I J A B E R, Vol. 14, No. 1, (2016): 439-455

CLUSTER & ROUGH SET THEORY BASED APPROACH TO FIND THE REASON FOR CUSTOMER CHURN

Mohammad Ahmar Khan[1], Mohammed Abdul Imran Khan[2]
Mohammed Aref[3] and Sarfaraz Fayaz Khan[4]

Abstract: *Data mining is the nontrivial process of extraction of interesting, implicit, potentially and previously unknown knowledge from large databases. There are many techniques used in data mining like: Statistical Analysis, Decision Tree, Neural Network, Clustering, Association Rule, Genetic Algorithms, Fuzzy Logic, and Rough Sets. Rough Set theory (RST), is a technique for dealing with uncertainty and for identifying cause-effect relationship in databases as a form of data mining and database learning. Customers become "churners" when they discontinue their subscription and move their business to a competitor. That is, churning is the process of customer turnover. This is a major concern for companies with many customers who can easily switch to other competitors. Examples include credit card issuers, insurance companies and telecommunication companies. This paper presents a mechanism to find the reason behind the churn. The proposed mechanism is based on the Clustering and Rough Set Theory. The objective of this chapter is to find the reason why a customer left the service provider. By knowing the reason behind the customer churn, service provider can take some preventive step to retain the customer.*

Keywords: *Knowledge, Database Learning, Customer Turnover, Clustering, Rough Set Theory*

1. INTRODUCTION

Service providers' focuses on two things first connect to the new customers and second retaining the old customers. In a saturated market connecting to new customer is somehow difficult but existing customers may be retained in order to increase the profit. The objective focused in this paper is to find the reason behind the churn. By knowing the reason behind the churn, service provider can take some preventive steps to retain the customer base. To find the churning reason Clustering Technique and Rough Set Theory approach has been used.

Clustering is a useful technique for the discovery of data distribution and patterns in the underlying data. The goal of clustering includes discovering both the dense and the sparse regions in the data set. The process of grouping a set of

[1] Asst. Professor, Dept of MIS, Dhofar University, Sultanate of Oman; *mkhan@du.edu.om*

[2] Asst. Professor, Dept of Accounting & Finance, Dhofar University, Sultanate of Oman; *imran@du.edu.om*

[3] Asst. Professor, Dept of MIS, Dhofar University, Sultanate of Oman; *mohammed_aref@du.edu.om*

[4] Asst. Professor, Dept of MIS, Dhofar University, Sultanate of Oman; *skhan@du.edu.om*

physical or abstract objects into groups based on similarity is called clustering. A cluster is a collection of data objects that are more similar to one another within the same cluster than to the objects in any of the other clusters. The areas that contribute to the advancements in research for clustering include data mining, statistics, machine learning, spatial database technology, biology and marketing.

2. LITERATURE REVIEW

2.1 Clustering - Various clustering approaches based on partitioning, hierarchical, grid-based and density-based. Clustering algorithms have several requirements as clustering has the potential to be applied in several fields. Some typical issues in this area are: scalability, ability to deal with different types of attributes, discovery of clusters of arbitrary shape, minimal domain knowledge to determine input parameters, ability to deal with noisy data, insensitivity to order of input records, high dimensionality, constraint based clustering and interpretability, and usability. The development and improvement of the main clustering techniques have been based either on the type of data or application domain. Therefore, of the several good clustering algorithms none display a satisfactory tradeoff between several criteria. All the approaches have some advantages and some disadvantages but none of them satisfies all the requirements.

Review of the Clustering Approaches - With the input *k,* the number of clusters, the partition-based algorithms partition the dataset *D* of *n* objects in a *d*-dimensional space into *k* groups such that the *cluster distribution* is minimized. The deviation of a point is computed differently in different algorithms and is more commonly called a *similarity function.* Partitioning algorithms use a two-step procedure. The first step involves identifying the k representative of clusters in such a way that the objective function is minimized. The second step pertains to assigning each object of the database to a cluster with its representative being *closest* to the considered object. The second step implies that a partition is equivalent to a Voronoi diagram and each cluster is contained in one of the Voronoi cells. Thus the shape of all clusters found by a partitioning algorithm is convex. This is a restrictive phenomenon that may not be suitable to a variety of real problems.

Hierarchical algorithms create a hierarchical decomposition of the database *D*. The hierarchical decomposition is represented by a dendrogram, a tree that iteratively splits *D* into smaller subsets until each subset consists of only one object. In such a hierarchy, each node of the tree represents a cluster of *D*. The dendrogram can either be created from the leaves up to the root or from the root down to the leaves by merging or dividing clusters respectively at each step. The process is terminated based on some condition related to the criteria for merge or divide. The algorithmic complexity of hierarchical algorithms are

O(n^2). The complexity of has been observed to be O(n) but it handles only numeric data. BIRCH is also order-sensitive. Being centric based approach BIRCH is observed not to perform well when the clusters do not have uniform size and shape while redistributing the data points in the final phase. CURE employs a combination of random sampling and partitioning to handle large databases. It identifies clusters having non-spherical shapes and wide variances in size by representing each cluster by multiple points. However, CURE is sensitive to some parameters. ROCK is a representative hierarchical clustering algorithm for categorical data. It introduces a novel concept called "link" in order to measure the similarity/proximity between a pair of data points. Thus, ROCK clustering method extends to non-metric similarity measures that are relevant to categorical data sets.

The grid-based clustering approach quantizes the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of the approach is its fast processing time, which is typically independent of the number of data objects. However, the processing time for this approach depends on the number of cells in each dimension in the quantized space.

Density-based approach produces clusters of arbitrary shape which are preferred for a number of application domains while the other approaches may not produce arbitrary shape clusters. Density-based clustering approach is founded on the idea that a cluster formation be continued on satisfying the density conditions, i.e. the density of the neighborhood of objects exceeds some threshold. This method is useful to filter out noise and outlier, and to discover clusters of arbitrary shape. Mathematically the notion of density is represented using two parameters namely, radius of neighborhood - Eps, and minimum number of points – MinPts. The values of these two parameters are user specified. These two parameters are global and are unable to reflect the local distribution of the data.

In most of the clustering algorithms all attributes are considered, to measure the similarity/dissimilarity between objects. In real life problem all attributes are not of equal importance. For example, consider the admission process for science group at graduate level. There may be various courses in science at graduate level viz. Physics, Chemistry, Mathematics etc. and each course may have different weights for the marks obtained in different papers in the qualifying examination. In, Devijver[1], weighted dissimilarity and other measures of dissimilarity have been proposed.

Parameters used in various clustering algorithms are either local or global. A good clustering must present the local as well as the global view. Therefore, the parameters used in the algorithms should be global as well as local. A good tradeoff between global and local view may be achieved by either of the two ways: let the parameters are local while the approach for cluster formation is global, or if the parameter be global then the approach for cluster formation is localized to the data in some perspective of the data.

Another major problem in the existing clustering algorithms is that the methods are applicable to datasets with either numerical or categorical attributes. As real world data may be represented by attributes of both the types clustering algorithms handling data with mixed type attribute may be desirable.

2.2 Rough Set Theory

Rough Set theory introduced by Pawlak[2] in early 1980's, is a technique for dealing with uncertainty and to identify cause-effect relationships in databases as a tool for data mining and database learning. It has also been used for improved information retrieval and for uncertainty management in relational databases. Some important concepts of RST are presented, defined and illustrated with the use of a representative Fruit dataset. RST is not limited to these concepts. However, only those which are used in this thesis are presented in this section.

Information System: A 3-tuple $S = (U, A, V_a)$ is called an *information system,* where U is a non-empty finite set of objects called the universe, A is a non-empty finite set of attributes, for $\forall\ a \in A,\ V_a$ is the value set of the attribute a.

Decision System: A *Decision system* is any information system of the form $S = (U, A \bigcup D, V_a)$, where $A \bigcap D = \phi$, D the set of decision attributes and A the set of conditional attributes. The fruit dataset as given in the table1.1 is an example of a decision system.

Following is the description of the fruit dataset as a decision system:

U= {O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8, O_9, O_{10}, O_{11}, O_{12}, O_{13}, O_{14}, O_{15}, O_{16}}

A= {Skin, Color, Size, Flesh}

D={Decision}

V_{Skin}= {hairy, smooth}

V_{Color}= {brown, green, red}

V_{Size}= {small, large}

V_{Flesh}= {soft, hard}

V_{Decision}={safe, danger}

Table 1
Fruit Data (Decision System)

Object	*Skin*	*Color*	*Size*	*Flesh*	*Decision*
O_1	Hairy	brown	large	Hard	Safe
O_2	Hairy	green	large	Hard	Safe
O_3	smooth	red	large	Soft	danger
O_4	Hairy	green	large	Soft	Safe
O_5	Hairy	red	small	Hard	Safe
O_6	smooth	red	small	Hard	Safe
O_7	smooth	brown	small	Hard	Safe
O_8	Hairy	green	small	Soft	danger
O_9	smooth	green	small	Hard	danger
O_{10}	Hairy	red	large	Hard	Safe
O_{11}	smooth	brown	large	Soft	Safe
O_{12}	smooth	green	small	Soft	danger
O_{13}	Hairy	red	small	Soft	Safe
O_{14}	smooth	red	large	Hard	danger
O_{15}	smooth	red	small	Hard	Safe
O_{16}	Hairy	green	small	Hard	danger

Indiscernibility Relation: The indiscernibility relation is at the core of rough set theory. All concepts of rough set theory are based on indiscernibility relation. Any two objects are said to be indiscernible if the vectors representing the two objects are identical i.e. the two tuples are identical. Two objects may be indiscernible with respect to $B \subseteq A$ if the attribute values of the attributes in B for the two objects are identical. Indiscernibility relation in on information system S denoted by $IND_S(B)$ for any $B \subseteq A$ is a relation on U defined as,

$$IND_S(B) = \{(y, y') \in U \times U \mid \forall a \in B, a(y) = a(y')\} \quad (1)$$

If $(y, y') \in IND_S(B)$, then objects y and y′ are indiscernible from each other with respect to all attributes in B then $IND_S(B)$ is called the *B-indiscernibility* relation. It is trivial to prove that $IND_S(B)$ for any $B \subseteq A$ satisfies the reflexivity, symmetricity and transitivity conditions. Therefore, using the equivalence relation $IND_S(B)$, the set of equivalence classes yields partition of the universe U denoted by $\frac{U}{IND_S(B)}$.

The equivalence class of $y \in U$ with respect to B-*indiscernibility* relation is denoted by $[y]_{B.}$ The relation $IND_S(B)$ when applied to the entire universe may also be indicated as IND(B).

Consider B= {Skin, Color, Size, Flesh}

B_1= {Skin, Color, Size}

B_2= {Skin, Color, Flesh}

B_3= {Skin, Color}

The indiscernibility relations corresponding to these sets of attributes are illustrated below. In the following examples of indiscernibility relation only the distinct pairs are exhibited while trivial cases, the pairs indicating reflexivity i.e. (O_i, O_i), are not included.

$\text{IND}_S(B) = \{(O_6,O_{15}), (O_{15},O_6)\}$

$\text{IND}_S(B_1) = \{(O_2,O_4), (O_3,O_{14}), (O_4,O_2), (O_5,O_{13}), (O_6,O_{15}), (O_8,O_{16}), (O_9,O_{12}), (O_{12},O_9), (O_{13},O_5), (O_{14},O_3), (O_{15},O_6), (O_{16},O_8)\ \}$

$\text{IND}_S(B_2) = \{(O_2,O_{16}), (O_4,O_8), (O_5,O_{10})\ (O_6,O_{14})\ , (O_6,O_{15}), (O_8,O_4), (O_{10},O_5), (O_{14},O_6), (O_{14},O_{15}), (O_{15},O_6), (O_{15},O_{14})\ (O_{16},O_2)\}$

$\text{IND}_S(B_3) = \{(O_2,O_4), (O_2,O_8), (O_2,O_{16}), (O_3,O_6), (O_3,O_{14}), (O_3,O_{15}), (O_4,O_2), (O_4,O_8), (O_4,O_{16}), (O_5,O_{10}), (O_5,O_{13}), (O_6,O_3), (O_6,O_{14}), (O_6,O_{15}), (O_7,O_{11}), (O_8,O_2), (O_8,O_4), (O_8,O_{16}), (O_9,O_{12}), (O_{10},O_5), (O_{10},O_{13}), (O_{11},O_7), (O_{12},O_9), (O_{13},O_5), (O_{13},O_{10}), (O_{14},O_3), (O_{14},O_6), (O_{14},O_{15}), (O_{15},O_3), (O_{15},O_6), (O_{15},O_{14}), (O_{16},O_2), (O_{16},O_4), (O_{16},O_8)\}$

Lower and Upper Approximation: Consider a concept $X \subseteq U$. The dataset U is described by the values of all the attributes in A. A description of X may also be possible based on the information of $B \subseteq A$. The lower and upper approximations of the concept X with respect to the B offer a formulation for such a description. The B-lower approximation and B-upper approximation of X are represented as $\underline{B}(X)$ and $\overline{B}(X)$ respectively and are defined by,

$$\underline{B}(X) = \{x : [x]_B \subseteq X\} \text{ and,}$$

$$\overline{B}(X) = \{x : [x]_B \cap X \neq \phi\} \tag{2}$$

The approximation regions $\underline{B}X$ and $\overline{B}X$ of the concept X are defined using the equivalence classes of the indiscernibility relation IND(B). The objects in $\underline{B}(X)$ with certainty are the members of X (certainly describe X) on the basis of the knowledge in B, while the objects in $\overline{B}(X)$ are possible members of X (possibly describe X) based on the knowledge in B.

Consider the decision system represented by fruit dataset and let X= {O_1, O_2, O_4, O_6, O_7, O_9, O_{11}, O_{14}} and let B= {Skin, color} then the equivalence classes of U = {O_1, O_2 , …, O_{16}} with respect to B are given by,

$[O_1]_B = \{O_1\}$

$[O_2]_B = \{O_2, O_4, O_8, O_{16}\}$

$[O_3]_B = \{O_3, O_6, O_{14}, O_{15}\}$

$[O_4]_B = \{O_2, O_4, O_8, O_{16}\}$

$[O_5]_B = \{O_5, O_{10}, O_{13}\}$

$[O_6]_B = \{O_3, O_6, O_{14}, O_{15}\}$

$[O_7]_B = \{O_7, O_{11}\}$

$[O_8]_B = \{O_2, O_4, O_8, O_{16}\}$

$[O_9]_B = \{O_{9,} O_{12}\}$

$[O_{10}]_B = \{O_5, O_{10}, O_{13}\}$

$[O_{11}]_B = \{O_7, O_{11}\}$

$[O_{12}]_B = \{O_{9,} O_{12}\}$

$[O_{13}]_B = \{O_5, O_{10}, O_{13}\}$

$[O_{14}]_B = \{O_3, O_6, O_{14}, O_{15}\}$

$[O_{15}]_B = \{O_3, O_6, O_{14}, O_{15}\}$

$[O_{16}]_B = \{O_2, O_4, O_8, O_{16}\}$

The partition of U with respect to the equivalence relation IND(B) for B = {Skin, color} is,

$U / IND(B) = \{\{O_1\}, \{O_2, O_4, O_8, O_{16}\}, \{O_3, O_6, O_{14}, O_{15}\}, \{O_5, O_{10}, O_{13}\}, \{O_7, O_{11}\}, \{O_{9,} O_{12}\}\}$

In the above example the equivalence classes which are certainly contained in the X are $[O_1]_B$, $[O_7]_B$ and $[O_{11}]_B$. Therefore, the lower approximation of X with respect to B is $\underline{B}X = \{O_1, O_7, O_{11}\}$.

For the objects O_1, O_2 ,O_3,O_4,O_6,O_7,O_8,O_9,O_{11},O_{12},O_{14},O_{15}, and O_{16}, it may be observed that

$[O_1]_B \cap X \neq \Phi,$

$[O_2]_B \cap X \neq \Phi,$

$[O_3]_B \cap X \neq \Phi,$

$[O_4]_B \cap X \neq \Phi,$

$[O_6]_B \cap X \neq \Phi,$

$[O_7]_B \cap X \neq \Phi$,

$[O_8]_B \cap X \neq \Phi$,

$[O_9]_B \cap X \neq \Phi$,

$[O_{11}]_B \cap X \neq \Phi$,

$[O_{12}]_B \cap X \neq \Phi$,

$[O_{14}]_B \cap X \neq \Phi$,

$[O_{15}]_B \cap X \neq \Phi$ and

$[O_{16}]_B \cap X \neq \Phi$

therefore, the upper approximation of X with respect to B is computed to be,

$$\overline{B}X = \{O_1, O_2, O_3, O_4, O_6, O_7, O_8, O_9, O_{11}, O_{12}, O_{14}, O_{15}, O_{16}\}$$

Properties of Lower and Upper Approximation:

1. $\underline{B}(X) \subseteq X \subseteq \overline{B}(X)$
2. $\underline{B}(\Phi) = \overline{B}(\Phi) = \Phi, \quad \underline{B}(U) = \overline{B}(U) = U$
3. $\overline{B}(X \bigcup Y) = \overline{B}(X) \bigcup \overline{B}(Y)$
4. $\underline{B}(X \bigcap Y) = \underline{B}(X) \bigcap \underline{B}(Y)$
5. $X \subseteq Y$ implies $\underline{B}(X) \subseteq \underline{B}(Y)$ and $\overline{B}(X) \subseteq \overline{B}(Y)$
6. $\underline{B}(X \bigcup Y) \supseteq \underline{B}(X) \bigcup \underline{B}(Y)$
7. $\overline{B}(X \bigcap Y) \subseteq \overline{B}(X) \bigcap \overline{B}(Y)$
8. $\underline{B}(\underline{B}(X)) = \overline{B}(\underline{B}(X)) = \underline{B}(X)$
9. $\overline{B}(\overline{B}(X)) = \underline{B}(\overline{B}(X)) = \overline{B}(X)$

Boundary Region: The set $BN_B(X) = \overline{B}X - \underline{B}X$ is called the *boundary region* of X, which consists of those objects whose membership to X is not decisive on the basis of the knowledge in B. The set $U - \overline{B}X$ is said to be the *B-outside region* of X. It consists of objects which are with certainty classified as not belonging to X on the basis of knowledge in B.

In the previous example, *X*= {O_1, O_2, O_4, O_6, O_7, O_9, O_{11}, O_{14}} and *B*= {Skin, color}. Since the lower and upper approximations of *X* with respect to B are,

$$\underline{B}(X) = \{O_1, O_7, O_{11}\}$$

$$\overline{B}(X) = \{O_1, O_2, O_3, O_4, O_6, O_7, O_8, O_9, O_{11} O_{12}, O_{14}, O_{15}, O_{16}\}$$

The boundary region may be obtained,

$$BN_B(X) = \{O_2, O_3, O_4, O_6, O_8, O_9, O_{12}, O_{14}, O_{15}, O_{16}\}$$

Rough Set: A set is said to be *rough* if the boundary region is non-empty and *crisp* otherwise.

In the above example since $BN_B(X) \neq \Phi$ therefore, the set X is a rough set.

Positive Region: Rough Set Theory offers tools to measure the degree of significance of attributes and the dependencies amongst them. For a given set of conditional attributes B, the *B-positive region* $POS_B(D)$ with respect to the relation IND(D) is defined as,

$$POS_B(D) = \bigcup\{\underline{B}X : X \in [x]_D\} \tag{3}$$

The positive region $POS_B(D)$ contains all the objects in U that can be classified without any error into distinct classes defined by IND(D), based only on information in B. Greater the cardinality of $POS_B(D)$ higher is the significance of the attributes in the set B with respect to D.

Rough Membership Function: The Rough membership function $\mu_X^B(y)$ is a tool to express how certainly an element *y* belongs to the concept X by the information about the element with respect to the set of attributes B. The Rough membership function is also used as a measure of significance of an attribute and is defined by,

$$\mu_X^B(y) = \frac{card\ (X \cap [y]_{IND\ (B)})}{card\ ([y]_{IND\ (B)})} \tag{4}$$

2.3 Reduct Computation - Reduct is one of the most important concepts in application of rough set theory in data mining. A reduct is the minimal set of attributes preserving classification accuracy of the original dataset. The problem to compute the reducts of a dataset is similar to the problem of feature selection. As per Pal[13, 14], all the reducts of a dataset are obtained by constructing a discernibility function from the dataset. It has been shown that the problems of finding minimal reduct and all reducts are NP-hard problems. Therefore, efficient methods to solve this NP-hard problem play an important role in the development of rough set-based data mining. Some efficient algorithms with heuristics, GA approach, etc. have also been proposed. Starzyk[8] has used strong equivalence to simplify discernibility function. However, this is still an open problem in rough set theory.

The conventional reduct computational algorithms fall into two categories: the reduction algorithms based on heuristic information and the reduction algorithms based on random strategies. Nevertheless, these algorithms do not guarantee to find a complete set of reducts for the dataset.

(i) Heuristic Algorithms

Johnson's[9] strategy is based on Johnson approximation algorithm for computing minimal prime implicant of any Boolean function in conjunctive normal form (CNF) formula. The main idea of the algorithm is to find an attribute discerning the largest number of pairs of objects, i.e., an attribute that occurs most in the entries of discernibility matrix. This algorithm proceeds until a reduct set is found. The time complexity of this algorithm is $O(|A|^2|U|^2)$ and the space complexity of this algorithm is $O(|A| \ |U|^2)$, where A is the set of attributes and U is database.

Jue Wang[7] has proposed an attributes reduction algorithm based on significance of attributes in discernibility matrix. In this algorithm significance of attributes is define as the attributes frequency in discernibility matrix. Hence algorithm regards the number of occurrences of each attribute as the significance of each attribute. The algorithm selects the attribute with the largest frequency, and deletes the elements involved with the selected attribute in discernibility matrix. Then the frequency of other attributes is computed. The algorithm continues to select and compute the frequency of remaining attributes until a reduct set is found. The time complexity of this algorithm is also $O(|A|^2 \ |U|^2)$ and the space complexity of this algorithm is also $O(|A| \ |U|^2)$.

By making use of attribute frequency information in discernibility matrix, Keyun Hu[6] has developed a feature ranking mechanism. Hu has proposed the algorithm using feature ranking as heuristics for reduct computation. The time complexity of this algorithm is $O((|A|+\log|U|) \ |U|^2)$ and the space complexity of this algorithm is $O(|A| \ |U|^2)$.

Keyun Hu[6] have proposed a new rough sets model and defined the core and reducts based on relational algebra using efficient set-oriented database operations. They presented two new algorithms to calculate core and reducts respectively, for feature selections. However, the time complexity of the algorithm is $O(|A|^2|U|)$ for the best case in spite of the hashing and indexing mechanism provided by the database systems.

(ii) Random Reduct Algorithms

Vinterbo[11] has formulated the rough set based attribute reduction as 'minimal hitting set' problem. He has defined an r-approximate hitting set as a set that intersects with at least a fraction r of given sets. Approximations of reducts from rough set theory are defined by means of minimal r-approximate hitting sets. In this method r-approximate hitting sets is computed using GA. The time complexity of the algorithm is $O(|A|^2 |U| \log |U|)$ and the space complexity is $O(|U|)$. Obviously, reducts obtained by this algorithm are not guaranteed to be complete.

Bazan[10] opines that the above methods do not take into account the fact that part of reduct set is chaotic i.e. it is not stable in randomly chosen samples of a given decision table. He introduced the notion of dynamic reduct. Dynamic reducts are in some sense the most stable reducts of the given decision table, i.e., they are the most frequently appearing reducts in subtables created by random sampling of a given decision table. Computation of reduct of variable size dynamically can be extremely computationally intensive, even for decision tables of moderately size. This algorithm is quite stable in most cases, yet it does not compute all reducts.

Quick Reduct algorithm by Chouchoulas[12], is an attempt to calculate a minimal reduct without exhaustively generating all possible subsets. Starting with an empty set the algorithm constructs the set P by adding the attributes with highest value of the attribute dependency $g_p(D)$, for D the decision attribute, until a maximum possible value is reached for the dataset (usually 1). Where

$$\gamma_p(D) = \frac{|POS_p(D)|}{|U|} \tag{5}$$

3. METHODOLOGY

The proposed methodology works in twofold: In first stage clusters are make using a Hierarchical clustering algorithm and then clustering analysis mechanism is applied to find the reason for the churn. The novelness of a clustering technique may be evaluated based on the analysis of its outcome in line with following questions:

a. Why does each of the clusters exist? and

b. Why is one cluster different from other clusters?

The answers to the above questions help to enhance the cluster definition, quality and specific distinctions of each cluster if any. A rough set theory based method using reduct has been explored to answer the following two questions,

a. Why does each of the clusters exist? and

b. What distinguishes one cluster from other clusters?

By the definition of the cluster, cluster is the collection of the similar objects. Therefore, the cluster_id could serve the purpose of the class labels for the data. By assuming the cluster_id as class label, decision relative reduct of rough set theory can be obtained from the dataset. A reduct is the minimal attribute set preserving classification accuracy of all the attributes of original dataset. Therefore, the reducts is identified as a useful tool to characterize the existence of a cluster. In the next subsections two methods have been discussed to discriminate between clusters and to characterize of clusters.

Knowing the differences between the clusters we can say because of these differences loyal customer and churn-able customer are differing and hence same will be reason for churn. After clustering we can which clusters are of loyal customer and which are not. And hence knowing the reason of difference we find the reason of churn.

3.1 Differentiating Clusters

Reduct is the minimal set of attributes which preserves the classification accuracy of the original data. This property of the reduct has been used to identify and describe the difference between clusters.

```
Input     : Two clusters $C_1$, $C_2$
Output    : The rule differentiating the clusters $C_1$, $C_2$
  Step 1. Assign the cluster_id as class label to the object of
          corresponding cluster.
  Step 2. Compute the reduct for the object of both clusters.
  Step 3. If the reduct exist then rule to differentiate $C_1$, $C_2$ is
          induced from the values of attributes in reduct.
  Step 4. If reduct does not exist then find the significant attributes
          having significance more than 0.5 and the rule to
          differentiating $C_1$, $C_2$ is the significant attribute(s) values
          having in these clusters.
  Step 5. End
```

Figure 1: Procedure for Difference between Clusters

An attribute based approach to determine the difference between any two clusters say C_1 and C_2 has been implemented with the use of the reduct of the objects belonging to the two clusters using the cluster_id of each object. In case a reduct exist then the discriminating values of the attributes in the reduct may specify the difference between a pair of clusters. In case of absence of a reduct the set of significant attributes obtained by other techniques may be used for this purpose. The set of attributes with significance higher than a threshold value θ, $0 < \theta \leq 1$, may be considered as significant attribute, a suitable value for θ is 0.5.

3.2 Existence of a Cluster

The use of reduct provides scope for an attribute based approach to characterize a cluster and to represent the basis of the existence of the cluster.

```
Input   : The set clusters $C_1, C_2, \ldots, C_n$
Output  : The reason of existence of clusters $C_1, C_2, \ldots, C_n$
  Step 1. Assign the cluster_id as class label to the object of
          corresponding cluster.
  Step 2. Compute the reduct for all the objects.
  Step 3. If the reduct exist then as reason for existence clusters $C_1$,
          $C_2, \ldots, C_n$ is the reduct and the values of attributes of
          reduct in the corresponding clusters.
  Step 4. If reduct does not exist then find the significant attributes
          having significance more than 0.5 and the Reason for
          existence clusters $C_1, C_2, \ldots, C_n$ is the significant attribute(s)
          and the values of significant attributes in the
          corresponding clusters.
  Step 5. End
```

Figure 2: Procedure for Existence of Clusters

The reduct for the given dataset considering cluster_id as the class data is to be computed. The reducts ensure a desired percentage of the classification accuracy. Therefore, the attribute values of the attribute in the reduct or the set of attributes with the significance above a certain threshold is proposed to be used to analysis the existence of the cluster.

4. RESEARCH FINDINGS

Proposed method of finding the reason behind churn has been applied on customer data with different attributes. Set of attributes can be categorized into different subset of attribute. The categorized set of attribute has been given in the table 2. In the initial research, the following churn-related factors were suggested :

- seniority (currently understood as the number of years the customer has been using the services),
- reaction to customer's complaints (the dominant reaction of the service provider to the complaints forwarded by the customer – acceptance or rejection),
- charge dynamics (*e.g.* the percentage relation of customer charges in the last monthly clearing period and the average charges in the previous three months),
- average charge (average monthly charge over the last 12 months),

Table 2
Attribute Detail

Subset of attributes	*Attributes*
Demographic data	Age sex Residence Area Population density of residence area Economy index of residence area
Loyalty data	Period of association Type of the last promotional offer (and the loyalty contract) Period of the last loyalty contract Months left before the end of the current loyalty contract Handing in the deactivation application Method by which Customer is Associated
Financial data	Monthly standing charges Changes in the monthly standing charges Monthly invoice amount Changes in the monthly invoice amounts
Visit data	Number of monthly visits Changes in the number of monthly visits Duration of monthly visits Changes in the duration of monthly visits
Additional Services	Catalog demand Number of monthly enquires Changes in the number of monthly enquires Number of monthly purchase Changes in the number of monthly purchase Number of monthly offers Changes in the number of offers
Market Competition	Any new offer in the market by other service provider Rate difference Services difference

- maximum charge (maximum monthly charge so far),
- customer's assessment of the provider's offer (in relation to other operators' offers),
- age,
- hometown and residential area,
- additional services (information on the use of extra services such as flayers, SMS *etc.*).

The proposed scheme for finding reason behind has been implemented in C language. To be able to differentiate clusters, pairs of clusters were considered for the experiment. The reducts of the data space was obtained in order to be able to differentiate the clusters using the cluster relative reducts. The significance of the attributes of the data space was also computed to help select one reduct from a set of reducts.

To find the cluster Attribute based Hierarchical Clustering Algorithm of Singh[15] has been used. This algorithm is capable in handling attributes of mixed type. As the data contained mixed type of attribute: nominal and continuous we have applied this algorithm. In the result there are many clusters are obtained for different set of parameters. For few result set, reduct has been found between pair of clusters. Attribute significance of some attribute are high in almost in all set of results. The significance of each attribute was computed by implementing the rough set theory concept $POS_p(Q)$, *p* being the attribute for which significance is measured and *Q* be the data space corresponding to the class.

Attributes having high significance are: Residence Area, Economy index of residence area, Period of association, Duration of monthly incoming calls, Number of monthly outgoing calls, Changes in the number of monthly outgoing multimedia messages, Number of monthly outgoing special services SMS, Any new offer in the market by other service provider, Call Rate difference, SMS rate difference etc. Moreover when we consider attributes in pair we higher significance of attribute pairs.

5. CONCLUSION AND FUTURE RECOMMENDATION

The finding demonstrates that a mechanism of finding the reason of churn among the customers. Proposed method based on clustering analysis and rough set theory to indicate the reason behind the churn. By knowing the reason behind the customer churn, service provider can take some preventive step to retain the customer.

The following churn-related factors were suggested :

- seniority (currently understood as the number of years the customer has been using the services),

- reaction to customer's complaints (the dominant reaction of the service provider to the complaints forwarded by the customer – acceptance or rejection),
- charge dynamics (*e.g.* the percentage relation of customer charges in the last monthly clearing period and the average charges in the previous three months),
- average charge (average monthly charge over the last 12 months),
- maximum charge (maximum monthly charge so far),
- customer's assessment of the provider's offer (in relation to other operators' offers),
- age,
- hometown and residential area,
- additional services (information on the use of extra services such as flayers, SMS *etc.*).

References

P. A. Devijver, and J. Kittler, eds. (1987), Pattern Recognition Theory and Applications. Berlin: Springer-Verlag.

Z. Pawlak (1982), "Rough sets." International Journal of Computer and Information Sciences 11, pp. 341-356.

Pawlak Z. (1998), Granularity of knowledge, indiscernibility and rough sets, Proceedings of 1998 IEEE International Conference on Fuzzy Systems, pp. 106-110.

S. K. Pal (1999), A. Skowron, Rough Fuzzy Hybridization- A new trend in decision making, Springer.

K. Hu, Y. Lu and C. Shi (2003), "Feature Ranking in Rough Sets", AI Communications, Special issue on Artificial intelligence advances in China, Volume 16 , Issue 1, pp. 41 – 50, May.

Hu, X., T. Y. Lin and J. Han (2004), A new rough set model based on database systems, Journal of Fundamental Informatics, vol. 59, pp. 135-152.

J. Wang and J. Wang (2001), "Reduction algorithms based on discernibility matrix: the ordered attributes method", Journal of Computer Science & Technology, vol.16, no.6, pp. 489-504.

J. Starzyk, D. E. Nelson, K. Sturtz (1998), "Reduct generation in information systems", Bulletin of international rough set society, volume 3, pp. 19-22.

D. S. Johnson (1974), "Approximation algorithms for combinatorial problems", Journal of Computer and System Sciences, pp. 256-278.

G. Bazan (1998), A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables, in Rough Sets in Knowledge Discovery 1: Methodology and Applications, Polkowski and Skowron (editors), Physica-Verlag, Heidelberg, Germany, Chapter 17, pp. 321-365.

S. Vinterbo and A. Ohrn (2000), "Minimal approximate hitting sets and rule templates", International Journal of Approximate Reasoning, vol. 25, no. 2, pp. 123-143.

A. Chouchoulas and Q. Shen (2001), Rough set-aided keyword reduction for text categorization. Applied Artificial Intelligence, Vol. 15, No. 9, pp. 843-873.

S. Mitra, S. K. Pal, and P. Mitra, "Data mining in soft computing framework: A survey", IEEE Transaction on Neural.

S. K. Pal (1999), A. Skowron, Rough Fuzzy Hybridization- A new trend in decision making, Springer.

Girish Kumar Singh, Sonajharia Minz (2007), "Attribute based Hierarchical Clustering Algorithm", National Conference on Trends in Advance Computing (NCTAC '07).