IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 3, Issue 5 (July-Aug. 2012), PP 44-51
www.iosrjournals.org

A Study in Employing Rough Set Based Approach for Clustering on Categorical Time-Evolving Data

H. Venkateswara Reddy[1], S. Viswanadha Raju[2]

[1]*(Department of CSE, Vardhaman College of Engineering, India)*
[2]*(Department of CSE, JNTUH College of Engineering, India)*

Abstract :-*The proportionate increase in the size of the data with increase in space implies that clustering a very large data set becomes difficult and is a time consuming process. Sampling is one important technique to scale down the size of dataset and to improve the efficiency of clustering. After sampling, allocating unlabeled data point into proper cluster is difficult in the categorical domain and in real situations data changes over time. However, clustering this type of data not only decreases the quality of clusters and also disregards the expectation of users, who usually require recent clustering results. In both the cases mentioned above, one is of allocating unlabeled data point into proper clusters after the sampling and the other is of finding clustering results when data changes over time which is difficult in the categorical domain. In this paper, using node importance technique, a rough set based method proposed to label unlabeled data point and to find the next clustering result based on the previous clustering result.*

Keywords:*-Categorical Data, Data Labeling, Node Importance, Rough Membership Function.*

I. INTRODUCTION

Extracting Knowledge from large amount of data is difficult which is known as data mining. Clustering refer to a method of finding a collection of similar objects from a given data set and objects in different collection are dissimilar. Most of the algorithms are developed for numerical data for clustering may be easy to use in normal conditions but not when it comes to categorical data [1, 3]. Clustering is a challenging issue in categorical domain, where the distance between data points is undefined [4]. It is also not easy to find out the class label of unknown data point in categorical domain. Sampling techniques accelerate the clustering [5, 6] and we consider the data points that are not sampled to allocate into proper clusters. The data which depends on time called time evolving data [7, 8]. For example, the buying preferences of customers may very with time, depending on the current day of the week, availability of alternatives, discounting rate etc [9]. Since data is modified and thus evolve with time, the underlying clusters may also change based on time by the data drifting concept [10, 11]. The clustering time-evolving data in the numerical domain [12, 13] has been explored in the previous literature though not in the categorical domain to the extent desired. It is a challenging problem in the categorical domain therefore to evolve a procedure for arriving at precise categorization.

As a result, a rough set [14] based method for performing clustering on the categorical time evolving data is proposed in this paper. This method find out if there is a drifting concept or not while processing the incoming data. In order to detect the drifting concept, the sliding window technique is adopted [15]. Sliding windows conveniently eliminate the outdated records while sifting through a mound of data. Therefore, employing the sliding window technique, we can test the latest data points in the current window to establish if the characteristics of clusters are similar to the last clustering result or not. This may be easy insofar as the numerical domain is concerned. However, in the categorical domain, the above procedure is challenging since the numerical characteristics of clusters are difficult to define.

Therefore, for capturing the characteristics of clusters, an effective cluster representative that summarizes the clustering result is required. In this paper, a mechanism called rough membership function-based similarity is developed to allocate each unclustered categorical data point into the corresponding proper cluster [16]. The allocating process goes by the name of data labeling and this method is also used to test the latest data point in the current sliding window depending on whether the characteristics of clusters are similar to the last clustering results or not[17, 19].

The paper is organized as follows. Section II supplies the relevant background to the study; section III deals with the basic definitions and data labeling for drifting concept detection, while the section IV, concludes the study with recommendations.

II. REVIEW OF RELATED LITERATURE

This section provides an exhaustive discussion of various clustering algorithms on categorical data along with cluster representatives and data labeling [10, 11, 16, 20]. Cluster representative is used to summarize and characterize the clustering result, which is not discussed in a detailed fashion in categorical domain unlike in the numerical domain [21, 23]. In K-modes algorithm [24], the most frequent attribute value in each attribute domain of a cluster represents what is known as a *mode* for that cluster. Finding modes may be simple, but the method of using only one attribute value in each attribute domain to represent a cluster is questionable.

ROCK clustering algorithm [25] is a form of agglomerative hierarchical clustering algorithm. It is based on links between data points, instead of distances between data points. The notion of links between data helps to overcome the problems with distance based coefficients. The link between point *i* (p_i) and point *j* (p_j), denoted as *link*(p_i,p_j), and is defined as the number of common neighbors between p_i and p_j. ROCK's hierarchical clustering algorithm accepts as input the set S of n sampled points as the representatives of those clusters, drawn randomly from the original data set to be clustered, and the number of desired clusters k. The procedure begins by computing the number of links between pairs of points. The number of links is then used in algorithm to cluster the data set. The first step in implementing the algorithm is to create a Boolean matrix with entries 1 and 0 based on adjacency matrix. The entry is 1 if the two corresponding points are adjacent neighbors or 0 if it is not. As this algorithm simply focuses on the adjacent of every data point, some data points may be left out or ignored; hence an algorithm based on entropy of the data points is assumed.

In the statistical categorical clustering algorithms [26, 28] such as COOLCAT [29] and LIMBO [30], data points are grouped based on the statistics. In algorithm COOLCAT, data points are separated in such a way that the expected entropy of the whole arrangements is minimized. In another algorithm LIMBO, the information bottleneck method is applied to minimize the information lost which resulting from summarizing data points into clusters. However, these algorithms perform clustering based on minimizing or maximizing the statistical objective function, and the clustering representatives in these algorithms are not clearly defined. Therefore, the summarization and characteristic information of the clustering results cannot be obtained by using these algorithms. A different approach is called for, which is the aim of the paper.

III. DRIFTING CONCEPT DETECTION

In this section discussed various notations with the problem definition and also Data labeling based on rough membership function.

III.1 Basic Notations

The problem of clustering the categorical time-evolving data is formulated as follows: if a series of categorical data points D is given, where each data point is a vector of q attribute values is $x_j=(x_j^1, x_j^2, \dots\ x_j^q)$. Let $A= (A_1, A_2 \dots A_q)$, where A_a is the ath categorical attribute, $1\leq a\leq q$. and N be window size. Divide the n data points into equal size windows call this subset as S^t, at time t. i.e. first N data points of D are located in the first subset S^1. The objective of our method is to find drifting data points between S^t and S^{t+1}.

Consider the following data set $D=\{x_1, x_2, \dots\dots\dots x_{30}\}$ of 30 data points shown below and if the sliding window size is 15 then S^1 contains first 15 data points and S^2 contains next 15 data points shown in table 1. We divide these first 15 data points into three clusters by using any clustering method shown in table 2. The data points that are clustered are called clustered data points or labeled data points and the remaining are called unlabeled data points. Our aim is to label the remaining 15 unlabeled data points which are belong to next sliding window S^2 and also to identify concept- drift occurred or not.

We define the following term which is considered in this method.

Node: A *Node* I_r is defined as attribute name + attributes value.

Basically a node is an attribute value, and two or more attribute values of different attributes may be identical, where those attribute domains intersection is non-empty, which is possible in real life. To avoid this ambiguity, we define node as not only with attribute value and but also with attribute name. For example, Nodes [height=60-69] and [weight=60-69] are different nodes even though the attribute values of attributes height and weight are same i.e.60-69. Because the attribute names height and weight are different, the nodes predictably are different.

Table 1: A data set D with 30 data points divided into two sliding windows S^1 and S^2.

S^1															
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
A_1	a	b	c	a	a	c	c	c	a	b	c	c	c	b	a
A_2	m	m	f	m	m	f	m	f	f	m	m	f	m	m	f
A_3	c	b	c	a	c	a	a	c	b	a	c	b	b	c	a
S^2															
	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	x_{30}
A_1	a	c	b	a	b	c	c	a	a	b	b	c	b	a	c
A_2	m	m	f	f	f	f	f	m	m	f	m	m	m	f	f
A_3	c	a	b	c	a	a	b	b	a	c	a	c	b	b	c

Table 2: Three clusters C_1, C_2 and C_3 after performing a clustering method on S^1.

C_1				
x_1	x_2	x_3	x_4	x_5
a	b	c	a	a
m	m	f	m	m
c	b	c	a	c

C_2				
x_6	x_7	x_8	x_9	x_{10}
c	c	c	a	b
f	m	f	f	m
a	a	c	b	a

C_3				
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
c	c	c	b	a
m	f	m	m	f
c	b	b	c	a

III.2 Data labeling

In this section, data labeling for unlabeled data points is done through rough membership function based on the similarity between existing cluster and this unlabeled data point.

To start with, some basic concepts are reviewed of rough set theory, such as *information system, the indiscernibility relation, rough membership function*. Then, a novel similarity between an unlabeled data point and a cluster is defined by considering the node importance values in a given cluster (i. e. the node frequency in a given cluster and the distribution of nodes in different clusters).

Generally, the set of data points to be clustered is stored in a table, where each row (tuple) represents the fact about an data point. We label this table as an information system. When all the attribute are categorical then this information system is termed categorical information system and is defined as a quadruple IS = (U, A, V, f), where U is the nonempty set of data points (or objects), called the universe, A is the nonempty set of attributes, V is the union of all attribute domains, i.e., V = $U_{a\in A}V_a$, where Va is the domain of attribute a and it is finite and unordered.

F: U XA→ V – a mapping called an information function such that for any $x \in$ U and $a \in$ A, f(x,a) $\in$ Va.

In 1982 Pawlak introduced rough set theory is a kind of symbolic machine learning technology for categorical information systems with uncertainty information [31, 32] and employing the notion of rough membership function the rough outlier factor for outlier detection has been defined by Jiang, Sui, and Cao in [33].

To understand the present method , we discuss the following definitions of rough set theory.

Definition 1: Let IS = (U, A, V, f) be a categorical information system, for any attribute subset P $\subseteq$ A, a binary relation IND(P), called indiscernibility relation, is defined as

$$IND(P)=\{(x,y)\in UXU \mid \forall a\in P,\ f(x,a)=f(y,a)\}$$

It is obvious that IND(P) is an equivalence relation on U and $IND(P)=\bigcap_{a\in P} IND(\{a\})$. Given P$\subseteq$A, the relation IND(P) induces a partition of U, denoted by U/ $IND(P)=\{[x]_P^U \mid x\in U\}$, where $[x]_P^U$ denotes the equivalence class determined by x with respect to P, i.e., $[x]_P^U=\{y\in U \mid (x,y)\in IND(P)\}$.

Definition 2: Let IS = (U, A, V, f) be a categorical information system, P $\subseteq$ A and X$\subseteq$U. The roughmembership function $\mu_{U,X}^P : U \to [0,1]$ is defined as $\mu_{U,X}^P(x)=\dfrac{\left|[x]_P^U \cap X\right|}{\left|[x]_P^U\right|}$

The rough membership function quantifies the degree of relative overlap between the set X and the equivalence class $[x]_P^U$ to which x belongs.

In classical set theory, an element either belongs to a set or it does not. The corresponding membership function is the characteristic function for the set, i.e., the function takes values 1 and 0, respectively. However, the rough membership function takes values between 0 and 1.

Definition 3: Let IS = (U, A, V, f) be a categorical information system, P $\subseteq$ A, U = S $\cup$ Q and S $\cap$ Q = ϕ. For any x $\in$ Q and X $\subseteq$ S, the rough membership function $\mu_{Q,X}^{P}: Q \to [0,1]$ is defined as

$$\mu_{Q,X}^{P}(x) = \{\frac{\left|[x]_P^S \cap X\right|}{\left|[x]_P^S\right|}, if\ [x]_P^S \neq \phi$$

= 0, otherwise,

Where $\left[x\right]_P^S = \{u \in S \mid \forall a \in P, f(u,a) = f(x,a)\}$

In Definition 3, th domain of the rough membership function is a subset Q of U, not the universe U. Moreover, we only consider the equivalence class of x on set S with respect to attribute set P.

Definition 4: Let IS = (U, A, V, f) be a categorical information system, P $\subseteq$ A, U = S $\cup$ Q and S $\cap$ Q = Φ. Suppose that a prior clustering result S = {c_1, c_2. . . c_k} is given, where c_i, 1≤ i ≤ k, is the i^{th} cluster. For any x∈Q, the similarity between an unlabeled object x and a cluster c_i with respect to P is defined as

$$S_P(x,c_i) = \sum_{a \in P} m_a * (1 - \frac{-1}{\log k}\sum_{j=1}^{k} w_a^{c_j} * \log w_a^{c_j})$$

Where $w_a^{c_j} = \frac{\left|[x]_a^S \cap c_j\right|}{\left|[x]_a^S\right|}$, $m_a = \frac{\left|\{u \mid f(u,a) = f(x,a), u \in c_i\}\right|}{|c_i|}$

And $|c_i|$ is the number of data points in the i^{th} cluster. m_a characterizes the importance of the attribute value f(x,a) in the cluster c_i with respect to attribute a. w_a^{cj} considers the distribution of attribute value f(x,a) between clusters [34]. Hence, $S_P(x, c_i)$ considers both the intra cluster similarity and the inter-cluster similarity.

Example 1: Let S = {c1, c2, c3}, where c_1, c_2 and c_3 are the clusters shown in table 2 and P=A= {A1, A2, A3}. According to the above Definition 4, it is clear that

$$S_P(x_{16},c_1) = \sum_{a \in P} m_a * (1 - \frac{-1}{\log 3}\sum_{j=1}^{3} w_a^{c_j} * \log w_a^{c_j})$$

$$= \frac{3}{5}(1 - \frac{-1}{\log 3}(\frac{3}{5}\log\frac{3}{5} + \frac{1}{5}\log\frac{1}{5} + \frac{1}{5}\log\frac{1}{5})) + \frac{4}{5}(1 - \frac{-1}{\log 3}(\frac{4}{9}\log\frac{4}{9} + \frac{2}{9}\log\frac{2}{9} + \frac{3}{9}\log\frac{3}{9})) +$$

$$\frac{3}{5}(1 - \frac{-1}{\log 3}(\frac{3}{6}\log\frac{3}{6} + \frac{1}{6}\log\frac{1}{6} + \frac{2}{6}\log\frac{2}{6})) = 0.1560$$

$$S_P(x_{16},c_2) = \sum_{a \in P} m_a * (1 - \frac{-1}{\log 3}\sum_{j=1}^{3} w_a^{c_j} * \log w_a^{c_j})$$

$$= \frac{1}{5}(1 - \frac{-1}{\log 3}(\frac{3}{5}\log\frac{3}{5} + \frac{1}{5}\log\frac{1}{5} + \frac{1}{5}\log\frac{1}{5})) + \frac{2}{5}(1 - \frac{-1}{\log 3}(\frac{4}{9}\log\frac{4}{9} + \frac{2}{9}\log\frac{2}{9} + \frac{3}{9}\log\frac{3}{9})) +$$

$$\frac{1}{5}(1 - \frac{-1}{\log 3}(\frac{3}{6}\log\frac{3}{6} + \frac{1}{6}\log\frac{1}{6} + \frac{2}{6}\log\frac{2}{6})) = 0.056$$

$$S_P(x_{16},c_3) = \sum_{a \in P} m_a * (1 - \frac{-1}{\log 3}\sum_{j=1}^{3} w_a^{c_j} * \log w_a^{c_j})$$

$$= + \frac{3}{5}(1 - \frac{-1}{\log 3}(\frac{4}{9}\log\frac{4}{9} + \frac{2}{9}\log\frac{2}{9} + \frac{3}{9}\log\frac{3}{9})) + \frac{2}{5}(1 - \frac{-1}{\log 3}(\frac{3}{6}\log\frac{3}{6} + \frac{1}{6}\log\frac{1}{6} + \frac{2}{6}\log\frac{2}{6})) = 0.079$$

Since the maximum similarity is found with cluster c_1. Therefore x_{16} can be allocated to the cluster c_1.

Table 3: Unlabeled data points similarities with clusters c_1, c_2 and c_3 and their class labels

Data point (x_i)	Cluster (c_j)	Similarity S_p (x_i, c_j)	Class Label
x_{16}	c_1	0.1560	C_1^*
	c_2	0.0560	
	c_3	0.0790	
x_{17}	c_1	0.0525	C_2^* or C_3^*
	c_2	0.0880	
	c_3	0.0880	
x_{18}	c_1	0.0265	C_2^*
	c_2	0.0583	
	c_3	0.0531	
x_{19}	c_1	0.1440	C_1^*
	c_2	0.1222	
	c_3	0.0905	
x_{20}	c_1	0.02270	C_2^*
	c_2	0.06016	
	c_3	0.04540	
x_{21}	c_1	0.04091	C_2^*
	c_2	0.11487	
	c_3	0.09114	
x_{22}	c_1	0.04377	C_2^*
	c_2	0.12060	
	c_3	0.09401	
x_{23}	c_1	0.12224	C_1^*
	c_2	0.05860	
	c_3	0.06236	
x_{24}	c_1	0.11671	C_1^*
	c_2	0.06361	
	c_3	0.04879	
x_{25}	c_1	0.05904	C_1^* or C_2^* or C_3^*
	c_2	0.05904	
	c_3	0.05904	
x_{26}	c_1	0.03578	C_2^*
	c_2	0.03676	
	c_3	0.02186	
x_{27}	c_1	0.08930	C_1^*
	c_2	0.08717	
	c_3	0.06062	
x_{28}	c_1	0.03856	C_1^*
	c_2	0.03160	
	c_3	0.03536	
x_{29}	c_1	0.10760	C_1^*
	c_2	0.08535	
	c_3	0.06947	
x_{30}	c_1	0.07733	C_2^*
	c_2	0.11392	
	c_3	0.11281	

Similarly, we can find similarity between the remaining unlabeled data point of sliding window S^2 with all the clusters of first clustering result C^1. The table 3 shows all those similarities and the class label of those unlabeled data points obtained with maximum similarity. And the graph in fig.1 shows unlabeled data points of sliding window S^2 similarities with clusters c_1, c_2 and c_3.

Obviously, x_{16}, x_{19}, x_{23}, x_{24}, x_{27}, x_{28} and x_{29} are allocated to the cluster c_1. x_{18}, x_{20}, x_{21}, x_{22}, x_{26} and x_{30} are labeled to the cluster c_2. But x_{17} can be allocated to either c_2 or c_3 and x_{25} can be allocated to any of the three clusters c_1, c_2 and c_3. The obtained temporal clustering result from sliding window S^2 is shown in table 4.

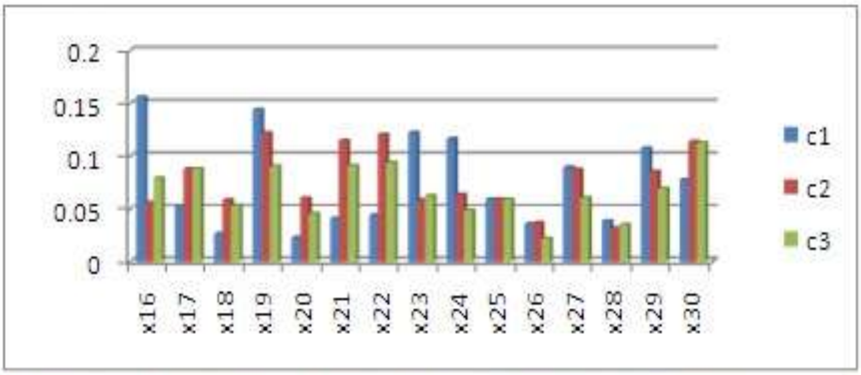

Fig1. Unlabeled data points of sliding window S^2 similarities with clusters c_1, c_2 and c_3

Table 4: The temporal clustering result with three clusters c_1, c_2 and $c_3$based on sliding window S^2.

c_1							c_2						c_2 c_3	c_1 c_2 c_3
x_{16}	x_{19}	x_{23}	x_{24}	x_{27}	x_{28}	x_{29}	x_{18}	x_{20}	x_{21}	x_{22}	x_{26}	x_{30}	x_{17}	x_{25}
a	a	a	a	c	b	a	b	b	c	c	b	c	c	b
m	f	m	m	m	m	f	f	f	f	f	m	f	m	f
c	c	b	a	c	b	b	b	a	a	b	a	c	a	c

Now we compare previous clustering result with current clustering result to decide the concept drift is occurred or not by using the equation (1).

$$\text{Concept drift} = \left\{ \begin{array}{l} \text{Yes, if } \dfrac{\sum_{i=1}^{k^{[t_e,t-1]}} d\left(c_i^{[t_e,t-1]}, c_i^t\right)}{k^{[t_e,t-1]}} > \eta, \\ \text{where } d\left(c_i^{[t_e,t-1]}, c_i^t\right) \\ = \left\{ \begin{array}{ll} 1, & \text{if } \left| \dfrac{m_i^{[t_e,t-1]}}{\sum_{x=1}^{k^{[t_e,t-1]}} m_x^{[t_e,t-1]}} - \dfrac{m_i^t}{\sum_{x=1}^{k^{[t_e,t-1]}} m_x^t} \right| > \epsilon \\ 0, & \text{otherwise} \end{array} \right\} \\ \text{No, otherwise} \end{array} \right.$$

---------------------(1)

Here we use two thresholds one η, a cluster difference threshold and other one is ε, a cluster variation threshold. Suppose the cluster difference threshold is set 0.25 and the cluster variation threshold set 0.3.

In our example, the variation of the ratio of the data points between clusters c_1 of S^1 and c_1 of S^2 is $|5/15-7/15|=0.133< \varepsilon$. The variation of the ratio of the data points between clusters c_2 of S^1 and c_2 of S^2 is $|5/15-6/15|=0.06< \varepsilon$ and the variation of the ratio of the data points between clusters c_3 of S^1 and c_3 of S^2 is $|5/15-0/15|=0.333< \varepsilon$. Therefore by equation (1) concept drift is yes because $(0+0+1)/3=0.33> \eta=0.25$.

III. 3 Presents data labeling algorithm based on rough membership function.

Algorithm 1:

Input: IS = (S ∪ Q, A, V, f), where S is a sampled data, Q is an unlabeled data set, k is the number of clusters;

Output: Each object of Q is labeled to the cluster that obtained the maximum similarity.

Method: 1. Generate a partition S = {c_1, c_2, . . . ,c_k} of S with respect to A by calling the corresponding categorical clustering algorithm;

```
        2. For i = 1 to |Q|
        3. For j = 1 to k
```

```
4. Calculate the similarity between the i^th object and the j^th cluster according to Definition 4, and the i^th
   object is labeled to the cluster that obtained the maximum similarity;
5. For end
6. For end
```

The runtime complexity of the above presented algorithm depends on the complexity of clustering algorithm which is used for initial clustering. Generally any clustering algorithm takes O(knq) runtime complexity, where k is the number of clusters, n is the total number of data points to cluster and q is the number of attributes. The runtime complexity for computing the similarity between arbitrary unlabeled data point and a cluster is O (|S||P|). Therefore, the whole computational cost of the proposed algorithm is O (S||P||Q||k|).

IV. CONCLUSION

In categorical domain, the problem of how to allocate the unlabeled data points into appropriate clusters has not been fully explored in the previous works. Besides, for the data which changes over time, clustering this type of data not only decreases the quality of clusters and also disregards the expectations of users, when usually require recent clustering results. In this paper, based on the rough membership function and the frequency of the node, a new similarity measure for allocating the unlabeled data point into appropriate cluster has been defined. Since this similarity measure has two characteristics, one pertaining to the distribution of the node in the different cluster and second is of the probability of the node in given cluster, which consider both the intra-cluster similarity and the inter- cluster similarity, the algorithm for remedying unaddressed problem has been presented for data labeling while making allowance for time complexity.

V. ACKNOWLEDGEMENTS

We wish to thank our supervisor Dr.Vinay Babu, JNTUniversity Hyderabad, India and all our friends who contributed to the paper.

REFERENCES

[1]. Anil K. Jain and Richard C. Dubes. “Algorithms for Clustering Data”, Prentice-Hall International, 1988.

[2]. Jain A K MN Murthy and P J Flyn, “Data Clustering: A Review,” *ACM Computing Survey,* 1999.

[3]. Kaufman L, P. Rousseuw,” Finding Groups in Data- An Introduction to Cluster Analysis”, Wiley Series in Probability and Math. Sciences, 1990.

[4]. Han,J. and Kamber,M. “Data Mining Concepts and Techniques”, Morgan Kaufmann, 2001.

[5]. Bradley,P.S., Usama Fayyad, and Cory Reina,” Scaling clustering algorithms to large databases”, Fourth International Conference on Knowledge Discovery and Data Mining, 1998.

[6]. Joydeep Ghosh. Scalable clustering methods for data mining. In Nong Ye, editor, “Handbook of Data Mining”, chapter 10, pp. 247–277. Lawrence Ealbaum Assoc, 2003.

[7]. Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani, and Liadan O’Callaghan,” Clustering data streams: Theory and practice”, IEEE Transactions on Knowledge and Data Engineering, PP.515–528, 2003.

[8]. Gibson, D., Kleinberg, J.M. and Raghavan,P. “Clustering Categorical Data An Approach Based on Dynamical Systems”, VLDB pp. 3-4, pp. 222-236, 2000.

[9]. Michael R. Anderberg,” Cluster analysis for applications”, Academic Press, 1973.

[10]. Chen H.L, M.-S. Chen, and S-U Chen Lin “Frame work for clustering Concept –Drifting categorical data”, *IEEE Transaction Knowledge and Data Engineering v21 no 5* , 2009.

[11]. Klinkenberg, R.,” Using labeled and unlabeled data to learn drifting concepts”, IJCAI-01Workshop on Learning from Temporal and Spatial Data, pp. 16-24, 2001.

[12]. Aggarwal, C., Han, J., Wang, J. and Yu P, “A Framework for Clustering Evolving Data Streams”, Very Large Data Bases (VLDB), 2003.

[13]. Aggarwal, C., Wolf, J.L. , Yu, P.S. , Procopiuc, C. and Park, J.S. “Fast Algorithms for Projected Clustering,”, ACM SIGMOD ’99, pp. 61-72, 1999.

[14]. Liang, J. Y., Wang, J. H., & Qian, Y. H. (2009). A new measure of uncertainty based on knowledge granulation for rough sets. Information Sciences, 179(4), 458–470.

[15]. Shannon, C.E, “A Mathematical Theory of Communication,” Bell System Technical J., 1948.

[16]. Chen, H.L., Chuang, K.T. and Chen, M.S. “Labeling Un clustered Categorical Data into Clusters Based on the Important Attribute Values”, IEEE International Conference. Data Mining (ICDM), 2005.

[17]. Venkateswara Reddy.H, Viswanadha Raju.S,” Our-NIR: Node Importance Representative for Clustering of Categorical Data”, International Journal of Computer Science and Technology , pp. 80-82,2011.

[18]. Venkateswara Reddy.H, Viswanadha Raju.S,” POur-NIR: Modified Node Importance Representative for Clustering of Categorical Data”, International Journal of Computer Science and Information Security, pp.146-150, 2011.

[19]. Venkateswara Reddy.H, Viswanadha Raju.S,” A Threshold for clustering Concept – Drifting Categorical Data”, IEEE Computer Society, ICMLC 2011.

[20]. Venkateswara Reddy.H, Viswanadha Raju.S, “Clustering of Concept Drift Categorical Data Using Our-NIR Method”, International Journal of Computer and Electrical Engineering, Vol. 3, No. 6, December 2011.

[21]. Tian Zhang, Raghu Ramakrishnan, and Miron Livny,” BIRCH: An Efficient Data Clustering Method for Very Large Databases”,ACM SIGMOD International Conference on Management of Data,1996.

[22]. S. Guha, R. Rastogi, K. Shim. CURE,” An Efficient Clustering Algorithm for Large Databases”, ACM SIGMOD International Conference on Management of Data, pp.73-84, 1998.

[23]. Ng, R.T. Jiawei Han “CLARANS: a method for clustering objects for spatial data mining”, Knowledge and Data Engineering, IEEE Transactions, 2002.

[24]. Huang, Z. and Ng, M.K, “A Fuzzy k-Modes Algorithm for Clustering Categorical Data” IEEE On Fuzzy Systems, 1999.

[25]. Guha,S., Rastogi,R. and Shim, K, "ROCK: A Robust Clustering Algorithm for Categorical Attributes", International Conference On Data Eng. (ICDE), 1999.

[26]. Vapnik, V.N," The nature of statistical learning theory", Springer,1995.

[27]. Fredrik Farnstrom, James Lewis, and Charles Elkan," Scalability for clustering algorithms revisited", ACM SIGKDD pp.:51–57, 2000.

[28]. Ganti, V., Gehrke, J. and Ramakrishnan, R, "CACTUS—Clustering Categorical Data Using Summaries," ACM SIGKDD, 1999.

[29]. Barbara, D., Li, Y. and Couto, J. "Coolcat: An Entropy-Based Algorithm for Categorical Clustering", ACM International Conf. Information and Knowledge Management (CIKM), 2002.

[30]. Andritsos, P, Tsaparas, P, Miller R.J and Sevcik, K.C."Limbo: Scalable Clustering of Categorical Data", Extending Database Technology (EDBT), 2004.

[31]. Liang, J. Y., & Li, D. Y. (2005). Uncertainty and knowledge acquisition in information systems. Beijing, China: Science Press.

[32]. Liang, J. Y., Wang, J. H., & Qian, Y. H. (2009). A new measure of uncertainty based on knowledge granulation for rough sets. Information Sciences, 179(4), 458–470.

[33]. Jiang, F., Sui, Y. F., & Cao, C. G. (2008). A rough set approach to outlier detection. International Journal of General Systems, 37(5), 519–536.

[34]. Gluck, M.A. and Corter, J.E. "Information Uncertainty and the Utility of Categories", Cognitive Science Society, pp. 283-287, 1985.