DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.4914

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A NEW ITJ METHOD WITH COMBINED SAMPLE SELECTION TECHNIQUE TO PREDICT THE DIABETES MELLITUS

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

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

Abstract- The purpose of this study was to generate more concise rule extraction from the ITJ method. The proposed algorithm replacing the c4.5 program currently employed in ITJ method. The algorithms that can provide further insight are needed. Rule extraction can provide such explanations. The research was consequently operated to determine twelve rules with data sets having discrete and continuous aspect. The rule derivation method recommended for strengthen ITJ method to carry out deeply classification rules. The J48 scion decision tree algorithm is generated and used for classification. ITJ method is combined with sample selection technique which is used to substantially better accuracy and provided a considerably fewer average number of rules and antecedents. The proposed method is suitable for decision making medical accept including the diagnosis of all type of diabetes mellitus. The conventional input scooping approach for the forecast of diabetes uses single classifier method for anticipate the disease, which have documented approximately high rate of efficiency. Thus the theoretical hybrid classifier capability is recommended to predict diabetes through feature relevance analysis with high accuracy rate.

Keywords: Classification rule, Data scooping, selection technique, Antecedents

I. INTRODUCTION

Diabetes is the second most common cause of death worldwide, accounting for 6% of global diabetes incidence and 9% of mortality. In 2012, 857,000 deaths were directly attributed able to diabetes [1]. It is the fifth most common type of diabetes among men (665,000 new cases, 8% of all cases) and the ninth most common among women(339,000 cases, 3% of all cases). Although substantial progress has been made regarding the knowledge and management of diabetes disease over the past several decades, approximately 40 million patients in India are suspected to have chronic diabetes condition [2]. Unfortunately, the evaluation of diabetes disease in the EU is limited due to difficulties in accessing data from individual countries.

In order to grasp a clear understanding of the actual burden of diabetes disease the prevalence of diabetes which represent the end stage of diabetes and therefore indicative of the associated mortality, need to be accurately assessed; however, such details have rarely been reported[3]. The existing method use of rule extraction algorithm with Re-RX techniques for preprocessing [4]. This combination is similar to sampling ITJ with j48 scion, however, based on the difficulty of extracting highly accurate rules, the use of sampling-selection ITJ method allowed us to achieve high accuracy while only sacrificing slightly less conciseness because although continuous ITJ provides higher accuracy, it also extracts a large number of rules [5]. The accuracy and interpretability of diagnostic rules extracted using sampling-selection ITJ were investigated based on a comparison with crisp rule extraction and two previous fuzzy rule extraction techniques. The Pima Indian Dataset from the UCI repository [6], which comprises 768 cases with two classes (diabetes or Non diabetes) and nine attributes used. Rule extraction can provide detailed explanations underlying assignments and is it therefore becoming increasingly popular; however, in the medical setting, extracted rules must be not only highly accurate, but also simple and easy to understand.

II. LITERATURE REVIEW

According to know the Pima Indians have 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. In this study were used to model the relationship between the onset of diabetes mellitus and previous risk factors for diabetes among Pima Indian data set [7].

Medical data mining discovers hidden patterns from datasets efficiently and accurately. These patterns can then be utilized for disease diagnosis and treatment. Following research method focus on using different data mining techniques for medical datasets [8] proposed a framework for diagnosis of diabetes in female patients. It is established on an altogether of neural network and support vector machine. The example dataset is from UCI data repository. Empirical conclusion display excellent classification and forecast certainty. It treated a great tool for identifying diabetes patients [9]. It is completed from decision that contemporary population of diabetes is placed on old population and can be used to retrive high chromosomal certainty [10]. The suggested exemplary handling kNN classification facility along with k-means in the existence of several pre-processing steps.

Experimental results show high accuracy with different k values [11] proposed a framework for diabetes diagnosing for Pima Indian diabetes dataset. It handling SVM classifier to anticipate the diabetic victim. The aspect selection is behave using F-score and k-mean clustering

design to obtain optimal set of appearances. High accuracy of proposed technique recommended it for disease diagnosis [12] created a soft switcher in Bayesian framework by combining elements from both averaging and switching techniques. The datasets are used to check out achievement of the proposed technique [13]. An analysis of the new literature show that broad analysis has been charged for diabetes diagnosis using ensembles[14]. Yet to be decision trees have not been analyzed appropriately for the purpose. This exploration arrangement ensemble based approach to evaluate the performance of decision tree ensembles for diabetic datasets.

Enhanced Regular Covering Technique (ERCT) algorithm recycled to guideline can be derived precisely from the training data (without having to generate a decision tree first) using regular covering technique [15]. The flag appear from the concept that the guideline is learned continuously; where each law for a given class will attractively dress many of the class's tuples. Sequential covering algorithms are the better universally used access to mining disjunctive arranged of classification rules form the topic of this division. Each time a rules is learned the tuples covered by the rule are removed and the process repeats on the remaining tuples[16]. 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 regular covering algorithm rules are studied for one class that time.The information a rule for a class C would relate the rule to cover all of the training tuples of class C and none of the tuples from other classes. The way of rules studied should be of high efficiency. If the rules need not necessarily be of high coverage, the action extend until the terminating condition is met, such as when there are no more training tuples or the character of rule exchanged is a user described doorstep. This learn one rule action asset the best rule for the recent class given the current set of training tuples[17].

Typically rules are grown in a general to specific manner. ERCT technique append by adding the attribute test as a logical consent to the existing condition of the rule antecedent, suppose in this training set D consists of Pima Indian Diabetes data. 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.

III. TO INTRODUCE ITERATION TECHNIQUE WITH J48 SCION (ITJ) TECHNIQUE

The elementary technique of choosing convenient trees have been frequently treated desirable, j48 scion tree differs in this process works on the assumption that similar objects are very likely in the same class. Hence j48 scion tree experiment to develop superior classification models at the expense of producing highly trees.

The j48 scion decision trees can experience grafting as a panel development meant broadly to reclassify parts of the detail space in which there are no training data or there are only misclassified data, as a means of decreasing prediction errors. As such this method identifies leaf regions that should be pruned and subsequently generates new leaves with novel classifications via a branching out process, necessarily producing a more complex tree. This process only allows branching that avoids the introduction of classification errors into data that have been correctly classified. As a result, rather than introducing errors, the scion technique eliminates them.

To provide a more efficient means of evaluating the supporting evidence, This algorithm is associated with grafting from the all regions of a leaf defined as those regions that result from removing all surrounding decision surfaces. Eliminating can be thought of as the contrary process to scion because it diminish decision tree intricacy while continuing a suitable degree of prediction accuracy. The comparison, scion technique increases the complexity of the tree. Although ITJ decisive that the coordinate use of these two approaches generates good results. The eliminating examines detail internal to analyzed leaves, during grafting takes into account information external to the leaves. These approaches are complementary and their combined use generally produces a lower prediction error than their separate applications.

The sample selection technique removes these data samples before building a model that distinguishes between diabetes and non-diabetes. An Neural Network (NN) ensemble is trained to identify potentially irregular or mislabeled data samples that are consistently misclassified by the majority of NNs in the ensemble are removed. The Proposed algorithm includes following techniques. 1) Ensemble creation: train an ensemble of M feed forward NNs using the available training data samples. 2) Sample selection: select training data samples based on the predictions of the NN ensemble; 3) Model generation: use the selected samples to train an NN. 4) Rule extraction: apply an NN rule extraction algorithm to obtain concise and interpretable classification rules capable of distinguishing between diabetes and non-diabetes. Sample selection is a core component of the sampling selection technique. This technique employed an NN ensemble to identify outliers in the training dataset. Removing outliers and noise prior to learn has been shown to improve the predictive accuracy of numerous learning methods. A data sample is labeled as an outlier, and subsequently discarded, if it was incorrectly classified by a proportion of NNs exceeding the threshold 'p' otherwise the sample is retained in the training dataset.

A. Iteration Technique with J48 scion (ITJ) method

Step1 : Train the dataset from the databases through10 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: Generate decision tree by using both discrete and continuous C' attributes .

Step6: For each rule Ri is generated.

Step7: Check for any base cases
Step8: For each attribute
Step9: Find the normalized information gain from splitting
Step10:Let a_best be the attribute with the highest normalized information gain
Step11: Create a decision node that splits on a_best
Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node.
Step12: If support Ki>t1 and error Ki>t2 then
Step13: Let Vi be the set of data samples that satisfies the condition of rule Ki, let Di be the set of discrete attributes and let Ci be the set of Continuous attributes that does not appear in rule condition Ki.
Step14: Call continuous ITJ(Si,Di, Ci)
Step15: Otherwise stop.

IV. ILLUSTRATION

A. Rules extracted using the ITJ method

R1: If OGTT≤ 135 then non diabetes
R2: If OGTT∈ (132,142) and BMI≤35 and DBP ≤92 then Non diabetes
R3: If OGTT ∈ (122,135) and BMI ∈ (27,35) and DBP ∈ (82,91) then Non diabetes
R4: If OGTT ∈ (138,148) and BMI >39 and DBP > 91 then diabetes
R5: If OGTT ∈ (122,132) and BMI >26 and DBP <86 then Non diabetes
R6: If OGTT>150 and BMI >36 and DBP>86 then diabetes
R7: If OGTT >153 and BMI >42 and DBP >92 then diabetes
R8: If OGTT <140 and BMI <32 and DBP>85 then Non diabetes
R9: If OGTT >152 and BMI >41 and DBP>92 then diabetes
R10: If OGTT<137 and BMI <31 and DBP<84 then Non diabetes
R11: If OGTT<134 and BMI<29 and DBP<89 then Non diabetes
R12: If OGTT>150 and BMI>39 and DBP>92 then diabetes

B. Confusion Matrix

If the person have diabetes is the predicted class, it will give the answer as “yes”. If the person have no diabetes is the predicted class, it 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 1 Confusion Matrix

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

The confusion matrix for the Pima Indian Diabetes dataset is tabulated in Table 1.It shows True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

C. Performance of ITJ method

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

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

Training Accuracy &

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

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

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

TABLE 2 Performance of ITJ method (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
ITJ method	94	93	12	3	1.73	1.80

The performance of ITJ method 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 ITJ method is compared with the Regular covering Technique and STAD model. The result shows that ITJ method produces higher percentage of accuracy. The regular covering technique, STAD model and ITJ method are tested using PID data set.

D. Histogram representation of ITJ method with regular covering technique

In this representation ITJ method is compared with STAD model and regular covering technique. The optimality of multi-objective optimization and economics 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 need to obtain extracted diagnostic rules

with reduced accuracy and more interpretability. Needless to say, if the optimal solution can be found only by best extracted rules the optimal solution is obtained using wider viable region which provides improvements in both diagnostic accuracy and interpretability rule extraction technique is used to compromise between both requirements and also building a simple rule set. The results show that the method is to take decision is well performed complex models.

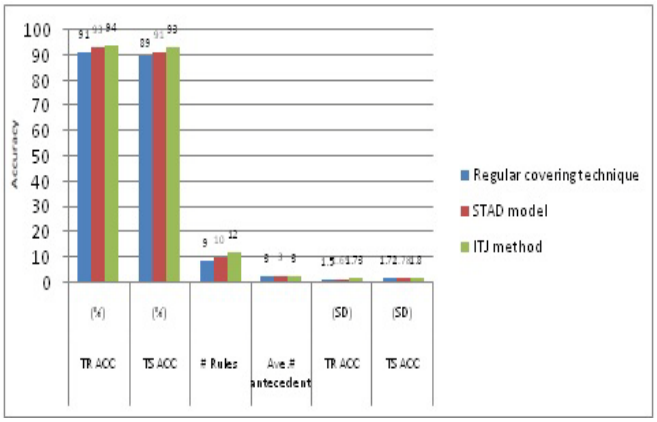

Fig.1 Histogram representation of comparison between Regular covering technique, STAD Model and ITJ method with accuracy, number of rules, antecedents.

V. CONCLUSION

The ITJ method is more accurate, concise and interpretable and therefore more suitable for decision making is medical environment. Actually high accuracy, conciseness and interpretability are achieved simultaneously by the proposed ITJ method. The ITJ method is used to be particularly in patients with diabetes mellitus with relatively high fracture risk. Hence, the diagnosis of diabetes mellitus remains a complex problem; therefore ITJ method 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] Beloufa F,ChikhMA. Design of fuzzy classifier for diabetes disease using modified artificial bee colony algorithm Comput Methods Prog Biomed 2013;112:92–103.

[2] Sapna, S., Tamilarasi, a: Data mining – Fuzzy Neural genetic algorithm in predicting diabetes. In: Research Journal on computer Engineering.(2013)

[3] NirmalaDevi M., Appavu, S., Swathi, U.V: An amalagam KNN to predict diabetes mellitus. In: International conference on Emerging Trends in computing, Communication and Nanotechnology (ICECCN), (2015)

[4] Gandhi, K.K., prajapati, N.B, Diabetes prediction using feature selection and classification. International journal of advance Engineering and Research Development (2014)

[5] Stahl, F., Johansson, R., Renard,E. Ensemble Glucose prediction in Insulin-Dependent diabetes. Data driven modeling for diabetes springer(2014)

[6] ParkJ, Eding ton DW.A sequential neural network mode l for diabetes pre- diction. ArtifIntell Med 2001;23:277–93.

[7] Gadaras,I, Mikhailov L. An interpretable fuzzy rule-based classification methodology for medical diagnosis. Artif Intell Med 2009; 47:25–41.

[8] GhazaviSN, LiaoTW. Medical data mining by fuzzy modeling with selected features. Artif Intell Med 2008; 43:195–206.

[9] Chavas ADF, Vallasco MMBR, Tanscheit R. Fuzzy rules extraction from support vector machines for multi-class classification. Neural Comput Appl2013;22:1571–80.

[10] Lekkas S, Mikhailov L. Evolving fuzzy medical diagnosis of Pima Indians dia- betes and of dermatological disease. Artif Intell Med 2010;50:117–26.

[11] Manaswini Pradhan, Dr.Ranjit kumar sahu, Predict the onset of diabetes disease using Artificial Neural network, International Journal of Computer Science & Emerging Technonolgies (E-ISSN: 2044-6004) 303 volume 2, Issue 2, April 2011.

[12] YilmazN, InanO, UzerMS. New data preparation method based on clustering algorithms for diagnosis system s of heart an diabetes diseases. JMed Syst 2014;38:48–59.

[13] MansourianM, FaghihimaniE, AminiM, FarinaD.Ahybriintelligent system for diagnosing microalbuminuriain type2 diabetes patients without having to measure urinary albumin.Comput Biol Med2014;45:34–42.

[14] 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 .

[15] 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.

[16] HunterDJ.Gene-environment interactions in human diseases.NatRevGenet2005; 6:287–98.

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