Mohamed College (Autonomous)
Tiruchirappalli.

Abstract: Diabetes-mellitus refers to the metabolic disorder that happens due to less insulin secretion action. It is characterized by hyperglycemia. The persistant hyperglycemia of diabetes leads to damage, malfunction and failure of different organs such as kidneys, eyes, nerves, blood vessels and heart. Detection and diagnosis of diabetes at an early stage is the need of the day. Diabetes disease diagnosis and interpretation of the diabetes data is an important classification problem. A variety of data mining techniques are used to discover new patterns of disease and promote the early detection and diagnosis of complex diseases such as diabetes Rule extraction is on among them. The rules are extracted from the dataset. The extracted rules may not only be highly accurate, but also simple and easy to understand. Therefore in this study, The rule extraction algorithm Enhanced STAD model is proposed to achieve highly accurate, concise, and interpretable classification rules for the pima Indian diabetes(PID) dataset, which comprises 768 samples with two classes(diabetes or non-diabetes) and eight attributes. The advanced decision tree algorithm is generated and used for classification. STAD model achieved substantially better accuracy and provided a considerably fewer average number of rules and antecedents. These results suggest that proposed algorithm, is more suitable for medical decision making including the diagnosis of all type of diabetes mellitus.

Keywords: Rule Extraction, Type 2 Diabetes mellitus, Pima Indian diabetes, Data mining

I. INTRODUCTION

Diabetes is often called a modern society disease. The lack of regular exercise and rising obesity rates are some of the main contributing factors for diabetes. It is a very serious disease that if not treated properly and on time, can lead to very serious complications, including death[1]. Detection and diagnosis of diabetes at an early stage is the need of the day. Diabetes disease diagnosis and interpretation of the diabetes data is an important classification problem[2]. Data classification problem is studied by statisticians and machine learning researchers. Data classification is widely used in variety of Engineering and scientific disciplines such as biology, psychology, medicines, marketing, computer vision, and artificial intelligence[3]. The goal of the data classification is to classify objects into a number of categories or classes. For a given dataset, the task of classification is to assign a class to the data object.

In 2011 there were 347 million diabetics worldwide and by 2030 this number is expected to increase to 552 million. About 4.6 million deaths were caused by diabetes in 2011 and by 2030; it is projected to be the seventh leading cause of death [4]. According to the centers for disease control and prevention, an estimated 29.1million people or 9.3% of the US population, have diabetes [5], 8.1 million of whom remain undiagnosed. In 2010, diabetes was listed as the underlying cause of death on 90,000 death certificates and a cause of death another 3,44,525, making it the fourth leading cause of death in India[6].The peak age of onset of type 2 diabetes mellitus which was previously known as non-insulin dependent diabetes mellitus or adult-onset diabetes is typically later than that of type 1 diabetes and accounts for about 80-90% of all diagnosed adult cases of diabetes[7]. Type 2 diabetes mellitus usually starts with insulin resistance, a disorder in which cells primarily within the muscles, liver and fat tissue do not utilize insulin lose the ability to produce properly. The beta cells in the pancreas begin to gradually lose the ability to produce sufficient quantities of insulin as the need for the hormone increases[8]. In contrast to individuals some primarily have insulin resistance and only a minor defect insulin secretion and only slight insulin resistance. An increasing amount of data is being collected in medical databases and historical data on complex disease such as patient's blood glucose levels is becoming more widely available therefore traditional methods of manual analysis have become inadequate[9]. As a result a variety of data mining are being applied in order to discover new patterns of disease and promote the early detection and diagnosis of complex diseases such as diabetes [10].

In this study, Stipulation Technique with Advanced Decision tree (STAD) is applied for rule extraction. It is tested with Pima Indian Diabetes dataset (PID)[11]. The environmental attributes such as hereditary, life style are also considered in this study. It was observed that the proposed STAD model gave better results with respect to accuracy.

II. LITERATURE REVIEW

The Pima Indians Dataset [PID] has the highest reported incidence of diabetes in the world. Smith used the same dataset to test a model for prediction the onset of diabetes mellitus. This study is modeled to find the relationship between the onset of diabetes mellitus and previous risk factors for diabetes among Pima Indian data set [12].

In 2012 shanker[13] evaluated the effectiveness of artificial NN classifiers in predicting the onset of non-insulin

dependent diabetes mellitus among the pima Indian female population[14]. According to knowler et al., the pima Indians have the highest reported incidence of diabetes in the world. Smith et al.[15] used the same dataset to test a model for predicting the onset diabetes mellitus. A study on semi-supervised fuzzy classification was conducted by lekkas and mikhailov[16] for the diagnosis of two medical problems. In their system, two domains contain records of actual patients with a known diagnosis were used.

They proposed the use of a new evolutionary approach to derive compact fuzzy classification systems directly from the data without any prior knowledge or assumptions regarding the distribution of the data.[17] The fuzzy membership functions are assigned to fuzzy variables. Rules and membership functions are then automatically created and optimized in an evolutionary process. A recent rule extraction algorithm that works in discrete and continuous data set by Rabybak et.al was proposed. The algorithm applies genetic programming to generate a syntactic tree representing a set of rules that mimics the functioning of the tree.

The objective of the Re-Rx algorithm is to achieve highly accurate concise and interpretable classification rules for the PID dataset. The most important aim of Re-Rx is to improve the conciseness and interpretability of extracted rules for physicians, because the competition for achieving only better classification accuracy for the PID dataset. The existing Re-Rx algorithm is used to extract a set of concise and interpretable diagnostic rules for the PID. The number of rules extracted by Re-Rx is more compared to the proposed model.

III. THE PROPOSED STIPULATED TECHNIQUE WITH ADVANCED DECIDION TREE (STAD) MODEL

The STAD model extracts the If-then rules directly from the training Data using the advanced Decision tree. The rules are learned from decision tree, where each rule for a given class will ideally cover many of the class's tuples. Rules are learned one at time. Each time a rules is learned, the tuples covered by the rule are removed and the process repeats on the remaining tuples. Since the basic decision tree learns the rules one at a time, the rules learned are at high accuracy. The rules need not necessary be of high coverage.

The process continues until the terminating condition is met. For example when there are no more training tuples or the quality of rule returned is a user specified threshold. The learn one rule procedure finds the best rule for the current class given the current set of training tuples.

Typically rules are grown in a general to specific manner. This technique append by adding the attribute test as a logical consent to the existing condition of the rule antecedent. Consider the training set as Pima Indian Diabetes data., Attributes regarding each applicant include their BMI, OGTT, and DBP data set. The classifying attribute is BMI level, which indicates whether a diabetes or Non diabetes. To start with, the rule antecedent is empty gradually the other attributes are incorporated. For example, in this study the BMI, OGTT and DBP are considered as attribute to detect the diabetes.

A. Stipulation Technique with Advanced Decision tree (STAD) Algorithm

Step1: Train and prune an NN using the dataset S and all of its D and C attributes
Step2: Let D' and C' be the sets of discrete a continuous attributes,
respectively, still present in the network and let S' be the set of data
samples correctly classified by the pruned network.
Step3: Generate decision tree by using both discrete and continuous C' attributes .
Step4: For each rule Ri is generated.
Step5: Stad (Examples, Target_Attribute, Attributes)
Step6: Create a root node for the tree
Step7: If all examples are positive, Return the single-node tree Root, with label =+.
Step8: If all examples are negative, Return the single-node tree Root, with label = -.
Step9: If number of predicting attributes is empty then
Return the single node tree
Root with label ← most common value of the target attribute in the examples.
Step10: Otherwise
Begin A← The Attribute that best classifier examples. Decision Tree attribute for Root = A.
Step11: For each possible value, v_i, of A,
Step12: Add a new tree branch below Root, corresponding to the test A = v_i.
Step13: Let Examples(v_i) be the subset of examples that have the value v_i for A
Step14: If Examples(v_i) is empty Then
below this new branch add a leaf node with
label = most common target value in the examples
Step15: Else below this new branch add the subtree AD (Examples(v_i),Target_Attribute,Attributes – {A})
Step16: End
Step17: Return Root
Step18: If support Ri>si and error Ri>s2 then
Step19: Let Si be the set of data samples that satisfies the condition of rule Ri, let Di be the set of discrete attributes and let Ci be the set of continuous attributes that does not appear in rule condition Ri.
Step20: Call continuous stad(Si,Di, Ci)
Step21: Otherwise stop.

IV.ILLUSTRATION

A. Rules Extracted Using the Proposed STAD Model

R1: If OGTT≤ 130 then non diabetes
R2: If OGTT∈ (130,140) and BMI≤33 and DBP ≤90 then Non diabetes
R3: If OGTT ∈ (120,135) and BMI ∈ (25,35) and DBP∈ (80,85) then Non diabetes
R4: If OGTT ∈ (140,150) and BMI >35 and DBP>92 then diabetes
R5: If OGTT ∈ (120,130) and BMI >33 and DBP <85 then Non diabetes
R6: If OGTT>152 and BMI >35 and DBP>95 then diabetes
R7: If OGTT >155 and BMI >40 and DBP>90 then diabetes

R8: If OGTT <155 and BMI <32 and DBP>85 then Non diabetes
R9: If OGTT >150 and BMI >35 and DBP>90 then diabetes
R10: If OGTT<140 and BMI <30 and DBP<93 then Non diabetes

B. Confusion Matrix

It is predicted that the person have diabetes is the predicted class will give the answer as "yes". It is predicted that the person have no diabetes is the predicted class will give the answer as "No". The classifier made a total of 768 predictions(e.g. patients were being tested for the presence of that disease). The classifier predicted "yes" 532 times, and "no" 236 times. In reality, 105 patients in the sample have the disease, and 60 patients do not have diabetes.

TABLE I. Confusion Matrix

	Classified as Healthy	Classified as not Healthy
Actual Healthy	TP	FN
Actual Not Healthy	FP	TN

C. Performance of STAD model

Training Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ (1)

Testing Accuracy (Test set) = $\frac{TP+TN}{TP+TN+FP+FN}$ (2)

Training & Testing Accuracy(SD) = $\sqrt{\frac{1}{N}\sum_{i=1}^{n}(xi-\bar{x})\,2}$ (3)

TPR = $\frac{TP}{P}$ (4)

FPR = $\frac{FP}{N}$ (5)

The performance of STAD model is tabulated in table II. The Training Accuracy, Testing Accuracy, Number of rules, Average number of antecedents and standard deviation of TR, TS are listed in table II. The proposed STAD model is compared with the Regular covering Technique and it was observed that STAD model produces higher percentage of accuracy.

TABLE II. Performance of STAD model
(average of 10 runs of 10-fold cross validation [CV])

	TR ACC (%)	TS ACC (%)	# Rules	Ave.# antecedent	TR ACC (SD)	TS ACC (SD)
Regular covering technique	91.11	89.62	9	3	1.59	1.72
STAD model	93	91	10	3	1.65	1.78

D. Histogram representation of STAD model compared with regular covering technique

In this representation, STAD model is compared with regular covering technique. In multi-objective optimization and economics, pare to optimality is always an important issue. In the case of medical rule extraction there is a tradeoff between high diagnostic accuracy and the interpretability of extracted rules. Physician may want to obtain extracted diagnostic rules with reduced accuracy and more interpretability. Needless to say, if the optimal solution can be found then the best extracted rules can be obtained. Ideally to extend the optimal solution to obtain a wider viable region that provides improvements in both diagnostic accuracy and interpretability, the rule extraction technique is used to find compromise between both requirements by building a simple rule set that mimics how the well-performing complex model makes decisions. The comparative analysis of Regular covering technique and STAD model for PID dataset is shown in Figure 4.

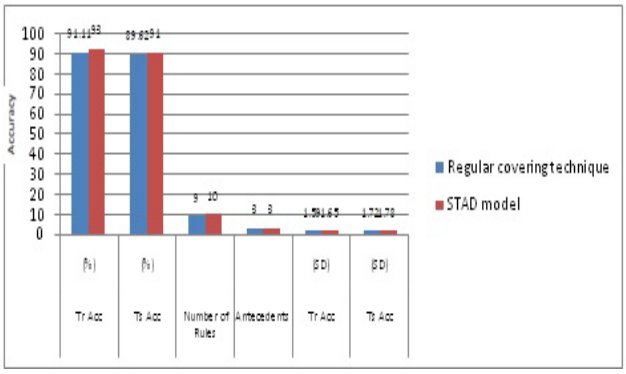

Fig.4 Histogram representation of STAD Model compare with Regular covering technique.

V. CONCLUSION

The STAD model is more accurate, concise and interpretable and therefore more suitable for medical decision making. Actually high accuracy, conciseness and interpretability are achieved simultaneously by the proposed STAD model. The use of STAD model is expected to be particularly useful in patients with diabetes mellitus whose fracture risk is relatively high. Needless to say the diagnosis of diabetes mellitus remains a complex problem; therefore STAD model should be tested on more recent and complete diabetes datasets in future studies in order to ensure that the most highly accurate rules can be extracted for diagnosis.

VI. REFERENCES

[1] YilmazN, InanO, UzerMS. New data preparation method based on clustering algorithms for diagnosis systems of heart an diabetes diseases. JMedSyst 2014;38:48–59.
[2] BeloufaF,ChikhMA. Design of fuzzy classifier for diabetes disease using modified artificial bee colony algorithm Comput Methods Prog Biomed 2013;112:92–103.
[3] MansourianM, FaghihimaniE, AminiM, Farina D.Ahybriintelligent system for diagnosing microalbuminuriain type2 diabetes patients without having to measure urinary albumin.ComputBiol Med2014;45:34–42.
[4] Centers for Disease Control and Prevention. National Diabetes statistics Report: Estimate o f Diabetes and its Burden in the United States, 2014. Atlanta, GA: Department of Health and Human Services 2014 .

[5] Zhu J, XieQ, ZhengK. An improve d early detection method of type-2 diabetes mellitus using multiple classifier system. In Sci 2015;292:1– 14.
[6] HunterDJ.Gene-environment interactions in human diseases.NatRevGenet 2005; 6:287–98.
[7] HommeMB,Reynolds KK, ValdesR, inder MW. Dynamicpharmacogenetic models in anti coagRegulationtherapy. ClinLab Med 2008;28:539–52.
[8] Ding S, ZhaoH, ZhangX, XuX, NieR.Extrem e learning machine:algorithm, theory and applications. ArtifIntell Rev 2015;44:103–15.
[9] YilmazN,InanO, UzerMS.A new data preparation method based on clus-tering algorithms for diagnosis system so heart and diabetes diseases. J Med Syst 2014;38:48–59.
[10] GürbüzE,Kılıç E.A new adaptive support vector machine for diagnosis of diseases. Expert Syst 2014;31:389–97.
[11] ShankerMS.Using neural networks to predict the on set of diabetes mellitus.JChemInfComputSci 1996;36:35–41.
[12] ParkJ, Eding ton DW.A sequential neural network mode l for diabetes pre- diction. ArtifIntell Med 2001;23:277–93.
[13] Gadaras,I, Mikhailov L. An interpretable fuzzy rule-based classification methodology for medical diagnosis.ArtifIntell Med 2009; 47:25–41.
[14] GhazaviSN, LiaoTW. Medical data mining by fuzzy modeling with selected features. ArtifIntell Med 2008; 43:195–206.
[15] Chavas ADF, Vallasco MMBR, Tanscheit R. Fuzzy rules extraction from support vector machines for multi-class classification. Neural Comput Appl2013;22:1571–80.
[16] Lekkas S, Mikhailov L. Evolving fuzzy medical diagnosis of Pima Indians dia- betes and of dermatological disease. ArtifIntell Med 2010;50:117–26.
[17] Ubeyli ED. Modified mixture of experts for diabetes diagnosis. J Med Syst 2009;33:299–305.