International Journal of Grid and Distributed Computing
Vol. 10, No. 9 (2017), pp.47-56
http://dx.doi.org/10.14257/ijgdc.2017.10.9.04

Predicting Diabetics Accuracy Using Rough Set Clusters

Shantan Sawa, H. Balaji[1], N. Ch. S. N. Iyengar[1] and Ronnie D. Caytiles[2]

SCOPE, VIT University, Vellore-632014
[1]Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, India
[2]Multimedia Engineering Department, Hannam University, Daejeon, Korea
shantansawa@gmail.com, rdcaytiles@gmail.com,
srimannarayanach@sreenidhi.edu.in

Abstract

The primary objective of this paper is to develop and propose a model using the concepts from Rough Set Theory to cluster the patients in the diabetic dataset. The model to be developed incorporates Rough Clustering of the dataset, and from the clusters formed, compute the accuracy on the testing data. Rough Clustering will help splitting the data into clusters of patients that suffer from Diabetes Mellitus and the ones which do not. As a result, the patients suffering from Diabetes Mellitus will be clustered together and will provide us with the average values of the features used in the model for data clustering. The results obtained will provide more depth in the field of rough clustering for diabetes as the number of studies done on diabetes using rough set theory are few to none.

Keywords*: Rough Set Theory, Diabetes Mellitus, Rough Clustering*

1. Introduction

Diabetes mellitus, commonly known as diabetes, is group of metabolic diseases which marks high blood sugar levels over a prolonged period. As per the survey done by International Diabetes Federation, in 2015, it was estimated that 415 million individuals are affected by diabetes around the world. And the number is estimated to rise to 642 million individuals by the year 2040. Type 1 diabetes mellitus (T1DM) itself comprises of roughly around 10% of the reported cases with diabetes. T1DM usually affects children where the causes are of the disease is unknown and no known ways to prevent the T1DM. The study [9] performed and published by Jane L. Chiang et al, estimated about 80000 children showing symptoms and developing the disease each year. The motivation behind pursuing a working model in the field of T1DM is to guide the doctors and the guardians of the affected children to catch the disease in its nascent stage and based on the severity of diabetes, accordingly provide apt treatment to the affected children.

Type 1 diabetes is a disorder in which the pancreas can no longer produce insulin. It is also called juvenile diabetes primarily because of it’s presence in both children and adults. However, statistics show that only 5% of all the type 1 cases are diagnosed in adulthood. It is sometimes referred to as insulin-dependent diabetes mellitus and has no cure. If one has it, they must take insulin to survive. According to the WHO (World Health Organization) the number of children having type 1 diabetes is very high as mentioned in the motivation. Hence it can be said that Diabetes is a serious chronic disease. Doctors usually take patients’ blood samples and check the sugar concentration in their blood in order to diagnose them with diabetes. This is a highly time- consuming process and there are many other features which need to be reviewed while attempting to detect whether a patient is diabetic or not. These other factors are, family history, body mass index and

Received (June 20, 2017), Review Result (August 30, 2017), Accepted (September 15, 2017)

ISSN: 2005-4262 IJGDC

age. If a patient's ancestors show a presence of diabetes, then there is a high chance of the patient exhibiting symptoms of diabetes as well. Presently, there is no tool to detect if a person has type 1 diabetes and how severe the diabetes will affect him/her. Hence the need arises for some sort of efficient application to predict the onset of type 1 diabetes in patients for a quicker diagnosis for better safety.

In the real world, data isn't crisp. But there exists some gray areas, some roughness among the objects in the data-set. As a result, in the real world applications, there are hardly any crisp data available. Thus, it becomes impossible to create crisp clusters from from real world data. Hence, in order to tackle this issue of lack of crispness, or involvement of roughness, in the data-set, implementing the concepts of Rough Set theory plays a significant part in developing models that take the rough nature of the data-set into account and accordingly form clusters of the data.

2. Literature Survey/Related Work

[1] In 2009, AsmaShaheen Khan, Waqas Ahmed proposed an Intelligent decision support system in diabetic ehealth care from the perspective of elders. The proposed system stores the patients' information and gives them optimal advices according to their condition entered by them. It also provides adequate and detail information about the patient to the health-care providers that help them to take an optimal decision about the patients. [2] In 2012, Tawfik Saeed Zeki, Mohammad V. Malakooti, Yousef Ataeipoor, TalayehTabibi proposed an expert system for diabetes diagnosis. After data acquisition and designing a rule-based expert system, this system has been coded with VP_Expert Shell and tested in ShahidHasheminezhad Teaching Hospital affiliated to Tehran University of Medical Sciences and final expert system was presented which could diagnose all kinds of diabetes. [15] In 2014, Gaganjot Kaur, Amit Chhabra proposed classification which used Decision tree algorithm to predict class whether patient is diabetic or not. The labelled data is feed as input. The leaf of the tree j48 acts as the class labels. The information gain is calculated for each attribute. Then the gain in information is calculated that would result from a test on the attribute. [12] In 2003, Margret Anouncia S., Clara Madonna L. J., Jeevitha P., Nandhini R. T. proposed a design for a Diabetic Diagnosis System using Rough Sets where the authors created a knowledge base from the existing data- set using upper and lower approximations and when a new incoming data is collected and evaluated to compute the equivalence classes. These equivalence are then compared against the knowledge, the system helped the user discern the type of diabetes. [4] In 2015, Rahman Ali, Jamil Hussain, Muhammad Hameed Siddiqi, Maqbool Hussain and Sungyoung Lee, proposed a Hybrid Rough Set Reasoning Model for Prediction and Management of Diabetes Mellitus. When a new incoming data is evaluated and compared against the knowledge base derived from the rough classification to classify the type of diabetes of the patient. [6] In 2015, AishwaryaIyer, S. Jeyalatha and RonakSumbaly conducted a comparative study where the authors used multiple data mining techniques to classify the PIMA Indians Diabetic dataset. The techniques used for feature extraction and classification were decision tree and Naive Bayesian Classifier. [8] In 2015, Áurea Celeste Ribeiro, Allan Kardec Barros, Ewaldo Santana, José Carlos Príncipe, proposed a redundancy reduction preprocessor that eliminated the redundant attributes from the data set. The authors used the PIMA Indians Diabetic dataset available on UCI's online repository for their model. In 1998, Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. have used the data set to develop a model implementing the ADAP learning algorithm to predict the onset of diabetes mellitus.

3. Motivation

Type 1 diabetes mellitus (T1DM) itself comprises of roughly around 10% of the reported cases with diabetes. T1DM usually affects children where the causes are of the disease is unknown and no known ways to prevent the T1DM. The study [9] performed and published by Jane L. Chiang et al, estimated about 80000 children showing symptoms and developing the disease each year. The motivation behind pursuing a working model in the field of T1DM is to guide the doctors and the guardians of the affected children to catch the disease in ascent stage and based on the severity of diabetes, accordingly provide treatment to the affected children.

4. Experimental Setup

Author names RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. RStudio was written in C++ language and uses the Qt framework for its graphical user interface for tasks such as plotting of graphs and charts. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux).

The following are the IDE Features offered by RStudio:

- RStudio has integrated support for Git and Subversion
- Rstudio supports authoring HTML, PDF, Word Documents, Slide Shows.
- RStudio supports interactive graphics with Shiny and ggvis.
- RStudio integrates the tools used in R into a single environment.

The following libraries have been used for the model development phase:

- RoughSets: Library consisting of functions for data analysis using RST and FSRT. The library is used for converting the data set into Decision Tables and to compute the discernability and indiscernability matrix from the decision tables.
- RoughSetKnowledgeReduction: Simplification of Decision Tables using RST. Incorporates the functions for reducts computation and feature selection.
- SoftClustering: The library contains various soft clustering algorithms to be implemented on the dataset. The function RoughKMeans_LW() is used to implement Lingras and Wests’ rough clustering algorithm to discern clusters from the dataset and the plotRoughKMeans() function to plot the rough clusters.
- caret: Classification and Regression Training. The library is used for training the dataset and testing the model for accuracy and prediction.

discretization: The library provides the user with various functions such as chi2() for data processing and discretization of numeric data tables for classification. The data set considered for the development analysis of the model is the Pima Indians Diabetes data set which is available online on University of California, Irvine, Machine Learning Repository. The data set has been provided by the National Institute of Diabetes and Digestive and Kidney Diseases. The data set consists of relevant information of patients of Pima Indian Heritage. There are 768 instances, with 500 testing negative for diabetes and the remaining 268 instances testing positive for diabetes.

5. Methodology

The proposed system follows the following steps as shown in the Figure 1, which is the system design for the model.

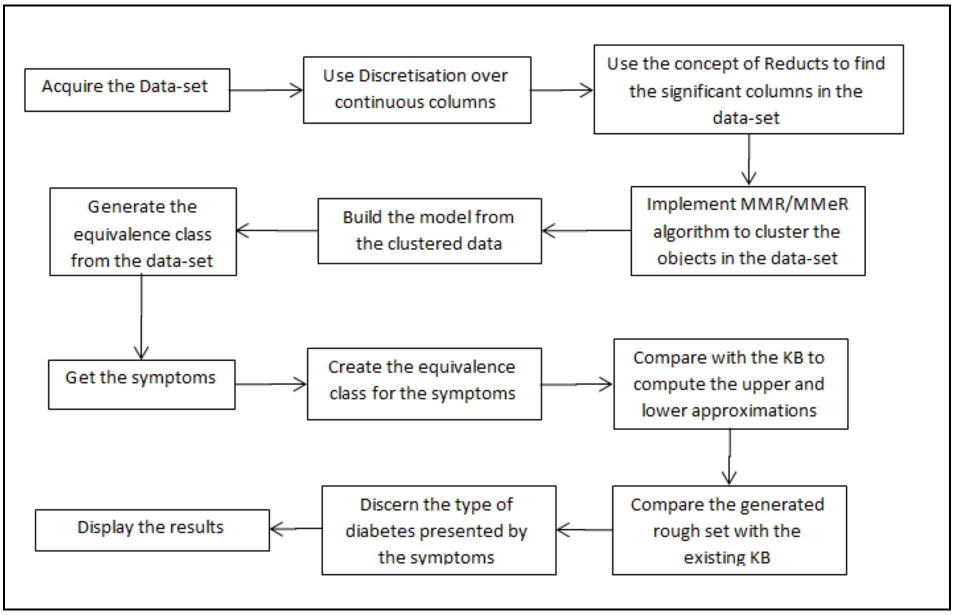

Acquire the Data-set
Use Discretisation over continuous columns
Use the concept of Reducts to find the significant columns in the data-set
Implement MMR/MMeR algorithm to cluster the objects in the data-set
Build the model from the clustered data
Generate the equivalence class from the data-set
Get the symptoms
Create the equivalence class for the symptoms
Compare with the KB to compute the upper and lower approximations
Compare the generated rough set with the existing KB
Discern the type of diabetes presented by the symptoms
Display the results

Figure 1. System Design Depicting the Processes Involved in the Rough Set Theory Model

Once the data set is loaded on the platform, we must discern the features that needed to be selected for the entire clustering process. First, we run the correlation matrix to eliminate the possibility of attributes having similar trends.

Pearson’s correlation coefficient methodology has been adopted for the computation of the correlation matrix. The method computes the correlation between two variables and provides the value within the range of -1 to +1. This algorithm is chosen as it is the traditional method of correlation computation and the accuracy increases with the increase in the sample size.

For two variables X and Y, with standard deviations σ_x and σ_y respectively and means μ_x and μ_y respectively, the correlation coefficient ρ(x,y) is given by the formula-

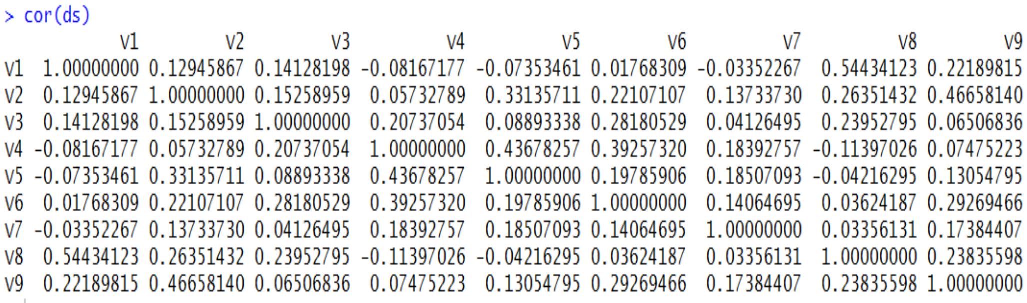
$$px, y = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

Where **E** is the expectation.

In order to interpret the values better, cut off values for the absolute coefficient have been established for better understanding.

Perfect 0	- No correlation
(0, 0.35]	- Weakly correlated
(0.35, 0.7]	- Moderately correlated
(0.7, 1)	- Strongly correlated
Perfect 1	- Perfectly correlated

Table 1. Correlation Matrix on the PIMA Indians Data-set



```
> cor(ds)
            v1         v2         v3          v4          v5         v6          v7          v8         v9
v1  1.00000000 0.12945867 0.14128198 -0.08167177 -0.07353461 0.01768309 -0.03352267  0.54434123 0.22189815
v2  0.12945867 1.00000000 0.15258959  0.05732789  0.33135711 0.22107107  0.13733730  0.26351432 0.46658140
v3  0.14128198 0.15258959 1.00000000  0.20737054  0.08893338 0.28180529  0.04126495  0.23952795 0.06506836
v4 -0.08167177 0.05732789 0.20737054  1.00000000  0.43678257 0.39257320  0.18392757 -0.11397026 0.07475223
v5 -0.07353461 0.33135711 0.08893338  0.43678257  1.00000000 0.19785906  0.18507093 -0.04216295 0.13054795
v6  0.01768309 0.22107107 0.28180529  0.39257320  0.19785906 1.00000000  0.14064695  0.03624187 0.29269466
v7 -0.03352267 0.13733730 0.04126495  0.18392757  0.18507093 0.14064695  1.00000000  0.03356131 0.17384407
v8  0.54434123 0.26351432 0.23952795 -0.11397026 -0.04216295 0.03624187  0.03356131  1.00000000 0.23835598
v9  0.22189815 0.46658140 0.06506836  0.07475223  0.13054795 0.29269466  0.17384407  0.23835598 1.00000000
```

From the above table, it is evident that no two attributes are strongly correlated. Hence, no attribute is dropped after the table. We continue to compute the reducts from the data-set for feature extraction to reduce the complexity during computing the clusters, at the same time, maintaining the integrity of the data-set.

In 1998, Jan Komorowski[16] documented the various concepts of Rough Set Theory in their paper, they have a well documented section dedicated for reducts. The authors discuss that reducts are a set of attributes that are used for feature selection process which reduce the dimensionality of the information system, but preserve the integrity of the data and help in eliminating redundant attributes from the data set.

Unfortunately, computing a minimal is a NP-Hard problem. But there are good heuristic algorithms ([17][19]) based on modified genetic algorithm to provide us with multiple reducts for the information system.

To find the reducts, we first compute the discernibility matrix. Discernibility matrix is a*n* x *n* matrix, where n is the number of attributes in the data set.

The entries of the matrix, c(i,j) are computed using the given formula- $c_{i,j} = \{a \in A \,|\, a(x_i) \neq a(x_j)\}\, for\, i, j = 1,...,n$. Where *A* is the set of attributes and *a* is an object of *A*. $x_i$is the value of the object for the corresponding attribute *a*.

After the discernibility matrix is computed, we proceed to discern the discernibility function *fA*, which is given by-

$$f_A(a_1^*,...,a_m^*) = \wedge\{\vee c_{i,j}^* \,|\, 1 \le j \le i \le n, c_{i,j} \neq \phi\}$$

A set of all the prime implicants of the discernibility function *fA* determines the reducts of the data set. Once the reducts are computed and the data-set is reduced, we deploy the rough clustering algorithm as proposed by P. Lingras [20] and Georg Peters. The rough clustering algorithm takes the lower and upper approximations of the features and to incorporate the roughness in the data and accordingly cluster the similar data points to discern the centroids in the model.

The algorithm to compute the centroids as given by P. Lingras, In which the clusters formed, we split the data-set into 70% for training the model and 30% to validated and verify the results from the trained model. The testing set is validated on the basis of distance from the centroids of the clusters.

6. Simulation and Results

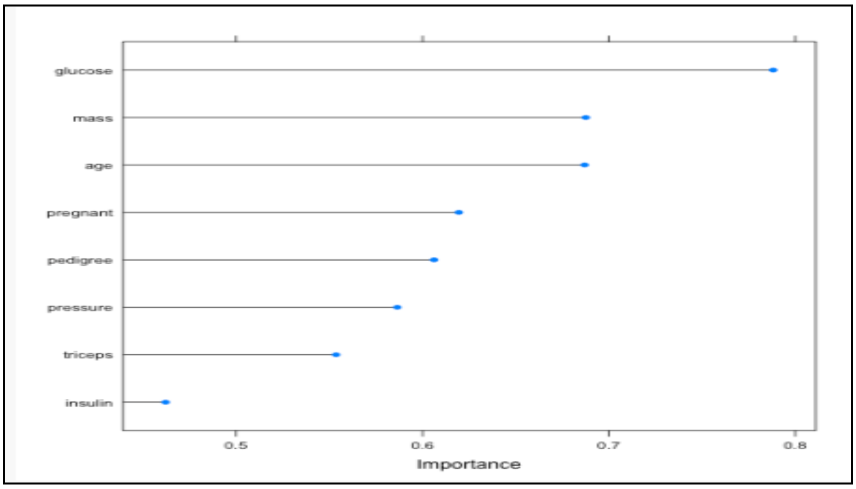


Figure 2. Ranking the Attributes for Feature Selection

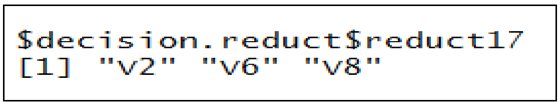
Table 2. Attributes Arranged by Priority with Assigned Weights

Attribute	Assigned Weight
Glucose	0.8
BMI	0.7
Age	0.7
Pregnancies	0.615
Diabetes Pedigree Function	0.61
Blood Pressure	0.59
Skin Thickness	0.55
Insulin	0.47

On the basis of correlation matrix and the feature selection process, the following attributes take precedence over the other features-

- Plasma Glucose Concentration
- Body Mass Index
- Age

On computing all the reducts to reduce the complexity of the execution, the data-set generates 28 unique reducts. 19 of which are three attribute reducts and the remaining 9 are four attribute reducts. After comparing with the correlation matrix and ranking the importance of the attributes, reduct 17 is selected for the clustering process of the data set.



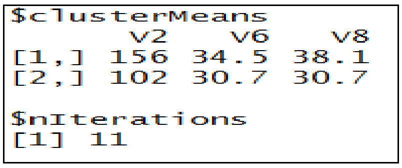
```
$decision.reduct$reduct17
[1] "V2"  "V6"  "V8"
```

Figure 3. Decision Reduct 17

On selecting the reducts from the data-set and dropping the non-significant features, we proceed with rough clustering to cluster the data-set. The data set is split into 70% for training and 30% for testing and validating the data-set.

For comparison purposes, we have even computed the centres implementing the classic Hard K-means clustering algorithm. The centroids for the Hard K-means algorithm were computed after 11 iterations over the data-set.

The instances are split into “Diabetic” and “Non- Diabetic” clusters. The first row in ClusterMeans is the value for “Diabetic” centre of the particular attribute. Meanwhile, the second row of ClusterMeans is the value for the “Non- Diabetic” centre of the attribute. The values of the centroids for the individual attributes in the reducts is listed in the figure below.

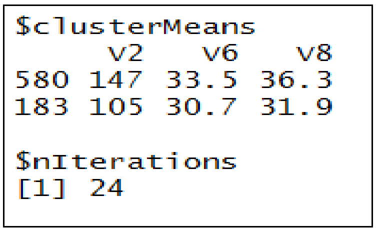


```
$clusterMeans
      V2   V6   V8
[1,] 156 34.5 38.1
[2,] 102 30.7 30.7

$nIterations
[1] 11
```

Figure 4. Cluster Means for the Attributes using Hard K-Means Algorithm

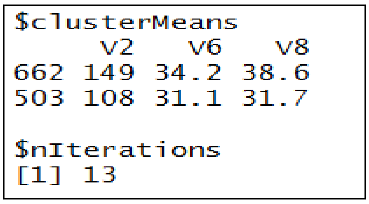
Meanwhile, the rough clustering algorithm as proposed by P. Lingras is implemented on the 70% of the data to compute the centres for the attributes. The same algorithm is then ran on the remaining 30% of the data to test and validate the values of the centres computed to determine the accuracy achieved by the rough clustering algorithms when implemented on a dataset.



```
$clusterMeans
     V2   V6   V8
580 147 33.5 36.3
183 105 30.7 31.9

$nIterations
[1] 24
```

Figure 5. Cluster Means for the Attributes using Rough K-Means Algorithm on the Training Set



```
$clusterMeans
     V2   V6   V8
662 149 34.2 38.6
503 108 31.1 31.7

$nIterations
[1] 13
```

Figure 6. Cluster Means for the Attributes using Rough K-Means Algorithm on the Testing Set

The plots for the rough clusters from the data set are plotted using the plotRoughKMeans() function which maps the data points on a 2D plane with the centres marked in the shape of a green circle which is the centre for a patient with no onset of Diabetes. Whereas, the red box is the centre for diabetic patients.

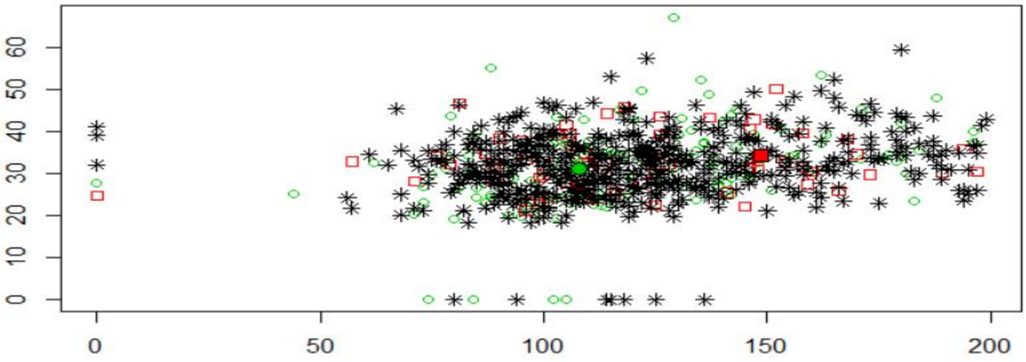


Figure 7. Rough Clusters Plot on a 2D Plane with the Centres for Training Data

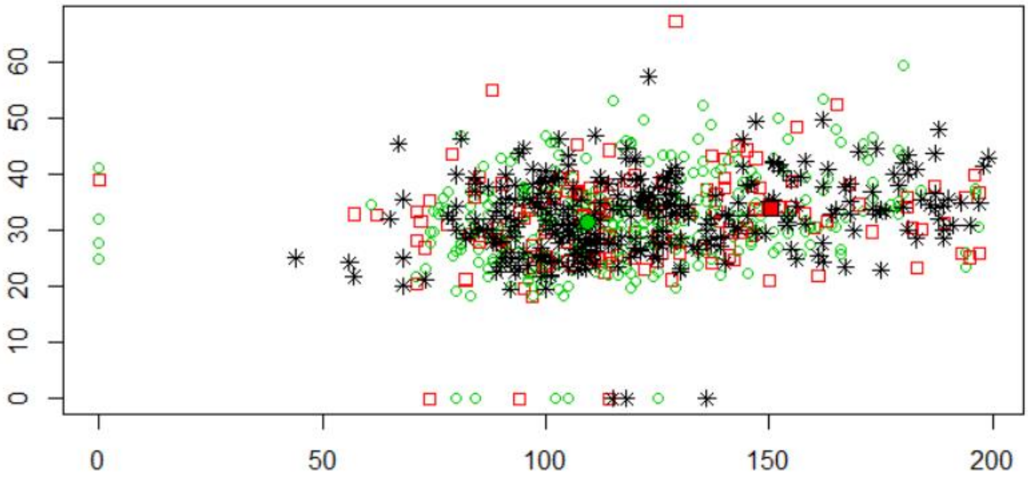


Figure 7. Rough Clusters Plot on a 2D Plane with the Centres for Testing Data

Table 3. Tabulating the Centre Values for Each Individual Attributes for the Various Methods

	V2	**V6**	**V8**
Hard K-means	(156, 102)	(34.5, 30.7)	(38.1, 30.7)
Rough K-means (Training)	(147, 105)	(33.5, 30.7)	(36.3, 31.9)
Rough K-means (Testing)	(149, 108)	(34.2, 31.1)	(38.6, 31.7)

The centre values for the Plasma Glucose Concentration (V2) for non-diabetic patients has an error of < 3%, and for diabetic patient is <1.5 %. The centre values for the Body Mass Index (V6) for non-diabetic patient has an error of < 1.5%, and for diabetic patients is < 2.1%. The centre values for Age (V8) for non diabetic patients has an error of < 0.7, and for diabetic patients is < 6.5%

Table 4. Tabulating the Error Percentage for Centres Values for Each Individual Attributes

Error %	V2	V6	V8
Non-diabetic	2.86	1.3	0.63
Diabetic	1.36	2.09	6.34

7. Conclusion

Type The Rough Clustering of the data-set was successfully trained, tested and implemented on the data set [5]. The results obtained were above satisfactory and can be further improved by increasing the size of the data set by adding to it the information gathered by incoming patients in a hospital or in a network of hospitals. Furthermore, the prediction of diabetes can also depend on other factors which are not present in said data set. Taking into consideration these factors will further help to improve the accuracy of the proposed system.

References

[1] A. Shaheen Khan and W. Ahmad, “Intelligent Decision Support System in Diabetic eHealth Care-From the perspective of Elders”, Blekinge Institute of Technology, **(2009)**, pp. 88.

[2] Z. Tawfik Saeed, M. V. Malakooti, Y. Ataeipoor and S. Talayeh Tabibi, “An expert system for diabetes diagnosis”, American Academic & Scholarly Research Journal, vol. 4, no. 5, **(2012)**, pp. 1.

[3] Margret Anouncia S, Clara Madonna L. J., Jeevitha P. and Nandhini R. T., “Design of a Diabetic Diagnosis System Using Rough Sets”, Cybernetics And Information Technologies, vol. 13, no. 3, **(2013)**, pp. 124-139.

[4] R. Ali, J. Hussain, M. Hameed Siddiqi, M. Hussain and S. Lee, “H2RM: A Hybrid Rough Set Reasoning Model for Prediction and Management of Diabetes Mellitus”, Sensors, vol. 15, no. 7, **(2015)**, pp. 15921-15951.

[5] A. Frank, “UCI machine learning repository”, http://archive. ics. uci. edu/ml, **(2010)**.

[6] Á. Celeste Ribeiro, A. Kardec Barros, E. Santana and J. Carlos Príncipe, “Diabetes classification using a redundancy reduction preprocessor”, Research on Biomedical Engineering, vol. 31, no. 2, **(2015)**, pp. 97-106.

[7] J. L. Chiang, M. Sue Kirkman, L. MB Laffel and A. L. Peters, “Type 1 diabetes through the life span: a position statement of the American Diabetes Association”, Diabetes care, vol. 37, no. 7, **(2014)**, pp. 2034-2054.

[8] Margret Anouncia S, Clara Madonna L. J., Jeevitha P. and Nandhini R. T., “Design of a Diabetic Diagnosis System Using Rough Sets”, Cybernetics and Information Technologies, vol. 13, no. 3, **(2013)**, pp. 124-139.

[9] R. Zolfaghari, “Diagnosis of Diabetes in Female Population of Pima Indian Heritage with Ensemble of BP Neural Network and SVM”, International Journal of Computational Engineering & Management, vol. 15, no. 4, **(2012)**, pp. 2230-7893.

[10] G. Kaur, “Improved J48 Classification Algorithm for the Prediction of Diabetes”, International Journal of Computer Applications, vol. 98, no. 22, **(2014)**.

[11] J. Lamb, “MATH 3220 Final Project”.

[12] M. Narasingarao, R. Manda, G. Sridhar, K. Madhu and A. Rao, “A clinical decision support system using multilayer perceptron neural network to assess well being in diabetes”, Journal Assoc. Phys. India, vol. 57, **(2009)**, pp. 127-133.

[13] M. Thirugnanam, P. Kumar, S. Vignesh Srivatsan and C. R. Nerlesh, “Improving the prediction rate of diabetes diagnosis using fuzzy, neural network, case based (FNC) approach”, Procedia Engineering, vol. 38, **(2012)**, pp. 1709-1718.

[14] H. Chen and C. Tan, “Prediction of type-2 diabetes based on several element levels in blood and chemometrics”, Biological trace element research, vol. 147, no. 1-3, **(2012)**, pp. 67-74.

[15] A. Sood, S. Diamond and S. Wang, “Type 2 Diabetes Mellitus Classification”, Department of Computer Science, Stanford University: Stanford, CA, USA, **(2012)**.

[16] J. Komorowski, L. Polkowski and A. Skowron, “Rough Set: A tutorial”.

[17] W. Jakub, “Genetic algorithms in decomposition and classification problems”, Rough Sets in Knowledge Discovery 2, Physica-Verlag HD, **(1998)**, pp. 471-487.

[18] S. Hoa Nguyen, A. Skowron and P. Synak, "Discovery of data patterns with applications to decomposition and classification problems", Rough Sets in Knowledge Discovery 2, Physica-Verlag HD, **(1998)**, pp. 55-97.

[19] U. Wybraniec-Skardowska, "On a generalization of approximation space", Bulletin of the Polish Academy of Sciences, Mathematics, vol. 37, no. 1-6, **(1989)**, pp. 51-62.

[20] P. Lingras and G. Peters, "Applying Rough Set Concepts To Clustering", Rough Sets: Selected Methods and Applications in Management and Engineering, **(2012)**, pp. 23-37.

Authors

Shantan Sawa, passed out student, of Bachelors of Technology in Computer Science and Engineering at VIT University, Vellore, Tamil Nadu, India. He has high enthusiasm and ambition towards the projects he works on. His academic interests include Artificial Intelligence (primarily Fuzzy Logic and Artificial Neural Networks) and Embedded Systems.

H. Bajaj currently working as an Associate Professor in Computer Science and Engineering Department at Sreenidhi Institute of Science and Technology, Hyderabad, Telangana India. His areas of research include Data Warehousing and Mining and Network Security.

Ronnie D. Caytiles, he had his Bachelor of Science in Computer Engineering- Western Institute of Technology, Iloilo City, Philippines, and Master of Science in Computer Science– Central Philippine University, Iloilo City, Philippines. He finished his Ph.D. in Multimedia Engineering, Hannam University, Daejeon, Korea. Currently, he serves as an Assistant Professor at Multimedia Engineering department, Hannam University, Daejeon, Korea. His research interests include Mobile Computing, Multimedia Communication, Information Technology Security, Ubiquitous Computing, Control and Automation

N. Ch. S. N. Iyengar (b 1961), he currently Professor at the Sreenidhi Institute of Science and Technology (SNIST) Yamnapet, Ghatkesr, Hyderabad, Telengana, India. His research interests include Agent-Based Distributed Computing, Intelligent Computing, Network Security, Secured Cloud Computing and Fluid Mechanics. He had 32+ years of experience in teaching and research, guided many scholars, has authored several textbooks and had nearly 200+ research publications in reputed peer reviewed international journals. He served as PCM/reviewer/keynote speaker/ Invited speaker.