ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Hybrid Genetic-Relative Reduct Algorithm for Pre-Processing the Diabetic Dataset

Karamath Ateeq
Research Scholar, School of Computer Science, Engineering and Applications,
Bharathidasan University, Tiruchirappalli, India.

Dr. Gopinath Ganapathy
Professor & Head, School of Computer Science, Engineering and Applications,
Bharathidasan University, Tiruchirappalli, India.

Abstract: Diabetes is a high sugar problem that occurs because of the inadequate secretion of Insulin in the human body. Nowadays, data are accumulated in digital form. To extract the required knowledge from the data, Data Mining is suggested to be the best tool. Since, the accumulated data contains a lot of noisy, irrelevant and redundant data; the dataset should be pre-processed before extracting required knowledge. In this paper, a hybrid algorithm combining Genetic Algorithm and Relative Reduct Algorithm from Rough Set Theory is proposed. This algorithm is proposed to remove noisy and unwanted data. The proposed Hybrid Genetic-Relative Reduct Algorithm is compared with existing algorithms. The proposed Hybrid Genetic-Relative Reduct Pre-Processing Algorithm has reduced the number of data attributes to minimum. The number of reduced attributes and time taken for execution of the algorithm is taken to evaluate the performance of the algorithm. The results obtained support the proposed hybrid Genetic-Relative Reduct algorithm as the best pre-processor than the existing algorithms

Keywords: Pre-processing, Diabetes, Hybrid Algorithm, Genetic Algorithm, Rough Set Theory, Relative Reduct Algorithm.

I. INTRODUCTION

Diabetes Mellitus is a serious health problem which affects people around the globe. It leads to various health issues like cardio vascular disease, visual impairments, leg amputation and renal failure if not treated or diagnosed in the right time [1]. Diabetes is caused due to the lack of insulin in the blood, where Insulin is a natural hormone concealed by the pancreas. It unlocks the body cells in a way that the sugar, starch and food molecules are absorbed and utilized by the cells to generate energy for the daily activities. Increased hunger, more thirst, slow healing of wounds, blurred vision, weight loss, fatigue and frequent urination are some of the symptoms of Diabetes [2]. It is of three types namely Type 1, Type 2 and Gestational Diabetes. The Type 1 is caused when the pancreas does not secrete the Insulin needed for the body. Type 2 occurs when the body cell does not respond to the Insulin produced by the body. The third type occurs for pregnant ladies. The cost for the treatment of disease and medical care is rising very quickly than the expectations [3]. Due to increase in storing the data in the digital form, a large amount of data is accumulated in a very short time. An expert cannot process that huge data in less time to make diagnosis, prognosis and treatment.

Data mining is a significant tool for diabetes diagnosis and research. It is used to discover hidden information from diabetes data base and helps to develop the eminence of treatment with respect to healthcare industry [4]. Data collected from the repository may contain redundant or irrelevant features. Data pre-processing is an important step in the Data Mining, as the decisions are based on the value of the data and knowledge obtained from it [5]. Data Cleaning, data integrations, data transformation and data reduction are the steps in data pre-processing. Figure 1 explains the basic steps involved in Data Pre-processing.

Feature selection is needed to remove irrelevant features. Feature selection has three approaches namely filter method, wrapper method and embedded method. A feature selection process consists of four basic steps like subset generation, subset estimation, stopping condition and result confirmation. Figure 2 explains the process involved in Feature Reduction.

Candidate feature subsets needed for evaluation is generated in subset generation using a definite search strategy. Based on the assured evaluation criterion, each candidate is evaluated and compared with the previous best one. If the new candidate is better than the previous candidate, then the previous one is removed. The process is repeated till the stopping condition is reached and the obtained candidate subset is validated by prior knowledge or different tests.

The remaining section of the paper is formed as follows. Related works are presented in Section II. Section III discusses the proposed methodology and the Results and discussions are explained in Section IV. Section V gives the conclusion and future direction of the work

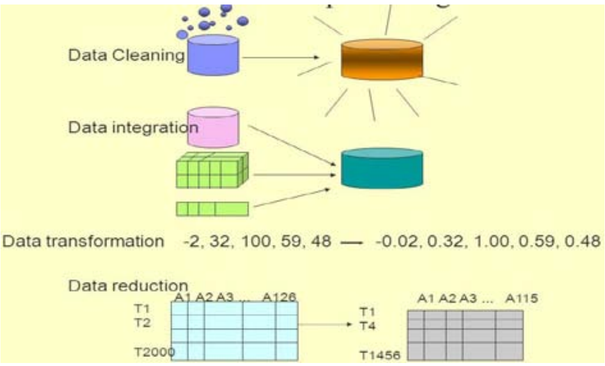

Figure 1: Steps in Pre-Processing

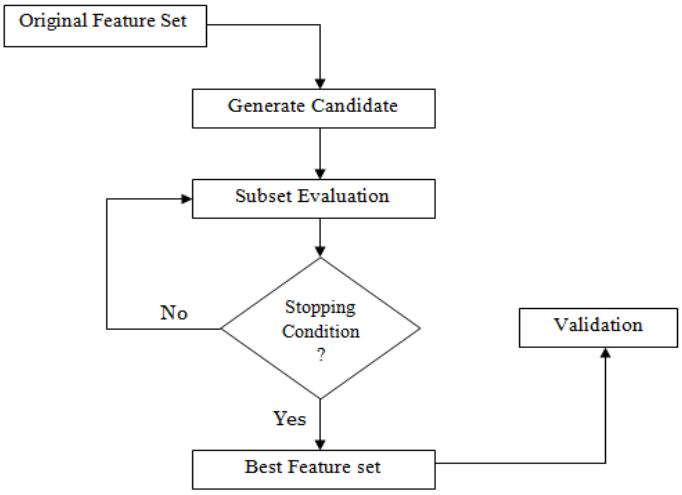

Figure 2: Steps in Feature Reduction

II. RELATED WORKS

Many automatic systems are generated in the area of data mining, machine learning, fuzzy logic and neural network for diagnosing various diseases [3]. To predict the treatment for young and old diabetic patients, Oracle Data Miner (ODM) tool with Support Vector Machine algorithm was used. Abdullah Aljumah A, et al. [6] advised that the treatment for the young diabetic patient should be postponed to avoid side effect and old diabetic patient should be started early as possible. Joseph L, et al. [7] applied Classification and Regression Tree (CART) in diabetic disease prediction and it was observed that the disease affects young people more than the old people. Azra Ramezankhani A, et al. [8] developed a model based on Decision Tree to predict the Type 2 diabetes. The model is tested with the highly populated sample dataset. Patil B, et al. [9] applied data pre-processing to improve the quality of the data. Apriori association algorithm was used to frame association rules for the class value “yes” as well as for the class value “no”.

Santhanam, et al. [10] used K-means algorithm to remove the noisiness in the data after pre-processing. Support Vector Machine classification algorithms were applied on the Pima Indian diabetes dataset. The research proved that K-means algorithm was better when compared to Support Vector Machine for diagnosing diabetes among the pregnant women of Pima. Elitist Generation Genetic Chromodynamics algorithm (EGGC) was proposed by Catalin Stoean, et al. [11] to diagnose the diabetes by using the multi model evolutionary algorithm. If–Then rules are built in present algorithm using the evolution concept. The obtained rules are high importance as they provide the reasoning rules underlying the decision-making and not only the results.

E.P. Ephzibah [12] developed a model to diagnose the diabetes based on Genetic Algorithm and Fuzzy Logic. Genetic Algorithm was used to produce the best feature subset. A new hybrid classification model was proposed by Mohammed Khanbabae, et al. [13] based on combination of clustering, feature selection, decision trees and genetic algorithm. Genetic algorithm was used for pre processing the input samples. A modified-Particle Swarm Optimization algorithm was hybrid with Least Squares Support Vector Machine (LS-SVM) by Omar S Soliman, et al. [14] for the classification of type 2 Diabetic patients. The modified Particle Swarm Optimization algorithm was used as an optimization technique for LS-SVM parameters to separate various classes. K.Rajeswari, et al. [15] proposed a new model for prediction of complications developing due to Diabetes Mellitus. Artificial Neural Network technique is used to predict the complications developing & focuses on modeling an effective Diagnosis of a special complication called neuropathy.

A Rule based genetic algorithm classifier was proposed to improve the fitness function parameter by Keshavamurthy B.N, et al. [16]. It compares the results with the probabilistic approach such as Naïve Bayes which always gives better results. M. Durairaj, et al. [17] presents the study to apply different data mining techniques in the healthcare application by using different tools on different types of disease that are commonly seen in many people. The algorithms and techniques play important role in diagnosing and predicting the disease in healthcare field. The mining techniques applied to the health data are classification, clustering, association rule mining and Naïve Bayes. After applying these methods on different kinds of disease it was observed that the accuracy was around 97.77% for cancer prediction. P. Yasodha, et al. [18] in a study includes the characteristics of diabetes and to find the number of people suffering from diabetes. This process is performed by considering the diabetic population of 249 instance and 7 unique attributes. The dataset of 249 instances are applied to WEKA tool and performed on algorithms such as Bayes network classifier, J48 Pruned tree, REP tree and Random forest. This survey was done to create awareness about the increasing population of diabetes among people all over the world and helps in knowing the status of the disease.

Rashedur M, et al. [19] presents the study that includes different data mining algorithms applied on diabetes data set; it also included the data mining measures/fitness criteria such as: sensitivity, specificity, accuracy,positive precision, negative precision and error rate. All this were performed to find out the best prediction measure for diabetes to classify tested positive or tested negative. The measure calculation is performed based on the confusion matrix that include “true positive, true negative, false positive and false negative”. D. Lavanya, et al. [20] created an awareness tool for diabetic foot problem which is often seen in diabetic patients. The model is built by using the ontology model, which includes the 4 main modules like: patient, support, report and results. The support includes the advice, action and reminder activities. The report includes the foot observations, fife style factors, medical test and symptoms. Based upon the foot observation the results like: immediate action or symptom advice or foot reminders are sent to the patients. This tool is helpful in raising the awareness of dangers developed as diabetic foot in diabetes patients.

K.R. Lakshmi, et al. [21] in the Diabetes Early Warning System considers the non laboratory data set of diabetic and non diabetic patients. The data includes the attributes such as age, gender, height, weight and family history. Found achieving the accuracy of 90% on hyper tension patients and 85% on diabetic patients. Also found out the number of people under risk were 75%. The algorithms used for this analysis are AdaBoost and C4.5. It was helpful in predicting of diabetes in new patients. The Temporal reasoning system developed by Akash Rajak [22] which consists of three subsystems: Nutria-Diet subsystem, Insulin-Glucose subsystem and Therapy Planner and Diagnosis subsystem. Artificial intelligence is used for temporal reasoning task. This system helps to plan the therapy for the diabetic patients based on the diagnosis performed on patient earlier, which is extracted from the database.

Vaishali Jain, et al. [23] used Fuzzy logic to improve the prediction rate of the diabetes. Based on the knowledge of patient's diagnosis and experience, the system will predict the diabetes. The crisp data is converted to fuzzy data, then by using the IF-THEN rule the fuzzy input is converted to fuzzy output. The decision making algorithms are used to perform the required operation and Defuzzification helps to get the crisp set from fuzzy set. Sivagowry S, et al. [24] used WEKA tool as a pre-processing tool to remove unwanted noisy data from dataset.

III. PROPOSED MODEL

The proposed model combines the feature of Genetic Algorithm and Relative Reduct Algorithm from Rough Set Theory. The attributes are classified as Conditional and Decision Attributes. At the initial state the variable R is initialized to null value. And the variable γ_{best} is assigned to 0. The γ_{best} value is stored temporarily in another variable γ_{tmp} and R is stored in T.

The consistency of the data set is checked after removing every. If the decision table is consistent, the attributes is removed and the reduced data set is stored. If the classification accuracy of the Conditional Attributes obtained is greater than the classification accuracy of the Decision attributes ($\gamma_{RU(X)}$ (C)>γ_X(D)) , then the first Generation of Offspring is constructed. The obtained attributes are selected and then mutation and crossover operations taken place.

The obtained Reduct set is stored in T. The Decision Attributes which are best are stored in the variable $\gamma_{best.}$ The obtained reduct set is stored in R. The process goes on repeating until the optimal data set is obtained. The obtained output is Reduct Set R. After obtaining the optimal data set, the algorithm terminates.

Figure 3 shows the flow of the proposed Hybrid Genetic – Relative Reduct Algorithm for pre-processing the diabetic dataset.

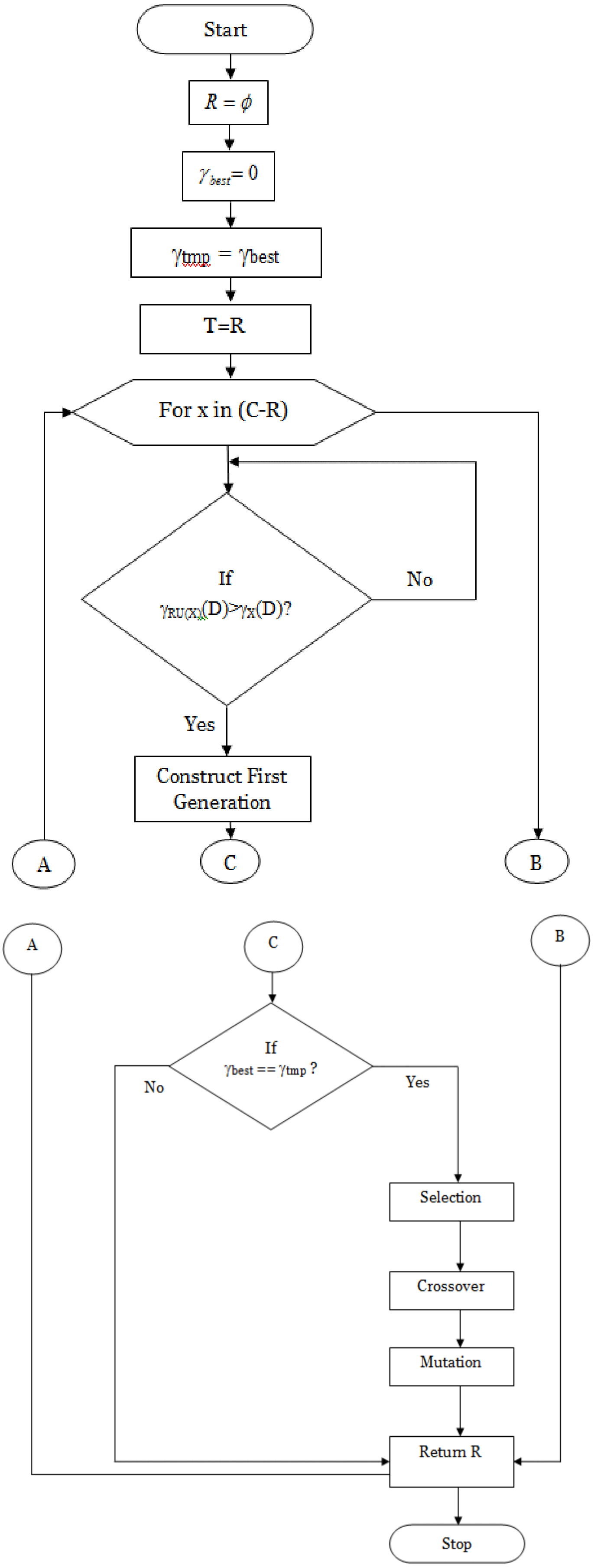
Start
$R = \phi$
$\gamma_{best} = 0$
$\gamma_{tmp} = \gamma_{best}$
T=R
For x in (C-R)
If
$\gamma_{RU(X)}(D) > \gamma_X(D)$?
No
Yes
Construct First Generation
A
C
B
A
C
B
If
$\gamma_{best} == \gamma_{tmp}$?
No
Yes
Selection
Crossover
Mutation
Return R
Stop

Figure 3: Proposed hybrid algorithm for Pre-Processing

The pseudo code for proposed Hybrid Genetic-Relative Reduct Algorithm:

C-> Conditional Attributes D-> Decision Attributes
Input: Data set

```
Step 1:   Start
Step 2:   R=$\phi$
Step 3:   $\gamma_{best} = 0$
Step 4:   do
Step 5:       $\gamma_{tmp} = \gamma_{best}$
Step 6:       T=R
Step 7:       for x in (C-R)
Step 8:           If $\gamma_{RU(X)}$ (C)>$\gamma_X$(D)
Step 9:   Construct the First Generation
Step 10:                  Selection
Step 11:                  Crossover
Step 12:                  Mutation
Step 13:              T=RU{x}
Step 14:              $\gamma_{best} = \gamma_c$(D)
Step 15:              R=T
Step 16:  Until $\gamma_{best} == \gamma_{tmp}$
Step 17:  Return R
Step 18:  End
```

Output: Optimal Data set

IV. Results and Discussion

The Dataset taken for experimentation is Pima Indian diabetic dataset with 793 instances and 9 attributes and US diabetic dataset with 500 instances and 50 attributes [25]. The proposed algorithm is applied on both the data set and results are observed. The observed results are compared with some existing work to prove the best of the proposed algorithm. The proposed algorithm works significantly in both the data set and reduced the number of attributes to minimum in minimal time. Obtained results are tabulated below. The Pima Indian diabetic dataset has consists of numerical values alone. The dataset is experimented with numerical data and the numerical data is converted to nominal data and again experimented the results obtained in both the cases are tabulated too.

TABLE I: Comparison of the reduced attributes using different algorithm

Dataset	No. of Reduced Attributes			
	Particle Swarm Optimization Algorithm	Genetic Algorithm	Relative Reduct Algorithm	Proposed Hybrid Algorithm
Pima Indian diabetic dataset (Numerical Value)	5	8	8	3
Pima Indian diabetic dataset(Nominal Value)	7	5	5	3
US diabetic dataset	10	8	7	6

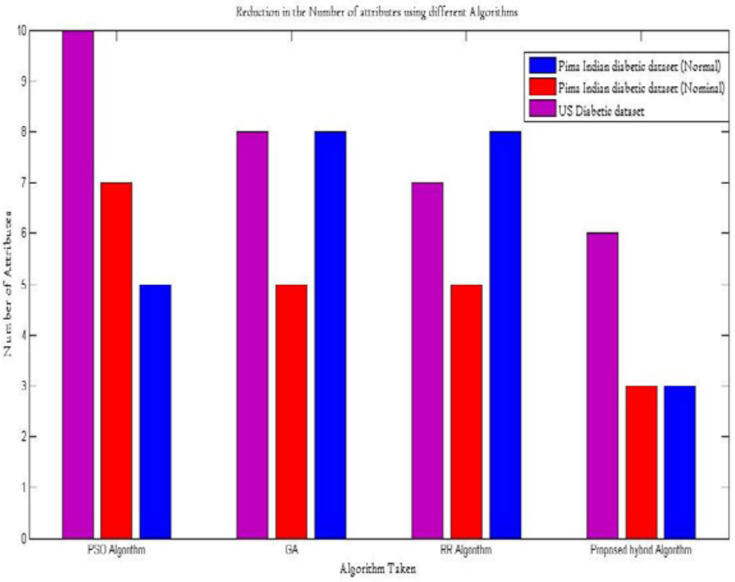

Figure 4: Comparison of reduced attributes using existing and proposed Algorithm

TABLE II: Comparison of the execution time using different algorithm

Dataset	Time taken for execution (sec)			
	Particle Swarm Optimization Algorithm	Genetic Algorithm	Relative Reduct Algorithm	Proposed Hybrid Algorithm
Pima Indian diabetic dataset (Numerical Value)	18	14	13	10
Pima Indian diabetic dataset(Nominal Value)	18	14	13	10
US diabetic dataset	56	50	52	43

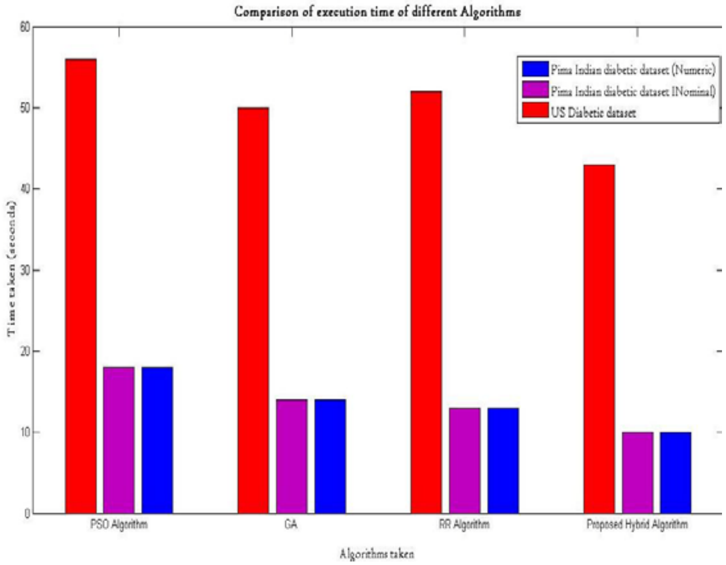

Figure 5: Comparison of Execution time

V. Conclusion

The proposed hybrid algorithm which combines the Genetic Algorithm and Relative Reduct Algorithm from Rough Set Theory performs well as pre-processor. Also it has reduced the number of features to minimum. It has produced an optimal dataset in minimum time. The algorithm is tested on Pima Indian diabetic dataset in two ways and US dataset. The promising results obtained support the hybrid algorithm in Feature Reduction of diabetic dataset. The future work is proposing a classification algorithm to classify the optimal dataset to suggest the required diet or medicine for the diabetic patients based on their disease type.

VI. REFERENCES

[1] World Health Organization, http://www.who.int/topics/diabetes_mellitus/en/, (Last access date: 30th November 2016)

[2] Srideivanai Nagarajan, RM.Chandrasekaran, “Diagnosing Diabetes using Data Mining Techniques”, International Journal of Engineering Sciences & Research Technology, Vol. 4, No. 11, November 2015.

[3] Priyanka khare, Dr. Kavita Burse, “Feature Selection using Genetic Algorithm and Classification using Weka for Ovarian Cancer”, International Journal of Computer Science and Information Technology, Vol. 7, No. 1, pp. 194-196, 2016.

[4] Miroslav Marinov, MS Abu Saleh Mohammad Mosa, M.S Illhoi Yoo, and Suzanne Austin Borea, “Data Mining Technologies for Diabetes: A Systematic Review”, Journal of Diabetes Science and Technology, Vol. 5, No. 6.

[5] E. Sreedevi and Prof. M. Padmavathamma, “Design and Development of Hybrid Genetic Classifier Model for Prediction of Diabetes”, International Journal of Modern Trends in Engineering and Research, pp. 260-265, 2016.

[6] Abdullah Aljumah A, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, “Application of data mining: Diabetes healthcare in young and old patients”, Journal of King Saud University – Computer and Information Sciences, Vol 25, pp. 127 – 136.

[7] Joseph L. Breaulta B, Colin R. Goodall.C.D, Peter J. Fose B, “Data mining a diabetic data warehouse”, Artificial Intelligence in Medicine, Vol -26, pp- 37–54, 2002.

[8] Azra Ramezankhani A, Omid Pournik B,C, Jamal Shahrabi D, Davood Khalili A, E, Fereidoun Azizi F, Farzad Hadaegh A, “Applying decision tree for identification of a low risk population for type 2 diabetes- Tehran Lipid and Glucose Study”, Diabetes research and C.linical Pract ice, vol 5, pp- 391 – 398, 2014.

[9] Patil.B.M, Joshi.R.C, Durga Toshniwal, “Associat ion rule for classificat ion of t ype-2 diabetic patients”, IEEE - Second International Conference on Machine Learning and Computing, DOI 10.1109/ICMLC, Pg-67, 2010.

[10] Santhanam T, Padmavathi. M. S, ”Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis”, Procedia Computer Science, Vol – 47, pp-76 – 83, 2015.

[11] Catalin Stoean, Ruxandra Stoean, Mike Preuss and D. Dumitrescu, “Diabetes Diagnosis through the Means of a Multimodal Evolutionary Algorithm”, Elsevier, pp.76-82, 2015.

[12] E. P. Ephzibah, “Cost Effective Approach on Feature Selection Using Genetic Algorithms and Fuzzy Logic For Diabetes Diagnosis”, International Journal on Soft Computing (IJSC), Vol.2, No.1, February 2011.

[13] Mohammad Khanbabaei and Mahmood Alborzi, “The Use Of Genetic Algorithm, Clustering and Feature Selection Techniques in Construction Of Decision Tree Models For Credit Scoring”, International Journal of Managing Information Technology (IJMIT), Vol.5, No.4, November 2013.

[14] Omar S Soliman et al “Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine”, International Journal of Computer Trends and Technology (IJCTT) – volume 8 number 1– Feb 2014.

[15] Rajeswari K, Vaithiyanathan V, Gurumoorthy T, “Modeling Effective Diagnosis of Risk Complications in Type 2 Diabetes - A Predictive model for Indian Situation”, European Journal of Scientific Research, Vol. 54, No. 1, June 2011.

[16] Keshavamurthy B. N, Asad Mohammed Khan & Durga Toshniwal, “Improved Genetic Algorithm Based Classification”, International Journal of Computer Science and Informatics (IJCSI), Volume-1, No. 3.

[17] M. Durairaj, V. Ranjani, “Data Mining Applications in Healthcare Sector: A Study”, International journal of scientific & technology research, Vol. 2, No. 10, October 2013.

[18] P. Yasodha, M. Kannan, “Analysis of a Population of Diabetic Patients Databases in WEKA Tool”, International Journal of Scientific & Engineering Research, Vol. 2, No. 5, May-2011.

[19] Rashedur M. Rahman, Farhana Afroz, “Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis”, Journal of Software Engineering and Applications, Vol. 6, 85-97, 2013.

[20] D. Lavanya, K. Usha Rani, “Performance Evaluation of Decision Tree Classifiers on Medical Datasets”, International Journal of Computer Applications, Vol. 26, No.4, 2011.

[21] K. R. Lakshmi, S.Prem Kumar, “Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability”, International Journal of Scientific & Engineering Research, Vol. 4, No. 6, 2013.

[22] Akash Rajak, “A Temporal Reasoning System for Diagnosis and Therapy Planning”, I.J. Information Technology and Computer Science, Vol. 12, pp. 23-29, 2015.

[23] Vaishali Jain, Supriya Raheja, “Improving the Prediction Rate of Diabetes using Fuzzy Expert System” I.J. Information Technology and Computer Science, Vol. 10, pp. 84-91, 2015.

[24] Durairaj M. and S. Sivagowry, “A pragmatic approach of preprocessing the data set for heart disease prediction”, International journal of Innovative Research in computer and communication Engineering, Vol. 2, No.11, 2014.

[25] https://archive.ics.uci.edu/ml/datasets/Diabetes (Last accessed on December 15, 2016).