Data Labeling method for genome DNA data based on Cluster similarity using Rough Entropy for Categorical Data Clustering

[1] Mr.G.Sreenivasulu [2] Dr.S.Viswanadha Raju [3] Dr.N.Sambasiva Rao

Associate Professor Department of CSE ACE Engineering College
Phone:+91-9908274630
E-Mail:gvsreenu@gmail.com
Professor of CSE JNTUHCEJ Hyderabad
svraju.jntu@gmail.com
Professor of CSE SRITW,Warangal
snandam@gmail.com

Abstract: Clustering is one of the major issues in data mining. Data labeling has been recognized as an important method in categorical clustering. Clustering is technique where all similar data point are grouped. However, with data labeling is applied on those points which are not labeled earlier. Although there are many approaches in the numerical domain, but very limited algorithms are available for categorical data. To address this problem of how to allocate those unlabeled data points into proper clusters remains as a challenging issue in the categorical domain. In this paper, a mechanism is proposed for labeling and keeping the similar data points into accurate clusters. We have a data set named Genome DNA where grouping of 'superfluous' Splice junctions on those points on a DNA sequence is a major challenge. The predicament posed in this dataset is to recognize, given a sequence of DNA, the limits between exons and introns. The new proposal is to allocate each unlabeled data point into the equivalent proper cluster with data labeling also. This method has two advantages: 1) The proposed method exhibits high execution efficiency. 2) This method can achieve quality clusters. The proposed method is empirically validated on DNA data set, and it is shown significantly more efficient than prior works while attaining results of high quality.

Keywords—Clustering; Categorical Data; Clustering; Data Labeling; Outlier; Entropy; Rough set;.

I. INTRODUCTION

In Data Mining [2] clustering is a major challenge. It is used to group similar objects as one [1,3]. These kinds of groups are often known as clusters. The extent of grouping mechanisms have been complete in Information Retrieval Systems, Medical diagnosis, statistics, and pattern recognition and machine learning, etc. The complete extent on clustering procedure can be originating in [3] various types. Numeric, Mixed and categorical data are the different types in data set. For Numeric data greater type of procedures are available when compared to other two [5][6] data types. In categorical data clustering is a complicated task, where the distance between data points is not accurate, when the data is increased on time. Clustering an enormous data set is a difficult concern in its intricacy it poses and the time it takes for the process. [7, 8] In clustering sampling is another method used to pick up the capability of clustering by selecting some data points arbitrarily for early clustering and regard as the data points which are un labeled (that are not sampled and are not clustered) to opt for customs and means to allot them into suitable clusters. This is called cluster labeling [9, 10, and 11].In categorical field numerical field is not that much straight forward in finding the class field. In Data Mining, concept Drift is time overwhelming. [12,16]. The time budding data in the numerical field for clustering [1, 5, 6, 10] has been explored in the last study literature, however not much more was addressed in categorical domain. So, still it is a main trouble in the categorical data. As a result, our research in changing methodology Ming-Syan Chen framework 2009 [8] uses any clustering method to find out the drifting practice.

This paper explains about the method and working results on a large data set for data labeling in Rough Set theory. It is an influential mathematical theory; it has been productively useful in Internet of Things (IOT), Wireless adhoc networks, dimensionality reduction, machine learning, pattern recognition and etc.,

The rest of the paper is as follows, Thrash out in section II is about analysis of related literature; Basic definitions in section III are discussing about the entropy model inside rough set theory, Data Labeling in section IV discusses strategy for categorical clustering, in section V investigational results are shown and in section VI the conclusions and future work are discussed.

II. REVIEW OF RELATED LITERATURE

We are going to offer whole representation of clustering methodology on categorical data for data labeling next to cluster council in this part[11, 12, 20]. Clustering huge data set is a time taking process and is not an ample task. To review and illustrate the clustering BIRCH, Balanced Reduced Iterative clustering Hierarchies (BRICH) is a speed clustering procedure to grasp 'noise' proficiently designed by Zhang[23]. BRICH will find a clustering and by using a little bit more scans it enriches the superiority of clusters in a first round scan of data. CLARAN[21] is uniformly inferior to BRICH, a clustering method probable for huge datasets. The best hierarchical clustering approach which avoids the troubles with non-Uniform calculated or shaped clusters is CURE [25]. In CURE, they are shrunk towards the mean of the cluster by partitions after identifying a fixed number and well marked objects of a cluster and these points are known as council of the cluster. Since, it is agglomerative and divisive (hierarchical) the clusters with the adjacent near pain of council are compounded in each step. It makes it partially sensitive to outliers since, it properly identifies the clusters. In next clustering approach K-modes [26], the quality purity value in each object domain of a cluster council the most frequent for that cluster. Pronouncement mode may be elegant, but the course of using only one attribute value in each attribute field to refer a cluster is uncertain.

The next procedure is ROCK. It is a process of agglomerative clustering algorithm. It functions based on data points and associations in between, to conquer the ambiguity the sequence of links in data are emerged. For grouping of categorical data values, (CACTUS) Categorical clustering using summaries is a summation based approach planned by Ramakrishnan [28]. The innovation at the back CACTUS is there," The complete dataset gross is enough to measure a set of runner clusters that can be validated to check the material set of clusters".

There are two additional clustering algorithms that are well-liked in categorical domain named COOLCAT and LIMBO. COOLCAT, The perfect entropy data standards is estranged in such a way that the conventional entropy of the complete preparations is minimized. LIMBO algorithm, in sequence bottleneck method is functional to reduce the information loss which consequential from summing data points into groups. However, these methods carry out clustering based on dipping or maximum the statistical goal function, and the clustering council in these methods are not properly specified. So, the summarization and quality information of the clustering points cannot be taken by using these algorithms. A dissimilar method based on rough set model with practical results is the objective of the paper.

The rough theory is aimed on the hypothesis that with every point of the world there is linked a convinced amount of information, articulated by means of some objects which are used for data point description. Objects having the same explanation are like with deference to the accessible information. The resemblance relation thus generated constitutes a mathematical model of the rough theory.

III. ENTROPY MODEL IN ROUGH SETS

Data is a real world Object, it is shown by categorical objects or values. In this context each tuple represents a lot of attributes. Formally a categorical data table can be defined as a quadruple Dataset D= (U,A,V,f), where

U –set of non empty objects, it is also called as universe;

A – a set of non empty attributes;

V – It is a union of all attribute fields, i.e., $V = \bigcup_{a \in A}$, where V_a is the field of quality a and it is limited and finite order;

f : U X A —>V – a mapping function is called an information function such that for any x belongs to U and a A, f(x,a) Va.

The clustering process of the categorical domain time-changing data is mentioned as this: Dataset D is a collection of n data values, where each data value is an array of q attribute values is $x_j=(x_j^1, x_j^2, \ldots x_j^q)$. For example A= $(A_1, A_2 \ldots A_q)$, where A_a is the a^{th} cate object, N be sliding window defined size. Conquer the n data points into equal width windows and call this group as S^t, at time interval t. So, first N data points of dataset D are situated in the first group S^1 and next sliding window data points of D are positioned in the second subset S^2 and so on. The aim of our proposed method is to take S^{t+1}, as an unknown data set and represent these data points into the clusters which are gained from S^t.

For example consider the following D as a Dataset=$\{x_1, x_2, \ldots x_{14}\}$ of 14 points shown in below and the sliding window size N is 7 then S^1 represents Initial 7 data points and next sliding window S^2 contains next 7 points shown in below TABLE I. Apply any one clustering approach on S^1 to break the data points into two clusters shown in below TABLE II. The points that are clustered are called grouped data points or labeled or classified data points and the remaining are called unlabeled points. Our methodology is to make the remaining 7 unlabeled data points keeping them into proper clusters which belong to next sliding window S^2.

Table I. A data set D with 14 data points divided into two equal size sliding windows S^1 and S^2

S1							
	x_1	x_2	x_3	x_4	x_5	x_6	x_7
F_1	F	H	F	F	R	S	R
F_2	N	N	N	N	N	N	L
F_3	I	I	J	J	Q	Q	Q

S2							
	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
F_1	F	H	F	S	H	I	S
F_2	N	E	L	L	L	E	N
F_3	I	G	J	Q	J	G	Q

A. Choose any existing clustering algorithm like K-means and K-modes on S1 to divide data points into clusters shown in the below TABLE II. The data points that are grouped are called clustered data points or labeled points and the left over are called unlabeled data points. Our objective is to label the remaining 7 unlabeled data points which belong to next sliding window S2.

TABLE II TWO CLUSTERS C_1^1 AND C_2^1 AFTER PERFORMING A CLUSTERING METHOD ON S^1

C_1^1					C_2^1		
x_1	x_2	x_3	x_4		x_5	x_6	x_7
F	H	F	F		R	S	R
N	N	N	N		N	N	L
I	I	J	J		Q	Q	Q

Rough Set theory Definition **:** Let D = (U, A, V, f) be an data base system. For any $P \subseteq A$, let $U/IND(P)=\{P_1, P_2, \ldots P_m\}$ be a partition for U induce by IND(P). The base system which can divided into k-cluster i.e $c_k=\{c_1, c_2, c_3,\ldots c_k\}$. For any $c_k \in c^k$. Rough Entropy RE(P) of similarity relation IND(P) is distinct as $RE(C_i, x_i) = \frac{1}{k}\left|\sum_{i=1}^{n} \frac{|x_i \backslash x_i[i]=x_j[i] \text{ and } x_i \in c_i|}{|x_i \backslash x_i[i]=x_j[i]|} \times \frac{|x_i \backslash x_i[i]=x_j[i]|}{|c_i \cup c_j|}\right|$. where $\frac{1}{k}$ denotes the number of clusters, with in the c_i cluster and $|x_i \backslash x_i[i] = x_j[i] \text{ and } x_i \in c_i|$ denotes the similarity of any element $x \in U$ being in likeness class P_i, with in the c_j cluster, and $|x_i \backslash x_i[i] = x_j[i]|$ denotes that comparison of attributes within S^i on C_i,and $|C_i \cup C_j|$ gives that number of diverse objects in S^i on C_i . It is noted that $0 \le RE(P) \le 1$ and RE(P) attains its maximum value 1 when $U/IND(P)=\{\{X\}:X \epsilon U\}$. This entropy assess is used in various levels and finds application in many domains . RE(P) has maintained its entrance value to 0.5.

IV. Rough Entropy based data labeling

In rough set theory, rough entropy is an uncertainty information measure. However, in rough set community there are few concerns about the problem of clustering the data based on data labeling. In the following subsection some basic definitions to perform data labeling using rough entropy calculating and how to label the data using this measure are discussed.

Let D= (U, A, V, f) be a base system equivalent to the cluster c_i which is obtained from sliding window S^1 for i=1 to k (k is the number of clusters obtain from sliding window S^1 by using any clustering technique) .

Example1:

Consider the data in below TABLE I with two sliding windows. After applying a clustering method on sliding window S^1, the clusters c_1^1 and c_2^1 are formed as shown in TABLE II. Now in view of the data points of sliding window S^2 , Data labeling of each data points are discussed in this section.

for each U with respect all attributes $a \in A_1$ is The partition calculated by using formula (1) as

$U/IND(a)=\{\{x_1, x_3, x_4, x_5\}\{x_2\}\}$

$U/IND(b)=\{\{x_1\}\{x_3, x_5\}\{x_2, x_4\}\}$

$U/IND(c)=\{\{x_1\}\{x_3\}\{x_5\}\{x_2, x_4\}\}$

The rough entropy for c_i with esteem all attributes $a \in A_1$ is calculated by using formula (3). By applying this formula on c_1 importance is each attribute is as follows

$RE(C_1^1, x_6)=\frac{1}{3}\left(\frac{1}{4} \times \frac{4}{5} + \frac{2}{5} \times \frac{2}{3} + \frac{1}{1} \times \frac{1}{3}\right)=1.208$

$RE(C_1^1, x_7)=\frac{1}{3}\left(\frac{3}{4} \times \frac{4}{5} + \frac{0}{5} \times \frac{0}{3} + \frac{1}{1} \times \frac{1}{4}\right)=0.875$

$RE(C_1^1, x_8)=\frac{1}{3}\left(\frac{0}{4} \times \frac{0}{5} + \frac{0}{5} \times \frac{2}{3} + \frac{0}{0} \times \frac{0}{3}\right)=0$

$RE(C_1^1, x_9)=\frac{1}{3}\left(\frac{0}{4} \times \frac{1}{5} + \frac{0}{0} \times \frac{0}{3} + \frac{0}{0} \times \frac{0}{3}\right)=0$

RE(C_1^1,x_8)=$\frac{1}{2}\left(\frac{0}{0}\times\frac{0}{2}+\frac{0}{0}\times\frac{0}{3}+\frac{0}{0}\times\frac{0}{2}\right)$=0

Similarly apply this method on cluster c_2

RE(C_1^2,x_6)=0.25

RE(C_1^2,x_7)=0.25

RE(C_1^2,x_8)=0.33

RE(C_1^2,x_9)=1.58

RE(C_1^2,x_{10})=0

Now consider the unlabeled data points surround in S^2 , that is shown Table II, now take object x_6 contains attributes as {B,E,Q} . Let a=A_1 , according to the resemblance of x_6 , it moves into the cluster c_1. Similarly object X_7 , contains the attribute as {B,M,Q} , having similarity with c_1 so it moves into c_1 cluster. Now consider the object x_8 which is not related to any cluster even though it is having similarity with c_2 but it is not reaching threshold value. So that it is considered as outlier. All data points in S^2 can move into their appropriate clusters shown in below Table V.

Table III: Data pints representing belonging clusters

Object	Cluster Label
X_6	C_1
X_7	C_1
X_8	Outlier
X_9	C_2
X_{10}	Outlier

Algorithm for data labeling based using Data Labeling.

Algorithm: Rough Entropy based Data Labelling Algorithm

Input: Dataset D with n data points, sliding window size N.

Output: Number of outliers.

Method:

Step 1: Divide the dataset D into equal size sliding windows based on given sliding window size N, say those sliding windows are S_1, S_2 ….

Step 2: Use any clustering algorithm on sliding window S^1 to obtain first clustering result C_1 with clusters $c_1^1, c_2^1 \ldots . c_k^1$. Let IS_i =(U_i, A_i, V_i, f_i) be an information system of the cluster $c^1{}_i$ of clustering result C_t for t=1, 2...Where t is a timestamp.

Step 3: out=0

Step 4: For every unlabeled data point $P_j \in S^{t+1}$ sliding window, start with t=1 begin.

Step 5: For every cluster $c^t{}_i$ begin

Step 6: For every a$\in A_i$ begin

Step 7: Find the partition U/IND({a}) using (1)

Step 8: Find the rough entropy RE({a}).(2)

Step 9: end

Step 10: Calculate rough entropy $RE_{p_j}(a)$ (3)

Step 11 : Set threshold to 0.5 .

Step 12 : End

Step 13: Take object of the next sliding window and move the objects into proper clusters based Importance of the each attribute of object.

Step 14**:** return out.

V. EXPERIMENTAL RESULTS

In this section, we demonstrate the performance of the proposed work on clustering categorical data by a thorough experimental study on the real dataset. In Section V.I, the test environment and the dataset used are described. Next section, the evolving processes of clustering results are visualized on the real dataset.

Table IV: Complete comparison of efficiency with various algorithms by taking consideration of Sliding window size, No. of attributes and Cluster size.

(Cluster size=10, Sliding window size=1000 and No. Of attributes=42)						
S.No	**Method**	**Data Size**	**C.S**	**S.W.S**	**Attribute**	**Time**
1.	REDLA	100000	10	1000	42	745869.6
2.	OUR	100000	10	1000	42	935869.6
3.	Pour-NIR	100000	10	1000	42	1045870
4.	REDLA	65000	10	1000	42	52756.79
5.	OUR	65000	10	1000	42	485567.80
6.	Pour-NIR	65000	10	1000	42	371526.30
7.	REDLA	50000	10	1000	42	44853.25
8.	OUR	50000	10	1000	42	406872.90
9.	Pour-NIR	50000	10	1000	42	320145.60
10.	REDLA	40000	10	1000	42	32351.85
11.	OUR	40000	10	1000	42	378841.70
12.	Pour-NIR	40000	10	1000	42	297546.30
13.	REDLA	30000	10	1000	42	23734.04
14.	OUR	30000	10	1000	42	297260.30
15.	Pour-NIR	30000	10	1000	42	207458.40
(Cluster size=10, Sliding window size=2000 and No. Of attributes=42)						
16.	REDLA	100000	10	2000	42	457869.5
17.	OUR	100000	10	2000	42	957869.5
18.	Pour-NIR	100000	10	2000	42	647869.5
19.	REDLA	65000	10	2000	42	195717.9
20.	OUR	65000	10	2000	42	245791.4
21.	Pour-NIR	65000	10	2000	42	209627.8
22.	REDLA	50000	10	2000	42	181928.8
23.	OUR	50000	10	2000	42	194687.6
24.	Pour-NIR	50000	10	2000	42	194242.8
25.	REDLA	40000	10	2000	42	147287.9
26.	OUR	40000	10	2000	42	178979.4
27.	Pour-NIR	40000	10	2000	42	157949.5
28.	REDLA	30000	10	2000	42	107548.9
29.	OUR	30000	10	2000	42	147989.9
30.	Pour-NIR	30000	10	2000	42	127548.9
(Cluster size=10, Sliding window size=3000 and No. Of attributes=42)						
31.	REDLA	100000	10	3000	42	35785.96
32.	OUR	100000	10	3000	42	75785.96
33.	Pour-NIR	100000	10	3000	42	58785.96
34.	REDLA	65000	10	3000	42	34578.67
35.	OUR	65000	10	3000	42	43745.41
36.	Pour-NIR	65000	10	3000	42	40670.46
37.	REDLA	50000	10	3000	42	28784.74
38.	OUR	50000	10	3000	42	32547.25
39.	Pour-NIR	50000	10	3000	42	30366.79
40.	REDLA	40000	10	3000	42	21476.17
41.	OUR	40000	10	3000	42	26748.97

42.	Pour-NIR	40000	10	3000	42	22538.86
43.	REDLA	30000	10	3000	42	18255.19
44.	OUR	30000	10	3000	42	20897.25
45.	Pour-NIR	30000	10	3000	42	20255.19
(Cluster size=10, Sliding window size=1000 and No. Of attributes=30)						
46.	REDLA	100000	5	1000	30	55915.5
47.	OUR	100000	5	1000	30	62725.27
48.	Pour-NIR	100000	5	1000	30	59205.24
49.	REDLA	65000	5	1000	30	46884.61
50.	OUR	65000	5	1000	30	50214.8
51.	Pour-NIR	65000	5	1000	30	49817.58
52.	REDLA	50000	5	1000	30	31794.8
53.	OUR	50000	5	1000	30	40908.56
54.	Pour-NIR	50000	5	1000	30	35898.75
55.	REDLA	40000	5	1000	30	27697.75
56.	OUR	40000	5	1000	30	37789.08
57.	Pour-NIR	40000	5	1000	30	36785.89
58.	REDLA	30000	5	1000	30	18584.85
59.	OUR	30000	5	1000	30	22654.75
60.	Pour-NIR	30000	5	1000	30	20584.85
(Cluster size=10, Sliding window size=2000 and No. Of attributes=30)						
61.	REDLA	100000	5	2000	30	49817.4
62.	OUR	100000	5	2000	30	52614.64
63.	Pour-NIR	100000	5	2000	30	51415.44
64.	REDLA	65000	5	2000	30	38745.54
65.	OUR	65000	5	2000	30	45979.57
66.	Pour-NIR	65000	5	2000	30	40789.51
67.	REDLA	50000	5	2000	30	34577.71
68.	OUR	50000	5	2000	30	39745.75
69.	Pour-NIR	50000	5	2000	30	38674.54
70.	REDLA	40000	5	2000	30	29578.47
71.	OUR	40000	5	2000	30	32745.21
72.	Pour-NIR	40000	5	2000	30	32367.68
73.	REDLA	30000	5	2000	30	22647.65
74.	OUR	30000	5	2000	30	28647.26
75.	Pour-NIR	30000	5	2000	30	25647.65
(Cluster size=10, Sliding window size=3000 and No. Of attributes=30)						
76.	REDLA	100000	5	3000	30	39687.63
77.	OUR	100000	5	3000	30	45291.26
78.	Pour-NIR	100000	5	3000	30	43783.43
79.	REDLA	65000	5	3000	30	30784.74
80.	OUR	65000	5	3000	30	35479.24
81.	Pour-NIR	65000	5	3000	30	34378.45
82.	REDLA	50000	5	3000	30	27487.57
83.	OUR	50000	5	3000	30	29974.55
84.	Pour-NIR	50000	5	3000	30	28854.17
85.	REDLA	40000	5	3000	30	19475.78
86.	OUR	40000	5	3000	30	25745.69
87.	Pour-NIR	40000	5	3000	30	20086.75
88.	REDLA	30000	5	3000	30	16393.08
89.	OUR	30000	5	3000	30	19574.54
90.	Pour-NIR	30000	5	3000	30	18393.08

(Cluster size=10, Sliding window size=1000 and No. Of attributes=20)						
91.	REDLA	100000	3	1000	20	48689.67
92.	OUR	100000	3	1000	20	53674.86
93.	Pour-NIR	100000	3	1000	20	51689.47
94.	REDLA	65000	3	1000	20	39788.87
95.	OUR	65000	3	1000	20	51880.31
96.	Pour-NIR	65000	3	1000	20	49889.08
97.	REDLA	50000	3	1000	20	30475.65
98.	OUR	50000	3	1000	20	45247.26
99.	Pour-NIR	50000	3	1000	20	38795.55
100.	REDLA	40000	3	1000	20	21589.36
101.	OUR	40000	3	1000	20	38548.7
102.	Pour-NIR	40000	3	1000	20	29695.89
103.	REDLA	30000	3	1000	20	17964.46
104.	OUR	30000	3	1000	20	35486.21
105.	Pour-NIR	30000	3	1000	20	20964.46
(Cluster size=10, Sliding window size=2000 and No. Of attributes=20)						
106.	REDLA	100000	3	2000	20	52795.74
107.	OUR	100000	3	2000	20	55153.62
108.	Pour-NIR	100000	3	2000	20	54645.82
109.	REDLA	65000	3	2000	20	26785.57
110.	OUR	65000	3	2000	20	31449.66
111.	Pour-NIR	65000	3	2000	20	30449.66
112.	REDLA	50000	3	2000	20	18748.82
113.	OUR	50000	3	2000	20	29157.57
114.	Pour-NIR	50000	3	2000	20	20634.75
115.	REDLA	40000	3	2000	20	16874.58
116.	OUR	40000	3	2000	20	22357.86
117.	Pour-NIR	40000	3	2000	20	19548.55
118.	REDLA	30000	3	2000	20	9675.387
119.	OUR	30000	3	2000	20	17358.12
120.	Pour-NIR	30000	3	2000	20	10675.39
(Cluster size=10, Sliding window size=3000 and No. Of attributes=20)						
121.	REDLA	100000	3	3000	20	31546.28
122.	OUR	100000	3	3000	20	37688.34
123.	Pour-NIR	100000	3	3000	20	34646.67
124.	REDLA	65000	3	3000	20	18546.28
125.	OUR	65000	3	3000	20	22368
126.	Pour-NIR	65000	3	3000	20	20274
127.	REDLA	50000	3	3000	20	14754.97
128.	OUR	50000	3	3000	20	18237.57
129.	Pour-NIR	50000	3	3000	20	16246.78
130.	REDLA	40000	3	3000	20	9145.756
131.	OUR	40000	3	3000	20	12127.63
132.	Pour-NIR	40000	3	3000	20	10127.3
133.	REDLA	30000	3	3000	20	8373.364
134.	OUR	30000	3	3000	20	10573.26
135.	Pour-NIR	30000	3	3000	20	9573.57

V.I Test Environment and Dataset:

All of our experiments are conducted on a PC with an Intel Corei5 processor with 4 GB memory and the Windows7 professional operating system. In the experiment, the k-modes [6] clustering algorithm is chosen to do the initial clustering and reclustering on the datasets. As the k-modes algorithm is dependent on the selection of initial cluster centers, we utilize an initialization method, which was proposed in [7], to obtain initial cluster centers before executing the k-modes. For developing this paper we use the .NET language and backend as MySql . DNA Splice junction points are on a DNA sequence at which `superfluous' DNA is terminated during the method of protein creation in higher organisms. The problem posed in this dataset is to recognize, given a sequence of DNA[39][40], the boundaries between exons and introns. [8], This dataset has been developed to help evaluate a "hybrid" learning algorithm that uses examples to inductively refine preexisting knowledge. Which has been used earlier to assess several stream-clustering methods and DCDAs, is used in our study. Therefore, this dataset is time-evolving data and is appropriate to asses our algorithms. We utilize the 10% subset version, which is provided from the DNA website for our experiments. In this dataset, there are 24,58,285 records, and each record contains 68 attributes (class label is included), such as the duration of the connection, We accept identical quantization on those numerical attributes where each attribute is quantized into five categorical values.

Evaluating scalability:

To test the scalability of the wk-modes algorithm, we use a synthetic data generator [38] to generate datasets with different number of objects and attributes. The number of objects varies from 10,000 to 100,000, and the dimensionality is in the range of 10–50. In all synthetic datasets, each dimension possesses five different attribute values. As the different clustering results will be obtained on the same dataset when we select different initial cluster. Table VI shows the execution times of records by applying the clustering algorithm, Ming chen method and Proposed method, Therefore, each value in Table VI is the average of 10 times experiments.

Table V: performance of various algorithms

Data Records	K-Modes algorithm	Ming-Chen Method	Proposed Method
10,000	0.866	0.7293	0.4425
20,000	1.732	1.5496	1.0254
30,000	2.598	2.1869	1.5847
40,000	3.464	2.9172	1.9657
50,000	4.334	3.6465	2.2548
100,000	9.576	8.3976	4.3596

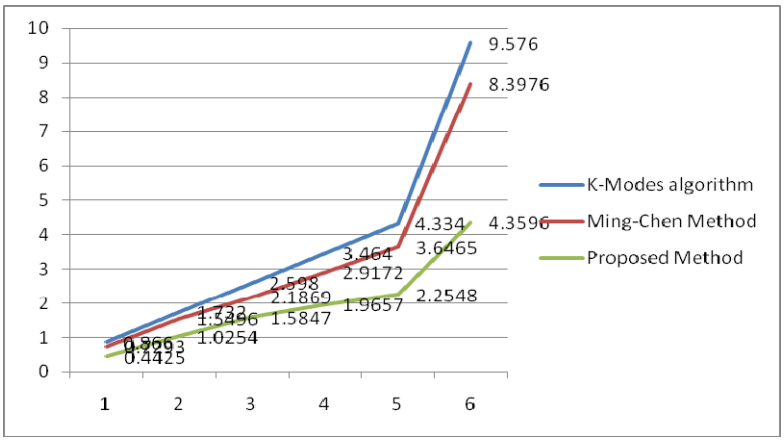

Fig: Representation of efficiency (Time in millisecond

This study fixed the dimensionality to10,and the cluster number to 3,and the data size varies from 10,000 to 100,000. It can be seen that the proposed algorithm is linear with respect to the data size. The execution time of the proposed method is very much effective then the k-modes algorithm and Ming-Chen Method. Therefore, the wk-modes algorithm can ensure efficient execution when the data size is large.

Fig. 1 shows the scalability with data size of three algorithms.

VI. CONCLUSIONS

In categorical domain, the problem of how to assign the unlabeled data points into suitable clusters has not been fully explored in recent research papers into data mining/clustering. Besides, for the data which changes eventually, clustering this kind of data not only decreases the excellence of clusters its also disregards the potential of users, when usually require recent clustering results. This paper, deals the method based on Rough Entropy correspondence measure for allocating the unlabeled data point into appropriate cluster has been defined. Outlier detection or clustering labeling is done based on variation in cluster similarity threshold using Rough Entropy. In future work, the concept drift can be detected using the above method whether it is occurred or not.

References

[1] Anil K. Jain and Richard C. Dubes. "Algorithms for Clustering Data", Prentice-Hall International, 1988.
[2] Clustering Categorical Data Using Summaries," ACM SIGKDD, 1999.
[3] Jain A K MN Murthy and P J Flyn, "Data Clustering: A Review," ACM Computing Survey, 1999.
[4] Fredrik Farnstrom, James Lewis, and Charles Elkan," Scalability for clustering algorithms revisited", ACM SIGKDD pp.:51–57, 2000.Barbara, D., Li, Y. and Couto, J. "COOLCAT: [5]An Entropy-Based Algorithm for Categorical Clustering", ACM International
[5] Kaufman L, P. Rousseuw," Finding Groups in Data- An Introduction to Cluster Analysis", Wiley Series in Probability and Math. Sciences, 1990.
[6] 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.
[7] http://archive.ics.uci.edu/ml/machine-learning-databases/molecular-biology/splice-junction-gene-sequences/splice.names.
[8] G.Sreenivasuluy, S.Viswanadha Raju et al." A Threshold for clustering Concept – Drifting Categorical Data", IEEE 3 rd International Conference on Machine Learning and Computing (ICMLC), Volume 3, Feb-2011, pp. 383-387.(ISBN: 978-1-4244-9253/11 IEEE).
[9] Han,J. and Kamber,M. "Data Mining Concepts and Techniques", Morgan Kaufmann, 2001.
[10] Ganti, V., Gehrke, J. and Ramakrishnan, R, "CACTUS—.Vapnik, V.N," The nature of statistical learning theory", Springer,1995.
[11] 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.
[12] G.Sreenivasulu, S.Viswanadha Raju et al. " A Comparative Study Of Node Importance In Categorical Clustering ", International Journal of Advanced Engineering and Global Technology (IJAEGT), Volume. 1, No. 1, July 2013, PP.784-788(ISSN: ISSN No:2309-4893 (print) | ISSN: 0975-397(online)).
[13] Bradley,P.S., Usama Fayyad, and Cory Reina," Scaling clustering algorithms to large databases", Fourth International Conference on Knowledge Discovery and Data Mining, 1998.
[14] 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.
[15] G.Sreenivasulu, S.Viswanadha Raju et al." Data Labeling Method Based On Rough Entropy For Categorical Data Clustering", International Conference On Electronics, Communication And Computational Engineering - ICECCE 2014, pp. 383-387.(ISBN: 978-1-1170-1175/11/ IEEE
[16] Chen. H. L., Chuang K.T. and Chen. M.S (2008), "On Data Labeling for clustering Categorical data", IEEE Transactions on knowledge and Data Engineering, 20(2011), 1458-1471.
[17] Fuyuan Cao, Jiye Liang, "A Data Labeling method for clustering categorical data", Elsevier Expert systems with applications, 38(2011), 2381-2385.
[18] 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.
[19] Xiangjun Li, Fen Rao, "An Rough Entropy Based Approach to Outlier Detection", Journal of Computational Information Systems 8: 24 (2012) 10501-10508.
[20] Klinkenberg, R.," Using labeled and unlabeled data to learn drifting concepts", IJCAI-01Workshop on Learning from Temporal and Spatial Data, pp. 16-24, 2001.
[21] Pawlak, "Rough sets ", International journal of computer and information sciences, 11(1982), 341-356.
[22] D. Parmer, T. Wu and J. Blackhurst, MMR, "An Algorithm for clustering data using rough set theory", Data and Knowledge Engineering, 63(3)(2007), 879-893.
[23] Gluck, M.A. and Corter, J.E. "Information Uncertainty and the Utility of Categories", Cognitive Science Society, pp. 283-287, 1985.
[24] G.Sreenivasulu, S.Viswanadha Raju et al. "A Review of clustering techniques ", International Conference on Data Engineering and Communication Technology (ICDECT), Springer, March-2016.J ISSN: 2250-3439).
[25] Shannon, C.E, "A Mathematical Theory of Communication," Bell System Technical J., 1948.
[26] Chun-Bao Chen, Li-Ya Wang, "Rough Set-Based Clustering with refinement Using Shannon's Entropy Theory", ELSEVIER Computers and Mathematics with Applications 52 (2006) 1563-1576.
[27] H.Venkateswara Reddy, S.Viswanadha Raju. "A Study in Employing Rough Set Based Approach for Clustering on Categorical Time-Evolving Data", IOSR Journal of Computer Engineering (IOSRJCE), Volume 3, Issue 5 (July-Aug. 2012), PP 44-51 (ISSN: 2278-0661) DOI number 10.9790/0661-0354451.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.
[28] 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.
[29] G.Sreenivasulu, S.Viswanadha Raju. "A Proficient approach for clustering of large categorical data cataloguing", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) March- 2016),IEEE, ISSN 978-1-4673-9939..
[30] G.Sreenivasulu,Venkateswara Reddy.H, Viswanadha Raju.S," A Threshold for clustering Concept – Drifting Categorical Data", IEEE Computer Society, ICMLC 2011.
[31] 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.
[32] Ng, R.T. Jiawei Han "CLARANS: a method for clustering objects for spatial data mining", Knowledge and Data Engineering, IEEE Transactions, 2002.
[33] Huang, Z. and Ng, M.K, "A Fuzzy k-Modes Algorithm for Clustering Categorical Data" IEEE On Fuzzy Systems, 1999.
[34] Guha,S., Rastogi,R. and Shim, K, "ROCK: A Robust Clustering Algorithm for Categorical Attributes", International Conference On Data Eng. (ICDE), 1999.

[35] Conf. Information and Knowledge Management (CIKM), 2002.
[36] Andritsos, P, Tsaparas, P, Miller R.J and Sevcik, K.C."LIMBO: Scalable Clustering of Categorical Data", Extending Database Technology (EDBT), 2004.
[37] G.Sreenivasulu, S.Viswanadha Raju et al." Graph Based Approach for Clustering Categorical Data", International Journal of Advanced Computing (IJAC), pp. 117-125.(ISBN: ISSN : 0975-7686.
[38] Chinta Someswara Rao, S. Viswanadha Raju etc., " Similarity analysis between chromosomes of Homo sapiens and monkeys with correlation coefficient, rank correlation coefficient and cosine similarity measures" in Elsvier, Genomics Data 7 (2016) 202–209. [40]. Chinta Someswara Rao , S. Viswanadha Raju etc., " Next generation sequencing (NGS) database for tandem repeats with multiple pattern 2°-shaft multicore string matching" in Elsvier, Genomics Data 7 (2016) 307–317.