International Journal of Pure and Applied Mathematics
Volume 118 No. 6 2018, 19-28
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

A New Monotony Advanced Decision Tree Using Graft Algorithm to Predict the Diagnosis of Diabetes Mellitus

D.I. George Amalarathinam[1] and N. Aswin Vignesh[2]

[1],[2]*Department of Computer Science*
Jamal Mohamed College (Autonomous),
Tiruchirappalli–620 012.
[1]*di_george@ymail.com,* [2] *aswin.pn@gmail.com*

Abstract

Diabetes mellitus is caused by hormone deficiency anemia called insulin. The insulin secreted naturally by the pancreas of human body keeps the blood sugar level in normal. When the insulin secretion is not sufficient, the blood sugar level increases, which leads to several disorders in human body like loss of eye sight, heart disease, improper function of kidney etc. Due to some of these disorders, nearly one million people lose their life in India. Hence it is better to prevent the occurrence of these disorders, before it gets aggressive. It is possible to predict the existence of diabetes mellitus using the technology. The proposed Monotony Advanced Decision Tree using Graft(MADG) algorithm uses the Data Mining techniques for predicting the Diabetes Mellitus. This algorithm uses three important parameters such as Oral Glucose Tolerance Test (OGTT), Body Mass Index(BMI), Diastolic Blood Pressure(DBP) for extracting the rule. The standard Pima Indian Diabetic Data set is used for rule extraction. Thus the proposed algorithm can predict the possible percentage of getting diabetes earlier rather than later stage.

Key Words and Phrases: Rule Extraction, Insulin, Hormone deficiency, Pima Indian diabetic data set, Cross validation.

1 Introduction

Diabetes is a metabolic disorder characterized by the inability of the pancreas to produce enough insulin or the body's lack of ability to effectively use insulin. The disease is contributed by multiple factors, including diet, lifestyle and genes. Despite

the advancement in health care, the prevalence of diabetes is still on the rise. More than 150 million people worldwide are affected with this debilitating disease [1]. Thus, the surveillance, prevention and control of diabetes and its complications are becoming increasingly important. The disease is recognized as a major global public health problem [2]. Generally, obesity and overweight induced diabetes mellitus [3] and most patients with type 2 diabetes, which account for greater than 90% of patients with diabetes are obese [4].

The Monotony Advanced Decision Tree with graft (MADG) technique is used to increase the accuracy of rule extraction. The proposed MADG algorithm is combined with sample selection technique to get the trained data as well as testing data. The design is to apply the advanced decision tree with graft technique to generate the tree formation and defines a classification rules. An instance defines a set of path in advanced tree. The rules are extracted through MADG algorithm. It provides more accuracy then the existing Regular Covering Technique. The accuracy is reduced since the extracted rule is in scattered form. In the proposed algorithm, the old flaws are fixed more accuracy than existing algorithms. It is tested using PID, which are collected from 768 tribal people living in the Arizona province of United States of America [5]. The new rules are extracted from this dataset using MADG algorithm.

The 5x2 cross validation technique is added in MADG algorithm. This method is used to select the data sample. By selecting sample, it can get an upgraded rule extraction method. This technique is considered better than the other cross validation because of overhead rules. Therefore any rule will give more accuracy to order this room to categorize. It should be followed by the method of confusion matrix to sort it. The number of True positive, True Negative, False Positive, False Negative is derived from this confusion matrix. Training accuracy, test accuracy, number of rules, number of antecedents and standard deviation are calculated by using confusion matrix.

2 Related works

The two groups have attempted to construct prediction models for Diabetic Nephropathy (DN). cho et al. utilized a support vector machine(SVM) approach to classify DN patients among all type of diabetics, but the model was trained on an irregular and unbalanced dataset [6]. On the other hand, Leung et al. compared various machine learning and statistical methods to develop a prediction strategy for the identification of genotype phenotype risk patterns in DN [7].

A detailed survey was conducted on the application of different soft computing techniques for the prediction of diabetes. This survey was aimed to identify and propose an effective technique for earlier prediction of the disease. The data set chosen for classification and experimental simulation was based on Pima Indian Diabetic set from (UCI) repository of machine learning databases. Anand et al [8] applied neural network techniques successfully for diagnosis of type II diabetes. It proposed to compute with PCA preprocessing and higher order neural network. The problem of missing data in the analysis and decision making process was handled through PCA. Rajesh et al [9] introduced new technique for medical professionals need a re-

liable prediction methodology to diagnose diabetes. Data mining is applied to find useful patterns to help in the important tasks of medical diagnosis and treatment. It aims for mining the relationship in diabetes data for efficient classification. The data mining methods and techniques will be explored to identify the suitable methods and techniques for efficient classification of diabetes dataset in mining useful patterns. Pradhan et all [10] have presented that support vector machine (SVM) is one of the most important machine learning algorithms that has been implemented mostly in pattern recognition problem for classifying the network traffic and also in image processing for recognition. Thirumal et al [11] have discussed that data mining looks through a large amount of data to extract useful information. The usage of data mining techniques in disease prediction is reducing. One of the most common diseases among young adult is diabetes mellitus. This develops at a middle age and more common in obese children and adolescents. In order to reduce the population with diabetes mellitus, it should be detected at an earlier stage. Hence a quick and efficient detection mechanism has to be discovered.

Usually standard statistical techniques have been used in classification difficulties when dependent variable is dichotomous. Data mining applications with higher accuracy and efficiency are used by researchers with popular classification techniques similar to artificial neural networks(ANN), decision trees (DT) and random forests (RF) for medical prediction [12]. These classification techniques determine the predictor associated with outcome in addition to predicting the outcome of a disease. For this principle they used artificial neural network (ANN), logistic regression (LR), DT and Bayesian model by comparing their performances. Marco et al. compared discriminate analysis, LR, RF, classification trees, support vector machines, ANN, Multilayer perception (MLP) and radial basis function (RBF) for prediction of dementia patients. Ture et al. compared a wide range of classification techniques to predict control and hypertension groups. Morteza et al. ANN and DT predicted albuminuria in patients with type 2 diabetes mellitus by using two diverse statistical models, MLP and conditional LR. Meng et al. compared the performance of LR, ANN and DT models for predicting diabetes or pre diabetes using common risk factors.

Enhanced Regular Covering Technique (ERCT) algorithm used to rules can be extracted directly from the training data (without having to generate a decision tree first) using regular covering technique. The name comes from the notion that the rules are learned sequentially; where each rule for a given class will ideally cover many of the class's tuples. Sequential covering algorithms are the most widely used approach to mining disjunctive sets of classification rules. Each time is learned the tuples covered by the rule removed and the process repeats on the remaining tuples. This regular learning of rules is in contrast to decision tree induction. Because the path to each leaf in a decision tree corresponds to a rule can consider decision tree induction as learning a set of rules simultaneously. A basic sequential covering algorithm rules are learned for one class that time. In this way the rules learned should be of high accuracy. The rules need not necessarily be of high coverage. The process continues until the terminating condition is met, such as 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. ERCT technique is appended with the attribute test as a logical consent to the existing condition of the rule antecedent. ERCT technique considers each possible attribute test that may be added to the rule. This can be derived from the parameter attribute values which contains a list of attributes with their associated values. Typically the training data will contain many attributes, each of which may have several possible values.

3 Monotony advanced decision tree using graft (MADG) technique: a proposed method

The sample selection technique is used in MADG algorithm. It allows data set to be split into several sample sets. The samples are selected which produces best potential and removed all other samples. The training data samples are selected based on the predictions of the neural network ensemble. Cross validation technique is also used in the MADG algorithm. Cross validation is a technique to evaluate the predictive models by partitioning the original sample into a training set to train the model and a test set to evaluate the same. After selecting the sample in the MADG algorithm, the training and pruning method is also applied to it. Pruning is the process that can be reduced the complexity of decision tree while retaining good predictive accuracy. The proposed algorithm is followed two additional method such as back propagation learning and multilayer perception. The backward propagation of errors or a back propagation is a common method of training artificial neural network used in conjunction with an optimization method such as gradient descent. A multilayer perception is a feed forward artificial neural network model that maps set of input data onto set of appropriate outputs. An MLP consist of multiple layers of nodes in a directed graph with each long fully connected to the next one.

Pima Indian diabetic data set is used in the proposed MADG algorithm. There are nine types of parameters in this data set. Only a three parameter are selected as a target parameter. Tree is constructed by advance decision tree method and their rules are extracted by the proposed MADG algorithm. Then extracted rules are trimmed by the grafting technique. Grafting is a post process that can be readily applied to decision trees. It's main objective is to reclassify regions of instance space where no training data exist or where there is only misclassified data and as a result to decrease prediction error and the number of support. The errors derived from these rules are compared with the threshold value. Finally, the last fifteen rules are obtained from it. The rules extracted by MADG is more effective to prevent the pre attack by diabetes mellitus.

3.1 Monotony Advanced Decision tree using graft (MADG) Algorithm

Step1: Train the dataset from the databases through 5×2 cross validation technique.
Step2: Selected the best samples from training dataset remaining samples discarded.
Step3: Train and prune an NN using the dataset S and all of its D and C attributes
Step4: Let D' and C' be the sets of discrete and continuous attributes, respectively, still present in the network and let S' be the set of data samples correctly classified by the pruned network.
Step5: MADG (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 classifies 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 MADG (Examples(v_i), Target_ Attribute, Attributes–$\{A\}$)
Step16: End
Step17: Return Root
Step18: If support $R_i > t_1$ and error $R_i > t_2$ then
Step19: Let S_i be the set of data samples that satisfies the condition of rule R_i, let D_i be the set of discrete attributes and let C_i be the set of continuous attributes that does not appear in rule conditioin R_i.
Step20: Call continuous MADG (S_i, D_i, C_i)
Step21: Otherwise stop.

Figure 1: Monotony Advanced Decision tree using graft (MADG) Algorithm

4 Illustration

The proposed Monotony Advanced Decision tree using graft (MADG) technique uses three important parameters such as Oral Glucose Tolerance Test (OGTT), Body Mass Index(BMI), Diastolic Blood Pressure(DBP) for extracting the rule. The fifteen rules are extracted from that technique.

4.1 Extracted rules for Monotony Advanced Decision tree using graft (MADG) technique.

4.2 Confusion Matrix

If the person have diabetes in the predicted class, it gives the answer as "yes". If the person have no diabetes in the predicted class, it gives 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.

The confusion matrix for the Pima Indian Diabetes dataset is tabulated in Ta-

R1: If OGTT $\leq$ 137 then non diabetes
R2: If OGTT $\in$ (133,145) and BMI $\leq$ 32 and DBP $\leq$ 95 then Non diabetes
R3: If OGTT $\in$ (123,133) and BMI $\in$ (23,37) and DBP $\in$ (80,90) then Non diabetes
R4: If OGTT $\in$ (133,144) and BMI $>$ 40 and DBP$\in$ (80,90) then diabetes
R5: If OGTT $\in$ (121,131) and BMI $<$ 30 and DBP $<$ 90 then Non diabetes
R6: If OGTT$>$ 152 and BMI $>$ 40 and DBP$>$ 95 then diabetes
R7: If OGTT $>$ 153 and BMI $>$ 42 and DBP $>$ 93 then diabetes
R8: If OGTT $<$ 135 and BMI $<$ 31 and DBP $<$ 89 then Non diabetes
R9: If OGTT $>$ 151 and BMI $>$ 42 and DBP$>$ 93 then diabetes
R10: If OGTT $<$ 135 and BMI $<$ 32 and DBP $<$ 85 then Non diabetes
R11: If OGTT $<$ 135 and BMI $<$ 35 and DBP $<$ 90 then Non diabetes
R12: If OGTT$>$ 152 and BMI$>$ 40 and DBP$>$ 93 then Diabetes
R13: If OGTT $<$ 130 and BMI $<$ 32 and DBP $<$ 89 then Non Diabetes
R14: If OGTT $<$ 158 and BMI$>$ 45 and DBP$>$ 95 then diabetes
R15: If OGTT $<$ 141 and BMI $<$ 30 and DBP $<$ 90 then Non diabetes

Figure 2: Extracted rules for Monotony Advanced Decision tree using graft (MADG) technique.

ble 1. It shows True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

Table 1: Confusion Matrix.

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

4.3 Performance of MADG method

The training accuracy of test data set is calculated by using the following equations.

$$\text{Training Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

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

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

$$\text{TPR} = \frac{TP}{P} \tag{4}$$

$$\text{FPR} = \frac{FP}{N} \tag{5}$$

The performance of MADG technique is tabulated in Table 2 The Training Accuracy (TR ACC), Testing Accuracy (TS ACC), Number of rules, Average number of antecedents and standard deviation of TR, TS are listed in Table 2. The proposed

Table 2: Performance of MADG Technique (average of 10 runs of 10-fold cross validation [CV]).

	TR ACC (%)	TS ACC (%)	# Rules	Ave.# antecedent	TR ACC (SD)	TS ACC (SD)
STAD model	93	91	10	3	1.65	1.78
ITJ method	94	93	12	3	1.73	1.80
MADG Method	96	95	15	3	1.78	1.88

MADG method is compared with the STAD model and ITJ method. The result shows that MADG method produces higher percentage of accuracy than ITJ and STAD.

4.4 Histogram representation of MADG technique with ITJ method and STAD model

In this representation, MADG technique is compared with ITJ and STAD model. The multi-objective optimization and economics are 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 need to obtain extracted diagnostic rules with reduced accuracy and more interpretability. Needless to say, the optimal solution is obtained by using wider viable region. It provides improvements in both diagnostic accuracy and interpretability. Hence, the rule extraction technique is used to compromise between both requirements and also build a simple rule set. The results show that the method helps to take decision in well performed complex models. Chart is drawn using such accuracy, the comparison of the MADG technique, ITJ method and the results of the STAD model in the chart and to calculate accuracy in X axis on the chart. Any kind of result is said in Y axis. Following this method you have to choose the highly accuracy rules.

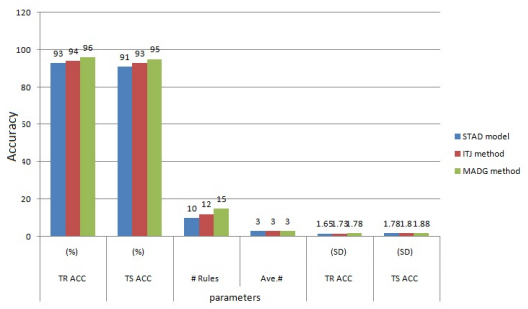

Figure 3: Histogram representation of comparison between STAD Model, ITJ method and MADG technique with accuracy, number of rules, antecedents.

5 Conclusion

The diagnosis of diabetes mellitus remains a complex problem in medical environment. The proposed MADG algorithm should be tested on more recent and complete diabetes datasets. The results of MADG is more accurate, concise and interpretable for diabetes. It is more useful for the patients who are affected by the diabetes. The rules extracted by MADG is more suitable for the prediction and prevention of diabetes mellitus.

References

[1] MansourianM, FaghihimaniE, AminiM, FarinaD. Ahybriintelligent system for diagnosing microalbuminuriain type2 diabetes patients without having to measure urinary albumin,*ComputBiolMed* **4)** (2014), 34–42.

[2] 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).

[3] Zhu J, XieQ, ZhengK. An improve d early detection method of type-2 diabetes mellitus using multiple classifier system. *In Sci* **292**,(2015),1–14.

[4] Homme MB,Reynolds KK, ValdesR, inder MW. Dynamic pharmacogenetic models in anti coag Regulationtherapy. ClinLab Med 2008;28:539–52.

[5] University of California, Irvine learning repository, `http://archive/ics.uci/edu/m/`:[last accessed 01.10.15].

[6] Cho BH, Yu H,Kim TH, Kim IY, Kim SI: Application of irregular and unbalanced data to predict diabetic nephoropathy using visualization and feature selection methods. Artif Intell Med 2014ama,**42(1)**:37–53.

[7] Leung RKK, Wang Y, Ma RCW, Luk AOY, Lam V, Ng M, So WY, Tsui SKW, Chan JCN: Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case-control chohort analysis. Bmc Nephrol (2013).

[8] Anand A. Chaudhari, S.P. Akarte,Fuzzy & datamining based disease prediction using K-NN algorithm,*International journal of innovations in Enginnering and technology* **3** (4), (2014).

[9] K. Rajesh V. Sangeetha, "Application of data mining methods and techeques for diabetes diagnosis,*international journal of engineering and innovative technology* **2** (3), (2012).

[10] Ashis Pradhan, Support vector machine—A survey,*International journal of emerging technology and advanced engineering* **2**(8), (2012).

[11] Thiurmal P.C. and Nagarajan N., "Utilizasation of data mining techniques for diagnosis of diabetes mellitus–A case stydt *ARPN journal of engneering and applied sciences* **10**(1),(2015).

[12] Yilmaz N., Inan O., Uzer M.S., "New data preparation method based on clustering algorithms for diagnosis systems of heart an diabetes diseases *J. Med. Syst* **38**, (2014), 48–59.

[13] Mansourian, M., Faghihimani, E., Amini, M., Farina, D., A hybrid intelligent system for diagnosing microalbuminuriain type2 diabetes patients without having to measure urinary albumin. Comput Biol Med. **45**, (2014), 34–42.