International Journal of Engineering & Technology, 7 (4.6) (2018) 68-72

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Data labeling method based on Cluster similarity using Rough Entropy for Categorical Data Clustering

B .Suresh Kumar[1] , Dr. H.Venkateswara Reddy[2] , Dr S.Viswanadha Raju[3], G Vijay Kanth[4]
[1,2,3]Department of C.S.E
[1,2] Vardhaman College of Engineering
[3]JNTUH College of Engineering
4Department of Computer Science and Engineering MLR Institute of Technology, Hyderabad

**Corresponding author E-mail: sureshkumargoud2006@gmail.com*

Abstract:

In present research, Data mining is become one of the growing area which deals with data. Clustering is recognized as an efficient methodology in data grouping; to improve the efficiency of the clustering many researchers have used data labeling method. Labeling method works on similar data points, into the proper clusters. In categorical domain applying data labeling is not so easy when compare with numerical domain. In numeral domain it is easy to find difference between to data points, but in categorical it is not easy. Since data labeling on categorical is a challenging issue till date and it is quite complex to implement. The proposed methodology is deals on this problem. According proposed method a sample data will be taken. That sampled data further divides sliding windows, and then a normal clustering algorithm will be applied on one sliding window and divides into clusters. Rough membership Entropy function is used to find the similarity between unlabelled data points to labeled data points. The proposed methodology has two important features those are 1) The Data points will moved into their proper clusters, means the quality clusters will take places, 2) Proposed methodology will execute with high efficiency rate. In this paper the proposed methodology is applied on KDD Cup99 data sets, and the results shows appreciably more proficient than earlier works.

Keywords: *Categorical Data; Clustering; Data Labeling; Outlier; Entropy; Rough set.*

1. Introduction

Similarity based data grouping is the major technology in data mining which is called clustering [1,2,3]. Mainly this cluster concept is used in statistics, machine learning algorithms, pattern recognition, etc…Survey on this clustering methods can be found in[2,3]. Data has different formation data can be in numerical, categorical, mixed type of data. Based on Numerical data lots of algorithms have been proposed [5] because it easy to find distance but when it comes to categorical it not that easy be find distance between two categorical data attributes [6]. In categorical field clustering is a demanding and critical task. Day to day in categorical domain the data is growing rapidly and clustering on this type of data is also difficult and time consuming process [7, 8], so in order to overcome this type of challenges sampling technique is used. Data sampling is deals with selecting of some random data points to perform the initial clustering; when the initial clustering is done clustering of unlabeled points which are not clustered has clustered into their proper clusters. This process of clustering unlabelled data points into proper clusters is called data labeling [9, 10, 11]. The existing method in [12] discussed about categorical data labeling is not an easy process unlike in numerical. Data increasing with time and the process of clustering is also change based on time using the concept drift [12, 16]. Ming-syan Chen in [11] has proposed a frame work, which uses any kind of clustering algorithms to detect the drifting on the data sets. The method proposed in [11] is referred as Chen method. In this paper the proposed methodology has adopted sliding window technique for initial cluster. This paper proposes a method on class labeling by introducing Rough Set Theory (RST) which is a powerful mathematical tool successfully applied in different algorithms. The organization of this paper is as follows; section II presents the pertinent background work; section III supplies essential description and entropy representation in rough sets; section IV

confer the clustering of unlabeled data points, In Section V experimental results are analyzed and finally conclusion of the paper has delivered with some future extension

2. Review Of Related Literature

The related work on clustering algorithms has been discussed in this section. In [11,12,20] methods converse about clustering in categorical domain along with representatives and data labeling. Clustering huge data is tricky and time taking process. Cluster representatives are utilized to differentiate the clustering result, which is not found complicated state in categorical area not like in the numerical domain [21, 22].
The algorithm proposed in [23] works fast and it can handle any type of noise successfully. The algorithm is called BRICH, at a time it will take data and finds appropriate clusters by using a few scans methods to improve the quality of the cluster. BRICH reliability is greater to CLARANS [24], a clustering

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

process projected for huge datasets. The method in [25], is a new clustering algorithm deals with the problem of non-Uniform sized clusters. In CURE[25], after selecting a fixed number and fine spotted point of a cluster, those can move towards the centric of the cluster by portion. Those points can be treated as cluster representatives.

This method proposed in[25] is hierarchical model means every time the closest points cluster representatives are mixed with clustered points. So the algorithms results to quality clusters and outliers. In[26] K-modes algorithm is proposed which is most used algorithm for clustering. In this most frequent value has take as representative in cluster that point termed as mode of the cluster. Detecting the mode in a data set is very is process but the drawback this method is representing one attribute value in the attribute domain. In [27] ROCK algorithm is proposed which is the type of agglomerative hierarchical. This algorithm is developed based on links between the data points. The main advantage of considering link between nodes, it helps to overcome problem with distance among the data points. link (pi, pj) is normal used notation for link between points, here pi denotes point i, and similarly pj denotes point j. In ROCK clustering algorithm takes S as input set and n sample data points and the number of clusters denoted by k. The method begins with calculating the amount of links linking pair of points. Those links are used to cluster the algorithms. Initial steps in the algorithm is creates the Boolean contains 1 and 0 depending on neighbor matrix. 1 represents is the points are adjacent, 0 represents points are not adjacent. The main focus the ROCK algorithm is on adjacency of the data points, drawback is some data points are ignored.

Ganti et.al has proposed clustering algorithm which works on categorical data poitns[28]. The algorithms works using summaries so that it is termed as Categorical clustering using summaries(CACTUS). The basic idea behinde this algorithm is that " the total set of data is ample to estimate a set of candidate clusters . In some stastical categorical clustering algorithms like[29,30] and COOLCAT proposed in[31] ,LIMBO[32] are used stastical analysis to cluster the data points. In COOLCAT algorithm , data points are divided in such a way that the entropy of the data grouping is minimized. In LIMBO, information holdup method used to minimize the loses in the data. All thses algorithms are not providing quality clusters and intra cluster similairty is also less. This paper aims to propose a rough entropy model which provides the quality and accurate cluster points.

3. Entropy Model In Rough Sets

The data points are stored in a table,in that every represents objects and coloumn represents attributes. Normally data in real world represented in a categorical domain. The categorical domain mainly contains four tuples which can be represented as D=(U,A,V,f) where

U – represents the universe which is a nonempty set of objects,;

A – Atteibutes ; nonempty set

V – combination of all attribute domains, i.e.,$V = \cup_{a \in A} V_a$ where V_a is the domain of attribute a with it is limited and unordered set;

f : U X A —> V – mapping function such that for any x ∈U and a∈A, f(x,a) ∈Va.

The time evolving categorical data is represented as follows: D is set of categorical data points. Every data points consists of q attribute values i.e. $x_j=(x_j^1, x_j^2, \ldots x_j^q)$. Now let A= ($A_1$, A_2 … A_q), in this A_a is represents the a^{th} categorical attribute, 1≤a≤q, and N be window size. Construct n data points into equivalent size windows and consider this as S^t, at time slice t. i.e. Initially 'N' data points of categorical domain D are placed in the primary subset $S^1$and remaining data points of D are situated in the subsequent subset S^2 and so on. The goal of proposed technique is take S^{t+1} unlabeled data points and label these data points into proper clusters which are derived from S^t. Now Consider the TABLE I data set which consist of 10 data points D={$x_1, x_2, \ldots x_{10}$}.

Table I : Data points

Attributes/ Objects	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
A_1	B	C	B	B	B	B	B	A	C	A
A_2	F	M	E	M	E	E	M	M	M	K
A_3	G	P	D	P	Q	D	Q	C	P	C

Suppose if the sliding window contain size as 5 then S^1 includes initial 5 data points and S^2 includes subsequently 5 data points.

I : Divided into two equal size sliding windows S^1 and S^2

S^1							S^2					
A/O	X_1	X_2	X_3	X_4	X_5		A/O	X_6	X_7	X_8	X_9	X_{10}
A_1	B	C	B	B	B		A_1	B	B	A	C	A
A_2	F	M	E	M	E		A_2	E	M	M	M	K
A_3	G	P	D	P	Q		A_3	D	Q	C	P	C

In the initial stage the proposed method apply any existing categorical clustering algorithm on S^1 which results the sliding window into two proper clusters shown in TABLE III. In TABLE III data points that which are set into clusters are referred as clustered data points and the points in sliding window S^2 are considered as unlabeled data points. The main aim of proposed method is to label the unlabeled data points of sliding window S^2. Lots of clustering algorithms are proposed on clustering in the field of categorical domain. All the algorithms gives the accurate results on limited amount of data, if we apply large amount data points to this cluster algorithms may lead to losing of information, for that reason in this proposed methodology we are applying a K-Mode algorithm on the sampled data points.

TABLE III :TWO CLUSTER C_1^1 AND C_2^1 AFTER PERFOMING CLUSTERING METHOD ON S^1

Attributes/ Objects	X_1	X_3	X_5		Attributes/ Objects	X_2	X_4
A_1	B	B	B		A_1	C	B
A_2	F	E	E		A_2	M	M
A_3	G	D	Q		A_3	P	P

3.1 Rough Set Theory

This section deals with fundamental perception of rough set theory. Rough set theory is used in many real time applications like Artificial intelligence, cognitive science and etc...Rough set theory proposed in 1982 by Pawlak , it is type of envoy machine learning method for categorical data set consist of ambiguous information[13] .Rough set theory resides among the pair of approximation called lower and upper approximations.

Definition 1: Let IS = (U, A, V, f) be a categorical data points domain and a binary relation in discernibility between the data points is represented as IND (P) for any attribute subset P ⊆ A.

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

It is apparent that IND(P) is an corresponding relation on U and $IND(P) = \bigcap_{a \in P} IND(\{a\})$. Given P ⊆ A, the relation IND(P) bring a division of U, represented by $U / IND(P)$ and it is defined by

$$U / IND(P) = \{ \left[X \right]_P^U \mid x \in U \}, \quad (1)$$

Where $\left[X \right]_P^U$ indicates the similarity class determined by x with respect to P, i.e.,

$$\left[X \right]_P^U = \{ y \in U \mid (x, y) \in IND(P) \} \ . \quad (2)$$

Definition 2: Let D = (U, A, V, f) be an data base system. For any $P \subseteq A$, let U/IND(P)={P_1, P_2, ... P_m} be a panel for U persuade by IND(P). The data base system which can separated into k-cluster i.e

c_k={c_1, c_2, c_3,...c_k}. For any $c_k \in c^k$. Rough Entropy RE(P) of equivalence relation IND(P) is defined as

$$\mathrm{RE}\left(C_i \middle| [X_i]_{A_i}\right) = \frac{1}{|A_i|} \frac{\left|[X_i]_{A_i} \cap C_i\right|}{\left|[X_i]_{A_i}\right|}$$

(3)

Where $\Pr\left(C_i \middle| [X_i]_{A_i}\right)$ denotes equality of the data point X_i with attribute A_i in cluster C_i. $|A_i|$ represents the number of attribute points in the cluster. $\left|[X_i]_{A_i} \cap C_i\right|$ represents the similarity between a unlabelled data point[X_i] with attribute A_i in cluster C_i. It is well-known that $0 \le RE(P) \le 1$ and RE(P) achieves its greatest rate 1 when U/IND(P)={{X}:XϵU}.This entropy determine is used in a variety of levels and locates in many domains . RE(P) has preserves its porch value to 0.5.

4. Rough Entropy based data labeling

Rough entropy is works on the uncertainty information. Even though, in rough set consist few drawbacks on clustering the data based on data labeling it works efficiently. The following sub part deals with some necessary explanations to execute data labeling using rough entropy measure and gives the clear idea about how to label the unlabeled data.

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

Example1:

Consider the data set showing in TABLE I consist of two sliding windows. After performing a clustering method on sliding window S^1, the clusters C_1^1 and C_2^1 are created, that are shown in TABLE II. Currently taking into consideration of data points in sliding window S^2 , data labeling of every data points are shown in this part.

The separation for every U with respect all attributes $a \in A_1$ is designed by using formula (1) as

$[X_6]_{A_1}$={X_1,X_3,X_4,X_5}
$[X_6]_{A_2}$={X_3,X_5}
$[X_6]_{A_3}$={X_3}

Rough entropy for c_i with respect all attributes $a \in A_1$ is calculated by applying formula (3). By this method on c_1 significance is every attribute calculation is as follows

$\mathrm{RE}\left(C_1 \middle| [X_6]_{A_i}\right) = \frac{1}{3}\left(\frac{3}{4} + \frac{2}{2} + \frac{1}{1}\right) = 0.91$
$\mathrm{RE}\left(C_2 \middle| [X_6]_{A_i}\right) = \frac{1}{3}\left(\frac{1}{4} + \frac{0}{2} + \frac{0}{1}\right) = 0.08$
$\mathrm{RE}\left(C_1 \middle| [X_7]_{A_i}\right)$=0.58 $\mathrm{RE}\left(C_1 \middle| [X_8]_{A_i}\right)$=0
$\mathrm{RE}\left(C_1 \middle| [X_9]_{A_i}\right)$=0 $\mathrm{RE}\left(C_1 \middle| [X_{10}]_{A_i}\right)$=0

Similarly apply this method on cluster c_2
$\mathrm{RE}\left(C_2 \middle| [X_7]_{A_i}\right)$=0.41 $\mathrm{RE}\left(C_2 \middle| [X_8]_{A_i}\right)$=0
$\mathrm{RE}\left(C_2 \middle| [X_9]_{A_i}\right)$=1 $\mathrm{RE}\left(C_2 \middle| [X_{10}]_{A_i}\right)$=0

The unlabeled data points in S^2 having the similarities with labeled data points shown in the following TABLE III .
The unlabelled data points shown TABLE III shows the exactly where data point has clustered

Table Iv : Object Similarities With C_1 And C_2

Object	C_1	C_2
X_6	0.91	0.08
X_7	0.58	0.41
X_8	0	0
X_9	0	1
X_{10}	0	0

In Table IV, it is clearly shown that object x_6 which contains attributes as {B,E,D} ,based on the similarity of x_6 is moves into the cluster c_1. likewise object X_7 , consists of the attribute as {B,M,Q} , contain similarity with c_1 so it moves hooked on c_1 cluster. Now take object x_8 which is not connected to any cluster , So that it is considered as outlier. All data points in S^2 can move into their appropriate clusters shown in Table V.

Table V: Unlabelled data points clustering

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

Algorithm for Rough Entropy based data labeling using similarity measure of cluster points

Algorithm: Rough Entropy based Data Labeling

Input: 'D' as a Data set consist of n data points and the size of sliding window is N.

Output: Quality Clusters and Number of outliers.

Method:

Step 1: Split the D into equivalent size sliding windows depending on specified sliding window size N and consider those sliding windows are S_1, S_2 ….

Step 2: Applying appropriate categorical clustering algorithm on S^1 to gain primary 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 data information system of $c^1{}_i$, clustering result C_t for t=1, 2...Where t represents timestamp.

Step 3: out=0

Step 4: For each unlabeled data point $P_j \in S^{t+1}$ sliding window, start with t=1 begin.
Step 5: Begin each and every cluster $c^t{}_i$
Step 6: For all $a \in A_i$
Step 7: Calculate U/IND({a}) using (1)
Step 8: Calculate the rough entropy RE({a}).using formula (2)
Step 9: end
Step 10: Find rough entropy $RE_{p_j}(a)$ using (3)
Step 11 : Fix the threshold value to 0.5 .
Step 12 : End

Step 13: Consider the next object of the sliding window and shift the objects into appropriate clusters based Importance of the each attribute of object.
Step 14: Repeat the procedure
Step 15: If concept drift finds then go to Step1.

5. Experimental Results

This section exhibit the presentation of the projected work on clustering categorical data by a methodical experimental study on the real dataset. In the Section IV.I, the test nature, characteristics are discussed. After that, the developing procedure of clustering outcomes is shown on the real dataset.

5.1 Test Environment and Dataset:

The proposed experiment is done on a Person computer with Intel Corei5 processor with 8 GB memory and the Windows7 professional operating system. In the presented work, the k-modes clustering algorithm is selected to do the primary clustering and reclustering on the datasets. As the k-modes algorithm is reliant on the collection of first cluster centers, we use an initialization method, which was proposed in [7], to obtain first cluster core point prior to performing the k-modes.
The network-intrusion-detection stream from the KDD-CUP'99 [8], is elected for build up the proposed methodology, the data set consist of 23 classes with the class for "normal connection". The subsequent experiments, all 22 attack-types are seen as "attack." We exploit the class label which designate the verification is a normal association or an assault to recognize the drifting concept. In this dataset consist of 494,021 records, and every record contains 42 attributes including class label, such as the period of the link, The number of data bytes passed on from starting place to end and vice versa, the percentile of acquaintances that have "SYN" errors, the number of "root" accesses, etc. Also, 34 attributes are continuous.

5.2 Evaluating scalability

Initially K-modes algorithm is used, after that to find the efficiency of the proposed method we uses synthetic data and that consists of different attributes and objects. There are objects taken into the experiment is from 10,000 to 100,000,the dimensionality varies from 10–50. In Table V the experimental results are stored with comparison Ming Chen method. The calculation of each value in Table VI is the average of 10 times testing.

Table VI: Comparison of Existing Methods

Data Records	Ming-Chen Method	Proposed Method
10,000	0.8524	0.4458
20,000	1.9568	1.1580
30,000	2.1574	1.2547
40,000	2.5896	1.5247
50,000	3.1533	2.1254
100,000	8.3215	4.3475

6. Conclusion

In categorical domain, the difficulty of how to allocate the unlabeled data points into suitable clusters has not been completely discovered in previous research in data mining/clustering. In addition, to that the data is transforming over period of time, clustering this type of data not only reduces the value of clusters and also ignores the potential of users, when usually require recent clustering results. This paper, deals the method based on Rough Entropy similarity 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]. 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]. Michael R. Anderberg," Cluster analysis for applications", Academic Press, 1973.
[5]. Han,J. and Kamber,M. "Data Mining Concepts and Techniques", Morgan Kaufmann, 2001.
[6]. 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.
[7]. Bradley,P.S., Usama Fayyad, and Cory Reina," Scaling clustering algorithms to large databases", Fourth International Conference on Knowledge Discovery and Data Mining, 1998.
[8]. 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.
[9]. 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.
[10]. Fuyuan Cao, Jiye Liang, "A Data Labeling method for clustering categorical data", Elsevier Expert systems with applications, 38(2011), 2381-2385.
[11]. 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.
[12]. Klinkenberg, R.," Using labeled and unlabeled data to learn drifting concepts", IJCAI-01Workshop on Learning from Temporal and Spatial Data, pp. 16-24, 2001.
[13]. Z. Pawlak, "Rough sets ", International journal of computer and information sciences, 11(1982), 341-356.
[14]. 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.
[15]. 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.
[16]. 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.
[17]. Gluck, M.A. and Corter, J.E. "Information Uncertainty and the Utility of Categories", Cognitive Science Society, pp. 283-287, 1985.
[18]. Shannon, C.E, "A Mathematical Theory of Communication," Bell System Technical J., 1948.
[19]. 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.
[20]. 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.
[21]. Xiangjun Li, Fen Rao, "An Rough Entropy Based Approach to Outlier Detection", Journal of Computational Information Systems 8: 24 (2012) 10501-10508.
[22]. Venkateswara Reddy.H, Viswanadha Raju.S," A Threshold for clustering Concept – Drifting Categorical Data", IEEE Computer Society, ICMLC 2011.
[23]. 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.
[24]. Ng, R.T. Jiawei Han "CLARANS: a method for clustering objects for spatial data mining", Knowledge and Data Engineering, IEEE Transactions, 2002.
[25]. 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.
[26]. Huang, Z. and Ng, M.K, "A Fuzzy k-Modes Algorithm for Clustering Categorical Data" IEEE On Fuzzy Systems, 1999.
[27]. Guha,S., Rastogi,R. and Shim, K, "ROCK: A Robust Clustering Algorithm for Categorical Attributes", International Conference On Data Eng. (ICDE), 1999.
[28]. Ganti, V., Gehrke, J. and Ramakrishnan, R, "CACTUS—Clustering Categorical Data Using Summaries," ACM SIGKDD, 1999.
[29]. Vapnik, V.N," The nature of statistical learning theory", Springer,1995.
[30]. Fredrik Farnstrom, James Lewis, and Charles Elkan," Scalability for clustering algorithms revisited", ACM SIGKDD pp.:51–57, 2000.
[31]. Barbara, D., Li, Y. and Couto, J. "COOLCAT: An Entropy-Based Algorithm for Categorical Clustering", ACM International Conf. Information and Knowledge Management (CIKM), 2002.
[32]. Andritsos, P, Tsaparas, P, Miller R.J and Sevcik, K.C."LIMBO: Scalable Clustering of Categorical Data", Extending Database Technology (EDBT), 2004.