ROUGH SET THEORY AND FUZZY LOGIC BASED WAREHOUSING OF HETEROGENEOUS CLINICAL DATABASES

R.SARAVANA KUMAR

Assistant professor
Department of Computer Science and Engineering
Jayam college of Engineering and Technology,
Dharmapuri, Tamilnadu
saravanakumar0681@gmail.com

G.THOLKAPPIA ARASU

principal
AVS Engineering College Salem,
TamilNadu.

Abstract: Large amounts of data about the patients with their medical conditions are presented in the Medical databases. Analyzing all these databases is one of the difficult tasks in the medical environment. In order to warehouse all these databases and to analyze the patient's condition, we need an efficient data mining technique. In this paper, an efficient data mining technique for warehousing clinical databases using Rough Set Theory (RST) and Fuzzy Logic is proposed. Our proposed methodology contains two phases – (i) Clustering and (ii) Classification. In the first phase, Rough Set Theory is used for clustering. Clustering is one of the data mining techniques for warehousing the heterogeneous data bases. Clustering technique is used to group data that have similar characteristics in the same cluster and also to group the data that have dissimilar characteristics with other clusters. After clustering the data, similar objects will be clustered in one cluster and the dissimilar objects will be clustered under another cluster. The RST can be reduced the complexity. Then in the second phase, these clusters are classified using Fuzzy Logic. Normally, Classification with Fuzzy Logic is generated more number of rules. Since the RST is utilized in our work, the classification using Fuzzy can be done with less amount of complexity. The proposed approach is implemented in MATLAB platform and evaluated using various clinical related databases from heart disease datasets – Cleveland, Switzerland and Hungarian. The performance analysis is based on Sensitivity, Specificity and Accuracy with different cluster numbers. The experimentation results show that our proposed methodology provides better accuracy result.

Keywords: Data mining, Warehousing, Rough Set Theory, Clustering, Fuzzy Logic, Classification.

1. Introduction

Clinical databases have accumulated large quantities of information about patients and their medical conditions [14]. In medical science, effectual tools are necessary to classify and systematically analyze giant amount of highly diverse medical records stored in heterogeneous databases. Also, there is an increasing demand for accessing those data. The volume, complexity and variety of databases used for data handling cause serious problems in manipulating this distributed information. Analyzing and warehousing all these heterogeneous medical data is a complex task in today's world. Data mining often

involves the analysis of data stored in a data warehouse [12]. Many data mining applications require partitioning of data into homogeneous clusters from which interesting groups may be discovered, such as a group of motor insurance policy holders with a high average claim cost, or a group of clients in a banking database showing a heavy investment in real estate [13]. Three of the major data mining techniques are regression, classification and clustering. The techniques in data mining are discovering new trends and patterns of behavior that previously went unnoticed [12].

Clustering analysis is an important research project in knowledge discovery and data mining (KDDM) [1]. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression [7]. In practical application, the data sets contain numerical and categorical (nominal) data in general. Accordingly, clustering algorithm is required to able to deal with both numerical data and categorical data [1]. Although classification is an effective means for distinguishing groups or classes of objects, it requires the often costly collection and labeling of a large set of training tuples or patterns, which the classifier uses to model each group [7]. A number of algorithms for clustering categorical data have been proposed such as K-Means, Expectation-Maximization (EM) Algorithm, Association Rule, K-Modes, K-Prototypes, CACTUS (Clustering Categorical Data Using Summaries), ROCK (Robust Clustering using Links), STIRR (Sieving Through Iterated Relational Reinforcement), LCBCDC (Link Clustering Based Categorical Data Clustering), fuzzy K-modes algorithm, fuzzy centroids algorithm etc., These algorithms require multiple runs to establish the stability needed to obtain a satisfactory value for one parameter. While these methods make important contributions to the issue of clustering categorical data, they are not designed to handle uncertainty in the clustering process. This is an important issue in many real world applications where there is often no sharp boundary between clusters [15].

There is a need for a robust clustering algorithm that can handle uncertainty in the process of clustering categorical data. This leads to another clustering algorithm named as Rough Set Theory (RST), which has received considerable attention in the computational intelligence literature [15]. Rough sets theory is a new mathematical tool to handle uncertainty and incomplete information. Polish mathematician Pawlak Z initially proposed it [3][4]. The theory consists of finite sets, equivalence relations and cardinality concepts [6]. A principal goal of rough set theoretic analysis is to synthesize or construct approximations (upper and lower) offsets concepts from the acquired data [5]. Rough set theory clarifies set-theoretic characteristics of the classes over combinatorial patterns of the attributes. This theory can be used to acquire some sets of attributes for classification and can also evaluate the degree of the attributes of database that are able to classify data [2]. Basically, when using rough set, the data itself is used to come up with the approximation in order to deal with the imprecision within. It can therefore be considered a self-sufficient discipline [6]. Unlike fuzzy set based

approaches, rough sets have no requirement on domain expertise to assign the fuzzy membership. Still, it may provide satisfactory results for rough clustering [15].

Fuzzy classification offers an alternative to crisp logic by evaluating data set based on their membership into each category [12]. Recently, fuzzy rule-based systems have often been applied to classification problems where non-fuzzy input vectors are to be assigned to one of a given set of classes. Many approaches have been proposed for generating and learning fuzzy IF–THEN rules from numerical data for classification problems. For example, fuzzy rule-based classification systems are created by simple heuristic procedures, neuro-fuzzy techniques, clustering methods, fuzzy nearest neighbor methods, and genetic algorithms [10]. Some heuristic criteria's are used for extraction of pre-specified number of fuzzy rule .Genetic algorithm (GA) based rule selection criteria improves classification ability of extracted fuzzy rule [9]. One of the interesting applications of GA's is in pattern classification problems in which the goal is to develop a classifier capable of dealing with different classes of a specific problem. Genetic algorithms have been used as rule generation and optimization tools in the design of fuzzy rule based systems. Genetic algorithms are search algorithms that use operations found in natural genetic to guide the journey through a search space [8]. The special term "Genetic Fuzzy System" (GFS) was coined by the community to refer to fuzzy systems that use a genetic algorithm to create or adjust one or more of their components. Specifically, the classifications of GFSs are (1) the genetic tuning of an existing knowledge base; (2) the genetic learning of components of the knowledge base [11].

The rest of the part is organized as follows: Section 2 discuss some of the previous work of data mining techniques that are used for warehousing various heterogeneous databases of medical field. Section 3 explains our proposed methodology for warehousing large amount of medical data by using two phases such as Clustering and Classification. The experimentation results and the performance evaluations for the proposed method are discussed in Section 4. By Section 5 our proposed work is summed up with the conclusion.

2. Related Works

Some of the brief reviews about warehousing large amount of heterogeneous databases using clustering algorithms are given below.

A fuzzy k-modes algorithm for clustering categorical data is very effective for identifying cluster structures from categorical data sets. However, the algorithm may stop at locally optimal solutions. In order to search for appropriate fuzzy membership matrices which can minimize the fuzzy objective function, G. Gan *et al.* [16] have presented the hybrid genetic fuzzy k-Modes algorithm for clustering categorical data sets. They have treated the fuzzy k-Modes clustering as an optimization problem and used GAs to solve the problem in order to obtain globally optimal solution. To speed up the convergence process of the algorithm, they have used the one-step fuzzy k-Modes algorithm in the crossover process instead of the traditional crossover operator. They have tested the algorithm using two real world data sets from UCI Machine Learning Repository (Blake & Merz, 1998) and the experimental results have shown that genetic fuzzy k-Modes is

very effective in identifying the inherent cluster structures in categorical data set if such structures exist.

Clustering categorical data is an integral part of data mining and has attracted much attention recently. Categorical data clustering technique has emerged as a new trend in technique of handling uncertainty in the clustering process. Tutut Herawan *et al.* [17] have focused their discussion on the rough set theory for categorical data clustering. They have proposed MADE (Maximal Attributes DEpendency), an alternative technique for categorical data clustering using rough set theory taking into account maximum attributes dependencies degree in categorical-valued information systems. They have proven that MADE technique is a generalization of MMR technique which is able to achieve lower computational complexity and higher clusters purity. Experimental results on two benchmark UCI datasets showed that MADE technique is better with the baseline categorical data clustering technique with respect to computational complexity and clusters purity. With this approach, they have believed that some applications through MADE will be applicable, such as for decision making, clustering very large datasets and etc.

Tutut Herawan [18] has presented the applications of rough set theory for clustering two cancer datasets. These datasets were taken from UCI ML repository. The proposed technique for selecting partitioning attribute was based on the maximum degree of dependency of attributes. To select a clustering attribute, the maximal degree of the rough attributes dependencies in categorical-valued information systems was used. It consists of four main steps. The first step deals with the computation of the equivalence classes of each attribute (feature). The second step deals with the determination of the dependency degree of attributes. The third step deals with selecting the maximum dependency degree. Finally, the attribute is ranked with the ascending sequence based on the maximum of dependency degree of each attribute. Further, he has used a divide-and-conquer method to partition/cluster the objects. The results showed that MDA technique can be used to cluster to the data. Further, he has presented clusters visualization using two dimensional plots. The plot results provided user friendly navigation to understand the cluster obtained. Moreover, he has succeeded in showing that the proposed technique is able to achieve lower computational complexity with higher purity as compared to the baseline method.

There are dozens of clustering algorithms that have been applied to gene expression data. But there is no single-best solution or a fit-all solution to clustering. The main goal in the analysis of large and heterogeneous gene expression datasets was to identify groups of genes that get expressed in a set of experimental conditions. J. Jeba Emilyn and K. Ramar [19] have proposed an intelligent clustering algorithm that is based on the frame work of rough sets. The main aim of their work was to develop a clustering algorithm that would successfully indentify gene patterns. The proposed novel clustering technique (RCGED) provided an efficient way of finding the hidden and unique gene expression patterns. It overcame the restriction of one object being placed in only one cluster. A more general rough fuzzy k means algorithm was implemented and experimented with different gene expression data sets. The proposed algorithm RCGED was also implemented and experimented with colon

cancer gene expression datasets. A comparison of the algorithms and their results were studied. The importance of upper and lower approximations of the rough clusters was optimized using DB index value. This algorithm seemed to prove better than the other rough set based clustering algorithms. The proposed algorithm was termed intelligent because it automatically determines the optimum number of clusters.

Data mining refers to the process of retrieving knowledge by discovering novel and relative patterns from large datasets. Clustering and Classification are two distinct phases in data mining that work to provide an established, proven structure from a voluminous collection of facts. A dominant area of modern-day research in the field of medical investigations includes disease prediction and malady categorization. Shomona Gracia Jacob, and R.Geetha Ramani [20] have focused to analyze clusters of patient records obtained via unsupervised clustering techniques and compare the performance of classification algorithms on the clinical data. Feature selection is a supervised method that attempts to select a subset of the predictor features based on the information gain. The Lymphography dataset comprises of 18 predictor attributes and 148 instances with the class label having four distinct values. Their paper highlighted the accuracy of eight clustering algorithms in detecting clusters of patient records and predictor attributes and highlighted the performance of sixteen classification algorithms on the Lymphography dataset that enables the classifier to accurately perform multi-class categorization of medical data. Their work asserted the fact that the Random Tree algorithm and the Quinlan's C4.5 algorithm give 100 percent classification accuracy with all the predictor features and also with the feature subset selected by the Fisher Filtering feature selection algorithm..

In many applications, data objects are described by both numeric and categorical features. Mixed data are ubiquitous in real world databases. Jinchao Ji et al. [21] have proposed a fuzzy c-mean type clustering algorithm to cluster these types of data. In their method, they have integrated the fuzzy centroid and mean to represent the prototype of a cluster, and used a new measure to evaluate the dissimilarity between data objects and the prototype of a cluster. In comparison with other algorithm, their algorithm has two main contributions: Firstly, by using the fuzzy centroid their algorithm could be preserved the uncertainty inherence in data sets for longer time before decisions are made, and was therefore less prone to falling into local optima in comparison with other clustering algorithms. Secondly, their algorithm took account of the significance of different attributes towards clustering by using the new measure to evaluate the dissimilarity between the data objects and the cluster's prototype. Because of these advantages their algorithm could achieved higher clustering accuracy, which has been demonstrated by experimental results. Then they have presented their algorithm for clustering mixed data. Finally, the performance of the proposed method was demonstrated by a series of experiments on four real world datasets in comparison with that of traditional clustering algorithms.

Xiaohui Yan et al. [22] have presented a Hybrid Artificial Bee Colony (HABC) algorithm, in which the crossover operator of GA was introduced in to improve the original ABC algorithm. With the new operator, information was exchanged fully between bees and the good individuals are utilized. In the early stage of the algorithm, the

searching ability of the algorithm was enhanced, and at the end of the algorithm, as the difference between individuals' decreases, the perturbation of crossover operator decreases and can maintain its convergence, too. To demonstrate the performance of the HABC algorithm, they have tested it on ten benchmark functions compared with ABC, CBAC, PSO, CPSO and GA algorithms. The results showed that the proposed HABC algorithm outperforms the canonical ABC and other compared algorithms on eight functions in terms of convergence accuracy and convergence speed. The test on rotated functions further proved that HABC is robust and can maintain its superiority on rotated functions while other algorithms are getting worse. According to its excellent optimization ability on numerical functions, they have applied HABC algorithm to the data clustering problem. Six well-known real datasets selected from the UCI machine learning repository were used for testing. Algorithms mentioned above as well as K-means algorithm were employed as comparison. The Results showed that HABC got the best total within-cluster variance value on five datasets, which prove that the HABC algorithm is a competitive approach for data clustering.

However, these algorithms will still trap in local minimum on a few functions, which can be seen both from the benchmark functions and data clustering. Finding the features of functions which HABC works not well on and improving the algorithm in solving these functions is important task, which is missing here.

3. Proposed Methodology During The Tenure Of The Research

In Today's Modern World, Clinical Fields Have Uncountable Data, Which Is Increasing Rapidly. With This Uncountable Data, Lots Of Problem Arises. Analyzing The Heterogeneous Databases Is A Complex Task. Thus In Our Previous Work, A Fast K-Modes Clustering Algorithm Was Used To Warehouse The Heterogeneous Databases. Modes That Used In K-Modes Algorithm Indicated The Values Of Attribute With High Frequency. According To The Frequency Of The Occurrence Of The Attribute Values, The Modes Were Used. A Dissimilarity Measure Was Used To Compare The Objects With The Modes And To Allocate Every Object To A Nearest Cluster. When Each Object Was Allocated To The Clusters, The Mode For Each Clusters Were Updated. After These Processes, All The Similar Objects Were Placed In One Cluster And The Dissimilar Objects Were Placed In Another Cluster. After The Clustering Process, The Classification Was Performed Using Fuzzy Logic. From This Method, The Related Medical Data Were Collected Effectively.

In That First Work, K-Modes Can Be Able To Handle Both The Numerical And Categorical Data And Also Cannot Be Able To Handle The Uncertainty. It Has Been Understood That The Boundary Between The Clusters Are Not Sharply. K-Modes Produce The Optimal Solutions, On The Basis Of Initial Modes And Also The Order Of The Objects In The Dataset. For Checking The Stability Of The Clustering Result Of Data, K-Modes Algorithm Must Be Run Multiple Times With Various Starting Values Of Modes. In Order To Overcome These Disadvantages Of The Previous Work, The Second Work Is Introduced. In Our Proposed Work, Rough Set Theory Is Used In The First Phase For Clustering. In This Process Also, After The Clustering Of Data, Similar

Objects Are Clustered In One Cluster And The Dissimilar Objects Are Clustered Under Another Cluster. The RST Can Be Reduced The Complexity. Then In The Second Phase, These Clusters Are Classified Using Fuzzy Logic. Normally, Classification With Fuzzy Logic Is Generated More Number Of Rules. Since The RST Is Utilized In Our Work, The Classification Using Fuzzy Can Be Done With Less Amount Of Complexity.

In This Work, There Are Two Phases Are Presented. These Two Phases Are Given As Follows.

Phase 1 - Clustering

Phase 2 - Classification

Phase 1 – Clustering Is Performed With The Help Of Rough Set Theory And The Phase 2 – Classification Is Worked Out With The Help Of Fuzzy Logic. The Structure For The Proposed Work Is Given In Fig. 1.

3.1. Phase 1 – Clustering

A partition of data into groups of similar categories or objects or is called as Clustering. Each of the groups with the categories or objects is called as clusters. Each of the categories in clusters is similar between them and is dissimilar to the categories of other groups. Some of the details from the data may be lose, because of the representation of data by fewer numbers of clusters; but can achieve simplification. The modeling of data makes the clustering in a historical perspective rooted in mathematics, numerical and statistics analysis. We can tell that the search for the clusters is unsupervised learning and the obtained system indicates a data concept by observing the machine learning perspective clusters related to hidden patterns. From this it is understand that the clustering is unsupervised learning of a hidden data concept. Data mining approach works with large databases that inflict on the analysis of clustering additional severe computational requirements.

According to the clustering approach, clusters are expressed in the following three different ways.

1. Identified clusters may be exclusive clusters; in this, any categories or objects belong to only one cluster.
2. Identified clusters may be overlapping; a category or an object may belong to many clusters.
3. Identified clusters may be probabilistic; a category or an object belongs to each cluster with a certain probability.

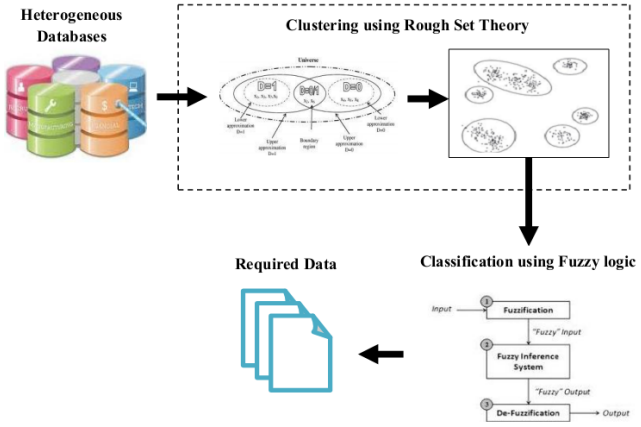

Fig. 1: Basic Structure for the proposed work

3.1.1. Using Rough Set Theory (RST) for Clustering

Rough Set Theory (RST) is a method for decision making in the presence of uncertainty. In RST, the general assumption is that the knowledge of human beings depends upon their capability to partitioning the objects such as classification of objects. RST classifies the information as imprecise, incomplete or uncertain that expressed in terms of the data that acquired from some experience. Each of the partitions or classifications of universe and the equivalence relations are notions that can be interchangeable one. So, the definition of Rough Sets depends on the equivalence relations for mathematical reasons.

3.1.2. Definitions and Notations in RST

Let $U(\neq \phi)$ be the finite set of all objects, called the Universe and A be the set of attributes. Let B be the non-empty subset of A and R be the equivalence relation over the set of all objects U. The concepts or categories of R can be defined with the notation U/R, by which the family of all the equivalence classes of R or the classification of U is represented. A category in R that contains an element $x \in U$ is denoted as $[x]_R$. From these notations and by a knowledge base, we can observe that the notation of a relation system is $K=(U,R)$, where K is the approximation space. In RST, a set of all the similar objects are defined as an elementary set. The elementary sets that are presented in K are the equivalence classes of K and the definable set in K is a finite union of the elementary sets in K.

For any subset $P(\neq \phi) \subseteq R$, the intersection of all the equivalence relations in the subset P is represented by $IND(P)$ and is known as Indiscernibility relation over P. In $IND(P)$, many equivalence classes are presented, these are called as P – basic Knowledge about U in K. For all $\phi \neq P \subseteq R$, the family of P – basic categories are called as the family of basic categories in Knowledge base K. The family of all equivalence relations that are defined in K is denoted by $IND(P)$. i.e. $IND(K)=\{IND(P) : \phi \neq P \subseteq R\}$. For any $Q \in R$, Q is known as Q–

elementary knowledge about U in K and the equivalence classes of Q are called as Q – elementary concepts of knowledge R.

For any $X \subseteq U$ and for an equivalence relation $R \in IND(K)$, $\underline{R}X = \cup \{Y \in U / R : Y \subseteq X\}$ and $\overline{R}X = \cup \{Y \in U / R : Y \cap X \neq \phi\}$ are the two subsets, which are called as $R - lower\ approximation\ of\ X$ and $R - upper\ approximation\ of\ X$, respectively. Upper approximation includes all the elements that possibly belong to the set, while the lower approximation includes only the elements that surely belong to the set. The notation of $R - boundary\ of\ X$ is denoted by $BN_R(X)$ and it is symbolically by $BN_R(X) = \overline{R}X - \underline{R}X$. The elements of $\underline{R}X$ are classified as the elements of X employing knowledge of R and also these are the elements of $\underline{U}$. In universe, we cannot decide the area of the borderline region. If and only if $\underline{R}X \neq \overline{R}X$ and $BN_R(X) \neq \phi$, then X is said to be rough with respect to R. And also if and only if $\underline{R}X = \overline{R}X$ or $BN_R(X) = \phi$, then X is said to be $\underline{R}X \neq \overline{R}X$ and $BN_R(X) \neq \phi$, then X is said to be rough with respect to $R - definable$ with respect to R. From this we can say that a set is said to be rough with respect to R, if and only if it is not $R - definable$.

Definition 1:- Indiscernibility Relation ($IND(B)$)

$IND(B)$ is a relation on U. Given two objects $x_i, x_j \in U$, these two objects are said to be indiscernible by the set of attributes B in A, if and only if $a(x_i) = a(x_j)$, $\forall a \in B$.

i.e. $x_i, x_j \in IND(B)$, if and only if $\forall\ a \in B$, where $B \subseteq A$ and $a(x_i) = a(x_j)$.

Definition 2:- Equivalence Class $\left([x_i]_{IND(B)} \right)$

Given $IND(B)$, the set of objects x_i having the same values for the set of attributes in B are said to be equivalence classes $[x_i]_{IND(B)}$, which also called as elementary set with respect to B.

Definition 3:- Lower Approximation

Given the set of attributes B in A and set of objects X in U, the lower approximation of X is defined as the union of all the elementary sets that presents in X.

i.e. $\underline{X_B} = U\, x_i \mid [x_i]_{IND(B)} \subseteq X$

Definition 4:- Upper Approximation

Given the set of attributes B in A and set of objects X in U, the lower approximation of X is defined as the union of all the elementary sets that have a non-empty intersection with X.

i.e. $\underline{X_B} = U \{x_i \mid [x_i]_{IND(B)} \cap X \neq \phi\}$

Definition 5:- Roughness

Roughness is defined as the ratio of the cardinality of the lower approximation and the cardinality of upper approximation.

$$\text{i.e. } R_B(X) = 1 - \frac{|X_B|}{|\overline{X_B}|}$$

If $R_B(X) = 0$, then we can say that X is precise with respect to B. If $R_B(X) < 1$, then X is rough with respect to B.

Definition 6:- Relative Roughness

Given $a_i \in A$, X is a subset of objects that have a specific value α of attribute a_i, $X_{a_j}(a_i = \alpha)$ and $\overline{X}_{a_j}(a_i = \alpha)$ indicates the lower and upper approximations of X with respect to $\{a_j\}$, then $R_{a_j}(X)$ is defined as the roughness of X with respect to $\{a_j\}$.

$$\text{i.e. } R_{a_j}(X / a_i = \alpha) = 1 - \frac{|X_{a_j}(a_i = \alpha)|}{|\overline{X}_{a_j}(a_i = \alpha)|}, \quad \text{where } a_i \in A \text{ and } a_i \neq a_j.$$

Definition 7:- Mean Roughness

Let A have n attributes and $a_i \in A$. X be the subset of objects that have a specific value α of the attribute a_i. Then $MeR(a_i = \alpha)$ defines the mean roughness for the equivalence class $a_i = \alpha$.

$$\text{i.e. } MeR(a_i = \alpha) = \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} R_{a_j}(X / a_i = \alpha) / (n-1) \right)$$

Definition 8:- Standard Deviation

Let A have n attributes and $a_i \in A$. X be the subset of objects that have a specific value α of the attribute a_i. Then $SD(a_i = \alpha)$ defines the standard deviation for the equivalence class $a_i = \alpha$.

$$\text{i.e. } SD(a_i = \alpha) = \sqrt{(1/(n-1)) \sum_{i=1}^{n-1} \left(R_{a_i}(X / a_i = \alpha) - MeR(a_i = \alpha) \right)^2}$$

Definition 9:- Distance of relevance

Given two objects B and C of categorical data with n attributes, then $DR(B,C)$ defines the distance of relevance of the two objects that is given below.

$$DR(B,C) = \sum_{i=1}^{n} (b_i, c_i)$$

where, b_i and c_i are the values of objects B and C respectively, under the i^{th} attribute a_i. We have

1. $DR(b_i, c_i) = 1, \quad if \ \ b_i \neq c_i$
2. $DR(b_i, c_i) = 0, \quad if \ \ b_i = c_i$
3. $DR(b_i, c_i) = \dfrac{|eq_{B_i} - eq_{C_i}|}{no_i} \quad if \ \ a_i \ is \ a \ numerical \ attribute$

where, eq_{B_i} - Number assigned to the equivalence class that having b_i.
eq_{C_i} - Number assigned to the equivalence class that having c_i.
no_i - Total number of equivalence classes in numerical attribute a_i.

3.1.3. Procedure for the Clustering of objects using RST

The whole data set is considered as the parent node U. According to the categories in the attributes, the whole data set is clustered. If the current number of cluster CNC is K, then our RST algorithm is applied for K times to get the desired cluster. Initially, the value of CNC is 1. So we can directly compute the roughness value, no need to calculate the average distance. In this, the relative roughness is calculated, in which the roughness of each attribute relative to the other attributes is computed. Each of the attributes a_i *and* a_j has N number of categories. At first, the first attribute a_1 $(i=1)$ is taken and also taken the first category α_1 of the first attribute a_1. Now the subset of objects X is obtained, which are having one specific value α_1 of attribute a_1. Then for the next attribute a_j $(j \neq i \ and \ j=2)$ with all the category values α, the equivalence classes are found. From this, the lower approximation $X_{a_j}(a_i=\alpha)$ and upper approximation $X_{a_j}(a_i=\alpha)$ are determined. The Roughness of a_i (when a_1 $(i=1)$ and α_1) is calculated using the formula,

$$R_{a_j}(X/a_i=\alpha) = 1 - \frac{|X_{a_j}(a_i=\alpha)|}{|X_{a_j}(a_i=\alpha)|}. \quad (1)$$

The roughness of all values of attributes are calculated for all the other attributes a_j $(j \neq i \ and \ j=2,3,4,...upto \ total \ number \ of \ attributes)$, by considering a_i (when a_1 $(i=1)$ and α_1) as constant one,. The Mean of all these values are found using

$$MeR(a_i=\alpha)=\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} R_{a_j}(X/a_i=\alpha)/(n-1)\right) \quad (2)$$

and then the Standard Deviation is calculated using the formula,

$$SD(a_i=\alpha)=\sqrt{(1/(n-1))\sum_{i=1}^{n-1}\left(R_{a_i}(X/a_i=\alpha)-MeR(a_i=\alpha)\right)^2} \quad (3)$$

This value is stored as a variable. For all the categorical value α of attribute a_1, the above process is done by keeping the same attribute a_1 $(i=1)$ as constant. All the Standard Deviation values are stored and the minimum value from these values is stored as another variable. For each of the attributes a_i $(i=2,3,4,...upto \ total \ number \ of \ attributes)$, the similar process is done every time and the smaller Standard Deviation value is used for the next computation.

Again the Standard Deviation is applied for these all smaller values to get the splitting attributes. If the value of Standard Deviation does not match with the smaller values then the nearest smaller value is taken as the splitting attribute and then the binary splitting is performed that splits the whole data set into two clusters. Now it is needed to

perform the above process on any one of the two clusters. In order to select the cluster, the Distance of Relevance formula

$$DR(B,C)=\sum_{i=1}^{n}(b_i,c_i) \tag{4}$$

is applied on the elements of these clusters. The Cluster which has larger Distance of Relevance is the input for the further calculation based on the above similar procedure. With this manner, our RST algorithm is continued, until the number of Clusters reach K.

RST Algorithm

Input *: Set of objects*
Output *: Clusters of objects*
Begin
Set Current Number of Cluster (CNC) = 1
Set ParentNode = U
Loop 1:
If $CNC<K$ *and* $CNC\neq 1$ *then*
ParentNode = Proc ParentNode (CNC)
End if
// Clustering the ParentNode
For each $a_i \in A$ *(i = 1 to n; where n is the number of attributes in A)*
Determine $[x_i]_{IND(a_i)}$ *(m = 1 to number of objects)*
For each $a_j \in A$ *(j = 1 to n; where n is the number of attributes in A,* $j\neq i$*)* *Determine* $Rough_{a_j}(a_i)$
Next
Find out $MeR(a_i=\alpha)=\left(\sum_{\substack{j=1\\ j\neq i}}^{n} R_{a_j}(X/\ a_i=\alpha)/(n-1)\right)$
Next
Calculate $SD(a_i=\alpha)=\sqrt{(1/(n-1))\sum_{i=1}^{n-1}\left(R_{a_i}(X/a_i=\alpha)-MeR(a_i=\alpha)\right)^2}$
Next
Set $SDR=SD\{\min\{SD(a_i=\alpha_1,...,SD(a_i=\alpha_{k_j})\}\}$*;(where* k_j *is the number of equivalence classes in* $Dom(a_i)$*). Determine splitting attribute* a_i *corresponding to the SDR*
Do binary split on the splitting attribute a_i
CNC = Number of leaf nodes
Go to loop 1:
End

Proc ParentNode (CNC)
Begin
Set i = 1
Do until i < CNC
If *Avg Dist of cluster i is calculated*
Go to label
Else
n = Count (set of elements in Cluster i)

$$Avg\ Dist(i) = \frac{2*\left(\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}(Distance\ of\ relevance\ between\ objects\ a_j\ and\ a_k)\right)}{(n*(n-1))}$$

End If
Label :
Increment i
Loop
Find Max (Avg Dist (i))
Return (Set of Elements in cluster i corresponding to Max (Avg Dist (i))
End

The output from the Rough Set Theory based clustering is two clusters, $C-1$ *and* $C-2$. Then these two clusters are classified using Fuzzy Logic.

3.2. Phase 2 – Classification

Each cluster from the clustering phase is classified in this second phase using a Fuzzy Logic. Fuzzy Inference is a method of generating a mapping from a given input to an output using fuzzy logic. Then, the mapping gives a basis, from which decisions can be generated or patterns discerned. Membership Functions, Logical Operations, and If-Then Rules are used in the Fuzzy Inference Process. The Stages of Fuzzy Inference Systems are,

1) Fuzzification
2) Fuzzy Rules Generation
3) Defuzzification

The Structure of the Fuzzy Inference System is given in the fig. 2. The three stages are also illustrated in the figure with the cluster as the input and the classification result as the output.

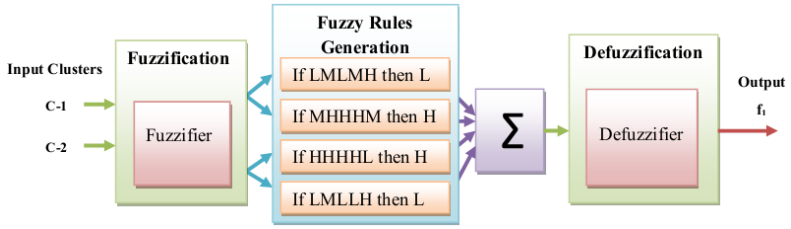

Fig 2: Structure of Fuzzy Inference System

3.2.1. Fuzzification

During the fuzzification process, the crusty quantities are converted into fuzzy. For the fuzzification process, the input is the two clusters, $C-1$ *and* $C-2$, that are the output of clustering algorithm using Rough Set Theory. After that, the minimum and maximum value of each cluster's are calculated from the input features. The process of fuzzification is computed by applying the following equations.

$$ML^{(C-1)} = \min + \left(\frac{\max - \min}{3} \right) \tag{5}$$

$$XL^{(C-1)} = ML + \left(\frac{\max - \min}{3} \right) \tag{6}$$

where, $ML^{(C-1)}$ - minimum limit values of the feature M .
$XL^{(C-1)}$ - maximum limit values of the feature M .

Use these equations (5) and (6), for calculating the minimum and maximum limit values for other cluster $C-2$ also. And also, three conditions are provided to generate the fuzzy values by using these equations.

Conditions

1. All the "Cluster 1 $(C-1)$" values are compared with "Minimum Limit Value ($ML^{(C-1)}$) ". If any values of Cluster 1 values are less than the value $ML^{(C-1)}$, then those values are set as L .
2. All the "Cluster 1 $(C-1)$" values are compared with "Maximum Limit Value ($XL^{(C-1)}$) ". If any values of Cluster 1 values are less than the value $XL^{(C-1)}$, then those values are set as H .
3. If any values of "Cluster 1 $(C-1)$" values are greater than the value $ML^{(C-1)}$, and less than the value $XL^{(C-1)}$, then those values are set as M .

Similarly, make the conditions for other cluster $C-2$ also for generating fuzzy values.

3.2.2. Fuzzy Rules Generation

According to the fuzzy values for each feature that are generated in the Fuzzification process, the Fuzzy Rules are also generated.

General form of Fuzzy Rule

"IF A THEN B"

The "IF" part of the Fuzzy Rule is called as "antecedent" and also the "THEN" part of the rule is called as "conclusion". The output values between L and H of the FIS is trained for generating the Fuzzy Rules.

3.2.3. Defuzzification

The input given for the Defuzzification process is the fuzzy set and the output obtained is a single number. As much as fuzziness supports the Rule Evaluation during the

intermediate steps and the final output for every variable is usually a single number. The single number output is a value L , M or H . This value of output f_1 , represents whether the given input dataset is in the Low range, Medium range or in the High range. The FIS is trained with the use of the Fuzzy Rules and the testing process is done with the help of datasets.

4. Results and Discussions

The experimental results obtained from the proposed methodology are given in this section. The proposed methodology is implemented using MATLAB. The data set description, clustering results and the performance analysis of our work are given in this section with the tables and graphical representations.

4.1. Dataset Description

In this section, the heart disease data sets – Cleveland, Hungarian and Switzerland are used for our work, which are taken from Data Mining Repository of the University of California, Irvine (UCI). The number of attributes used in these data sets are 76, but 14 of them only are used in general. 14 attributes are: Age, sex, chest pain type, resting blood pressure,serum cholesterol in mg/dl, fasting blood sugar, resting electro-cardiographic results, maximum heart rate achieved, exercise induced angina, ST depression, slope of the peak exercise ST segment, number of major vessels, thal and diagnosis of heart disease.

Cleveland dataset

Robert Detrano, M.D., Ph.D., collected these data at V.A. Medical Centre. All published experiments related to using a subset of 14 of the 76 attributes and 303 objects present in the processed Cleveland heart disease database. Specifically, ML researchers use only the Cleveland database till today. The existence of heart disease in the patient is indicated in the ‘‘goal’’ field by means of an integer that can take any value from 0 (no presence) to 4. Distinguishing disease existence (values 1–4) from non-existence (value 0) has been the focus of the experiments con-ducted in the Cleveland database. Six of the examples have been discarded because they had missing values. Class distributions are 54% heart disease absent, 46% heart disease present.

Hungarian data

Andras Janosi, M.D., collected these data at the Hungarian Institute of Cardiology, Budapest. Due to a huge percentage of missing values three of the attributes have been discarded but the format of the data is exactly the same as that of the Cleveland data. Thirty-four objects of the database were discarded on account of missing values and 261 objects were present. Class distributions are 62.5% heart disease absent and 37.5% heart disease present.

Switzerland data

William Steinbrunn, M.D., collected these data at the Univer-sity Hospital, Zurich, Switzerland. Switzerland data has more number of missing values. It contains 123 data objectss and 14 attributes. Class distributions are 6.5% heart disease absent and 93.5% heart disease present.

The following table 1 describes the sample description of dataset.

Table 1: A sample of Dataset Description

Age	Sex	chest pain type	resting blood pressure	serum cholesterol in mg/dl	fasting blood sugar	resting electro-cardiographic results	maximum heart rate achieved	exercise induced angina	ST depression	slope of the peak exercise ST segment	number of major vessels	thal	diagnosis of heart disease
63	male	1 (typ-angina)	145	233	t	Left-vent-hyper	150	no	2.3	down	0	Fixed-defect	<50
41	female	2 (atyp-angina)	130	201	f	Left-vent-hyper	172	no	1.4	up	2	normal	<50
37	Male	3 (non-anginal)	130	250	f	normal	187	no	3.5	down	1	normal	<50
57	female	4 (asympt)	120	354	f	normal	163	yes	0.6	up	0	normal	<50
48	male	0 (absense	112	230	f	normal	178	yes	2.5	flat	3	reversable_defect	>50_1

4.2. Evaluation metrics

An evaluation metric is used to evaluate the effectiveness of the proposed systems and to justify theoretical and practical developments of these systems. It consists of a set of measures that follow a common underlying evaluation methodology. Some of the metrics that we have chosen for our evaluation purpose are True Positive, True Negative, False Positive and False Negative, Specificity, Sensitivity, Accuracy.

Sensitivity

Sensitivity measures the proportion of actual positives which are correctly identified. It relates to the test's ability to identify positive results.

$$Sensitivity = \frac{Number\ of\ true\ positives}{Number\ of\ true\ positives + Number\ of\ false\ negatives} \quad (7)$$

Specificity

Specificity measures the proportion of negatives which are correctly identified. It relates to the ability of the test to identify negative results.

$$Specificity = \frac{Number\ of\ true\ negatives}{Number\ of\ true\ negatives + Number\ of\ false\ positives} \quad (8)$$

Accuracy

From the above results, we can easily get the accuracy value using the following formula,

$$Accuracy = 100 - \frac{[FP/(FP+TN)] / [FN/(FN+TP)]}{2} \quad (9)$$

Each person taking the test either has or does not have the disease. The test outcome can be positive (predicting that the person has the disease) or negative (predicting that the person does not have the disease).

$$True\ Positive\ (TP) = Unhealthy\ \ people\ \ correctly\ \ diagnosed\ \ as\ \ unhealthy$$
$$True\ Negative\ (TN) = Healthy\ \ people\ \ correctly\ \ identified\ \ as\ \ healthy.$$
$$False\ Positive\ (FP) = Healthy\ \ people\ \ incorrectl\ y\ \ identified\ \ as\ \ unhealthy$$
$$False\ Negative\ (FN) = Unhealthy\ \ people\ \ incorrectl\ y\ \ identified\ \ as\ \ healthy$$

4.3. Clustering Results

Using Rough Set theory, the clustering of whole objects is carried out in the clustering phase. As a result of this clustering phase, each similar object is clustered into one cluster. In this work, the datasets are clustered into two classes, since the type of heart disease values are taken into non-zero and zero valued. The non-zero values used here are 1,2,3,4, which show the presence and severity of heart diseases with its type. The indication of values are: Value 0 – absence of heart disease, Value 1–typical angina, Value 2 – atypical angina, Value 3 – non-anginal pain, Value 4 – asymptomatic. Thus our proposed work with the heart disease data set provides two clusters of the whole number of objects used. The clustering result is shown in the following fig. 3.

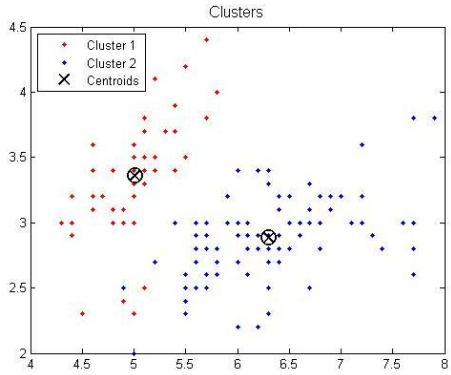

Fig. 3: Clustering result with two clusters

The red color dots in the above fig. 3 indicates the objects that are similar in cluster 1 and like that, the blue color dots indicates the similar objects that are presented in cluster 2. The cross mark on the red color dots and blue color dots, in the figure shows the centroids of the cluster 1 and cluster 2, respectively. For the heart disease datasets, totally 2 different classes are there, which makes the us to cluster the whole objects into two clusters using Rough Set Theory.

4.4. Performance Analysis of the proposed work with three datasets

Based on the evaluation metrics illustrated in the above section 4.2., the performance of the proposed system is analyzed. The evaluation metrics such as sensitivity, specificity

and the accuracy values are evaluated by our proposed system and the values are tabulated in the following tables. Three dataset values are used in this and each of which is tabulated in following tables. In table 2, the performance measure for sensitivity, specificity and accuracy of Cleveland dataset is tabulated.

Table 2: Performance evaluation for sensitivity, specificity and accuracy of Cleveland dataset

Iteration No	Sensitivity (in %)	Specificity (in %)	Accuracy (in %)
1	21	7	30
2	29	19	37
3	36	25	44
4	54	25	45
5	57	38	47
6	57	38	50
7	64	50	54
8	71	57	59
9	71	69	64
10	79	75	75

The graphical representation of the performance analysis of the proposed work with the Cleveland dataset is given in the following fig. 4.

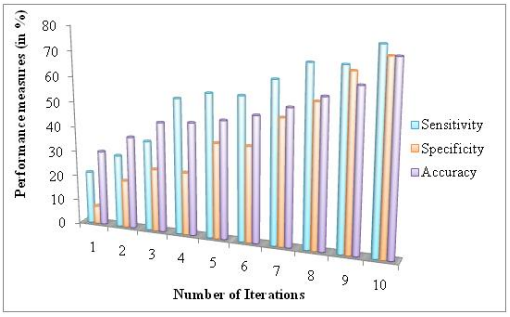

Fig. 4: Graph for the sensitivity, sensitivity and accuracy of Cleveland dataset

From the tabular values in table 2 and fig. 4, we can understand that each of the evaluation values sensitivity, specificity and accuracy are increased, when the iteration increased. In this work, totally 10 iterations are taken for the Cleveland dataset. These results show that the values varied with a slight increase in each of the iterations. From this, the accuracy result is high, when the iteration is high. From these result values of fig.4, we can understand that the proposed work is worked with better accuracy.

The performance evaluation for the Switzerland dataset is given in the following table 3.

Table 3: Performance evaluation for sensitivity, specificity and accuracy of Switzerland dataset

Iteration No	Sensitivity (in %)	Specificity (in %)	Accuracy (in %)

1	8	98	15
2	25	98	31
3	68	98	69
4	83	98	85
5	83	98	85
6	93	98	92
7	93	98	92
8	98	98	98

For the Switzerland dataset, the total number of iterations used for our proposed work is 8. But for the other two datasets, ten datasets are used. The reason for this is here we can obtain higher accuracy results in eighth iterations itself. This shows that our proposed work performs with good clustering and classification results. The following fig. 5 shows the graphical representation of the tabular values in table 3 for the Switzerland dataset.

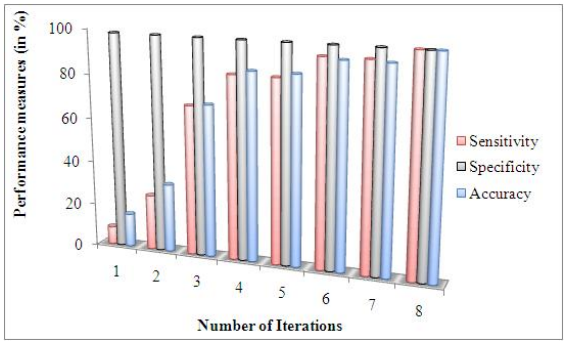

Fig. 5: Graph for the sensitivity, sensitivity and accuracy of Switzerland dataset

The observable points from the above table 3 and fig. 5 are the values of sensitivity, specificity and accuracy are increased, when the iteration is increased. For the Switzerland dataset, here 8 iterations are used. And also the variation for each of the iterations is highly varied for the first four iterations. After some increase in iteration only, the iteration gives higher value. From the result of Switzerland dataset evaluation, we can understand that the proposed work gives good clustering and classification of heterogeneous databases.

The Hungarian dataset is used as the third dataset in our work, which gives the following table 4 results for the performance evaluation of the proposed work.

Table 4: Performance evaluation for sensitivity, specificity and accuracy of Hungarian dataset

Iteration No	Sensitivity (in %)	Specificity (in %)	Accuracy (in %)
1	9	26	40
2	9	58	50
3	18	59	54
4	28	63	54
5	37	69	57
6	37	69	57

7	37	73	60
8	46	79	62
9	46	89	69
10	64	98	72

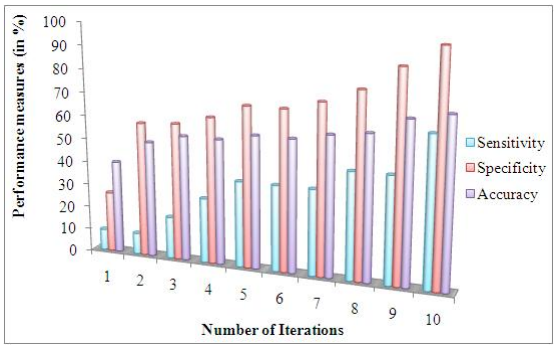

Fig. 6: Graph for the sensitivity, sensitivity and accuracy of Hungarian dataset

From the result of table 4 and fig. 6, we can observe that the sensitivity, specificity and accuracy values are varied with an increase for each of the iteration increases. Each of these evaluation metrics shows that the proposed work for the Hungarian dataset is worked with good accuracy result. Among the three datasets, Switzerland dataset is the best one for facilitating the higher clustering and classification result. Thus from the result of the three datasets used in our work, we can show that our proposed work is better one to cluster and classify the various databases in an effective manner. Even though, each of the dataset values is different in each of the iteration, the better result of our proposed work can be obtained from the higher iteration values.

5. Conclusion

Health information of every person related with the clinical field stores large amount of heterogeneous databases. Mining the required data for a query from these various databases is a difficult process. In order to warehouse the data, a proposed method was used in this paper. Initially, same objects were clustered into one cluster with a help of clustering algorithm, Rough Set Theory. After that, the clusters were classified using a Fuzzy logic, from which the required data could be extracted. The experimentation was carried out on heart disease datasets – Cleveland, Switzerland and Hungarian, using the MATLAB platform. The evaluation metrics of sensitivity, specificity and accuracy for the proposed work was also analyzed. From the results of the proposed work, the Switzerland dataset has provided better result, in compared with the other two datasets. However, the values increased in each iteration, while the iteration values increased. At the highest iteration level, we could achieved good clustering and classification results. The proposed method could be also able to deal with uncertainty problems. The boundary between the clusters was sharp to clearly cluster out the objects.

References

[1] Duo Chen, Du-Wu Cui, Chao-Xue Wang, and Zhu-Rong Wang, "A Rough Set-Based Hierarchical Clustering Algorithm for Categorical Data", *International Journal of Information Technology*, Vol.12, No.3, pp. 149-159, 2006.

[2] Show-Chin Lee and Mu-Jung Huang, "Applying AI technology and rough set theory for mining association rules to support crime management and fire-fighting resources allocation", *Journal of Information, Technology and Society*, Vol. 2, pp. 65-78, 2002.

[3] Licai Yang and Lancang Yang, "Study of a Cluster Algorithm Based on Rough Sets Theory", *In Proceedings of the sixth international conference on Intelligent Systems Design and Applications*, 2006.

[4] Tu Bao Ho, and Ngoc Binh Nguyen, "Nonhierarchical Document Clustering Based on a Tolerance Rough Set Model", *International Journal of Intelligent Systems*, Vol. 17, 199–212, 2002.

[5] Pabitra Mitra, Sankar K. Pal, and Md Aleemuddin Siddiqi, "Non-convex clustering using expectation maximization algorithm with rough set initialization", *ELSEVIER Pattern Recognition Letters*, Vol. 24, pp. 863-873, 2003.

[6] Chih-Cheng Hung, Hendri Purnawan, Bor-Chen Kuo, and Scott Letkeman, "Multispectal Image Classification Using Rough Set Theory and Particle Swarm Optimization", *Advances in Geoscience and Remote Sensing*, pp. 569-596, October 2009.

[7] A Priyadarishini, S Karthik, J Anuradha and B K Tripathy, "Diagnosis of Psychopathology using Clustering and Rule Extraction using Rough Set", *Advances in Applied Science Research*, Vol. 2, No. 3, pp. 346-362, 2011.

[8] Mohammad Saniee Abadeh, Jafar Habibi, and Emad Soroush, "Induction of Fuzzy Classification Systems Via Evolutionary Aco-Based Algorithms", *International Journal of Simulation Systems, Science and Technology*, Vol. 9, No. 3, September 2008.

[9] Dinesh P.Pitambare, Pravin M.Kamde, "Literature Survey on Genetic Algorithm Approach for Fuzzy Rule-Based System", *International Journal of Engineering Research*, Vol. 2, No.2, pp. 29-32, April 2013.

[10] Hisao Ishibuchi, and Tomoharu Nakashima, "Effect of Rule Weights in Fuzzy Rule-Based Classification Systems", *IEEE Transactions on Fuzzy Systems*, Vol. 9, No. 4, August 2001.

[11] Marcos Evandro Cintra, Maria Carolina Monard, Trevor P. Martin, and Heloisa de Arruda Camargo, "An Approach for the Extraction of Classification Rules from Fuzzy Formal Contexts", *Computer Science and Mathematics Institute Technical Reports*, pp. 1-28, 2011.

[12] A. Anushya, and A. Pethalakshmi, "A Comparative Study of Fuzzy Classifiers With Genetic On Heart Data", *International Conference on Advancement in Engineering Studies and Technology*, pp. 113-117, July, 2012.

[13] Zhexue Huang, "Clustering Large Data Sets With Mixed Numeric And Categorical Values", *In Proceedings of the first Pacific-Asia Conference on Knowledge Discovery and Data mining*, 1997.

[14] Jonathan C. Prather, David F. Lobach, Linda K. Goodwin, Joseph W. Hales, Marvin L. Hage, and W. Edward Hammond, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse", *In Proceedings of the AMIA Annual Fall Symposyum*, Vol. 101, No. 5, 1997.

[15] Darshit Parmar, Teresa Wu, and Jennifer Blackhurst, "MMR: An algorithm for clustering categorical data using Rough Set Theory", *ELSEVIER Data and Knowledge Engineering*, Vol. 63, pp. 879-893, 2007.

[16] G. Gan, J. Wu, and Z. Yang, "A genetic fuzzy k-Modes algorithm for clustering categorical data", *ELSEVIER Expert Systems with Applications*, Vol. 36, pp. 1615-1620, 2009.

[17] Tutut Herawan, Rozaida Ghazali, Iwan Tri Riyadi Yanto, and Mustafa Mat Deris, "Rough Set Approach for Categorical Data Clustering", *International Journal of Database Theory and Application*, Vol. 3, No. 1, March, 2010.

[18] Tutut Herawan, "Rough Clustering for Cancer Datasets", *International Journal of Modern Physics*, Vol. 1, No. 1, pp. 1-5, 2010.
[19] J. Jeba Emilyn and K. Ramar, "A Rough Set based Gene Expression Clustering Algorithm", *Journal of Computer Science*, Vol.7, No.7, pp. 986-990, 2011.
[20] Shomona Gracia Jacob, and R.Geetha Ramani, "Evolving Efficient Clustering and Classification Patterns in Lymphography Data Through Data Mining Techniques", *International Journal on Soft Computing (IJSC)* Vol.3, No.3, pp. 119-132, August 2012.
[21] Jinchao Ji, Wei Pang, Chunguang Zhou, Xiao Han, and Zhe Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data", *ELSEVIER Knowledge-Based Systems*, Vol. 30, pp. 129-135, 2012.
[22] Xiaohui Yan, Yunlong Zhu, Wenping Zou, and Liang Wang, "A new approach for data clustering using hybrid artificial bee colony algorithm", *ELSEVIER Neurocomputing*, Vol. 97, pp. 241-250, 2012.