Our - NIR : Node Importance Representative for Clustering of Categorical Data

[1]S.Viswanadha Raju, [2]H.Venkateswara Reddy, [3]N.Sudhakar Reddy, [4]G.Sreenivasulu, [5]Dr. KVN Sunitha
[1]Dept of CSE, JNTUH,Hyderabad, India
[2,3]Dept. of CSE, SVCE, Tirupati, India
[4]Dept of CSE, VCE,Hyderabad, India
[5]Dept. of CSE, GNITS, Hyderabad, India

Abstract
The problem of evaluating node importance in clustering has been active research in present days and many methods have been developed. Most of the clustering algorithms deal with general similarity measures. However In real situation most of the cases data changes over time. But clustering this type of data not only decreases the quality of clusters but also disregards the expectation of users, when usually require recent clustering results. In this regard Ming-Syan Chen proposed a method, which is related to calculate the node importance that is very useful in clustering of categorical data, but it has serious deficiency that is bias towards features with many outcomes. In this paper we proposed a new method evaluating of node importance by summarize rules which will be better than the Ming-Syan Chen proposed method by comparing the results.

Keywords
Clustering; Categorical; CNIR, Our-NIR

I. Introduction

Extracting Knowledge from large amount of data is difficult which is known as data mining. Clustering is a collection of similar objects from a given data set and objects in different collection are dissimilar. Most of the algorithms developed for numerical data may be easy, but not in Categorical data [1-2, 13- 14]. It is challenging in categorical domain, where the distance between data points is not defined. It is also not easy to find out the class label of unknown data point in categorical domain. Sampling techniques improve the speed of clustering 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. For example, the buying preferences of customers may change with time, depending on the current day of the week, availability of alternatives, discounting rate etc. Since data evolve with time, the underlying clusters may also change based on time by the data drifting concept [12, 16]. The clustering time-evolving data in the numerical domain [1, 5, 6, 10] has been explored in the previous works, where as in categorical domain not that much. Still it is a challenging problem in the categorical domain.

As a result, our contribution in modifying the frame work which is proposed by Ming-Syan Chen in 2009 [8] utilizes any clustering algorithm to detect the drifting concepts. Here Ming-Syan Chen NIR values are referred to as CNIR. We adopted sliding window technique and initial data (at time t=0) is used in initial clustering. These clusters are represented by using Our-NIR, where each attribute value importance is measured. We find whether the data points in the next sliding window (current sliding window) belongs to appropriate clusters of last clustering results or they are outliers. We call this clustering result as a temporal and compare with last clustering result to drift the data points or not. If the concept drift is not detected to update the Our-NIR otherwise dump attribute value based on importance and then re-clustering using clustering techniques.

The rest of the paper is organized as follows. In section II discussed related work, in section III basic notations and Node representation provided, in section IV new method for node importance representative discussed and also contains results with comparison of CNIR and our method, and finally concluded with section V.

II. Related Work

In this section, we discuss various clustering algorithms on categorical data with cluster representatives and data labeling. We studied many data clustering algorithms with time evolving.
Cluster representative is used to summarize and characterize the clustering result, which is not fully discussed in categorical domain unlike numerical domain. In K-modes which is an extension of K-means algorithm in categorical domain a cluster is represented by 'mode' which is composed by the most frequent attribute value in each attribute domain in that cluster. Although this cluster representative is simple, only use one attribute value in each attribute domain to represent a cluster is questionable. It composed of the attribute values with high co-occurrence. In the statistical categorical clustering algorithms [3,4] such as COOLCAT and LIMBO, 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 algorithm LIMBO, the information bottleneck method is applied to minimize the information lost which resulted from summarizing data points into clusters. Improved the accuracy of clustering concept drift categorical data by modified threshold value [16] which has still a deficiency that is finding the node importance value with purity. However, all of the above categorical clustering algorithms focus on performing clustering on the entire dataset and do not consider the time-evolving trends and also the clustering representatives in these algorithms are not clearly defined.

In this paper, first object of Our-NIR which is based on the idea of representing the clusters by the importance of the attribute values. This representation is more efficient than using the representative points. After thoroughly scanning the literature survey, it is clear that clustering categorical data is un touched many ties due to the complexity involved in it. A time-evolving categorical data is to be clustered within the due course hence clustering data can be viewed as follows: there are a series of categorical data points D is given, where each data point is a vector of q attribute values, i.e., $p_j=(p_j1,p_j2,...,p_jq)$. And $A = \{A_1,A_2 ,..., A_q\}$, where A_a is the ath categorical attribute, $1 \leq a \leq q$. The window size N is to be given so that the data set D is separated into several continuous subsets S_t, where the number of data points in each S_t is N. The superscript number t is the identification number of the sliding window and t is also called time stamp. Here in we consider the

first N data points of data set D this makes the first data slide or the first sliding window S_0. Our intension is to cluster every data slide and relate the clusters of every data slide with previous clusters formed by the previous data slides. Several notations and representations are used in our work to ease the process of presentation.

III. Node Representation

For categorical or mixed data can have several representations. But in our work we can take two sorts of data representations. In first kind of representation every data point present in the sliding window or data slide is divided into distinct points in which every distinct point is considered as the new node and each node has two parts, in this name or the categorical value is placed in the pre-part of the node where as the post-part contains the numerical value of that data point or the node. For example: nodes with attribute name "COMPOSE" which is a categorical part and the number of occurrences in the document '24' is a numerical part. This node is represented as follows:

Node [COMPOSE: 24]

This representation eventually reduces the ambiguity that may prevail among the attributes, as many attributes may have same value. By introducing the categorical part into the node we eliminate the risk of confusion. There is another form of representation of the data in our work. In this second representation we use a data description file that describes the data attributes and with a transitive relation we recognize the data attribute. This is the simplification of the above mentioned representation the only difference is that categorical part is kept in another file. This may look like the numerical representation of the data at an instance, but the value that is used to represent an attribute may be a numerical, binary, or categorical. However in this paper considered categorical data set. This eases the effort that is required. This representation is also useful for the importance of node in the data set used in our work.

IV. Importance of Node

The distribution of Node that is described in above section represents a cluster. As mentioned every node has attribute value, the same value is used to find the distribution of the data points. Hence the importance of the node plays a great role in finding clusters the importance of the node is evaluated with the following rules such as rule 1, rule 2 and rule 3. Here we considered a symbolic representation for the ith node in cluster i is N [i, r], Thenumber of data points in cluster Ci is mi , and k is number of clusters .

A. Rule1 (Probability of node N [i, r])

The probability of node (pi)in the cluster can be calculated as follows:

$$pi = \frac{\left| N_{[i,\,r]} \right|}{m_i}$$

Sliding window -1

A	A	A	X	Y	X
M	M	M	M	M	M
C	D	C	P	P	P

Sliding window-2

A	Y	Y	D	A
K	K	M	M	K
D	P	C	P	P

B. Rule 2 (Frequency of node N [i, r])

The distribution of the node in the clusters is calculated as follows.

$$d(N[r]) = \frac{\left| \sum_{y=1}^{k} p(N[y,r])^2 \right|}{2}$$

Where $p(N[y,r]) = \frac{|N[y,r]|}{\sum_{z=1}^{k} |N[z,r]|}$

C. Rule 3 (Weighted Function)

The importance of node N [i, r]. can be calculated by the product of Rule 1 and Rule 2: $W(c_i, N_{[i,r]}) = p_i * d(N_{[i,r]})$

The weighting function is designed to measure the distribution of the node between clusters based on the information theorem [15].

The weighting function measures the entropy of the node between clusters. Suppose that there is a node that occurs in all clusters uniformly. The node that contains the maximum uncertainty provides less clustering characteristics. Therefore, this node should have a small weight. Moreover, the maximum entropy value of a node between clusters equals. In order to normalize the weighting function from zero to one, the entropy value of the node between clusters is divided by 2 where as in Ming-Syan Chen method divide by logk.

The importance of the node $N_{[i,r]}$ in cluster c_i is measured by multiplying the rule1 and rule 2 i.e., the weighting function $W(ci, N_{[i,r]})$. Note that the range of both the probability of $N_{[i,r]}$ being in c_i and the weighting function $W(c_i, N_{[i,r]})$ is [0, 1], implying that the range of the important value $W(c_i, N_{[i,r]})$ is also in [0, 1].

The new method is related to the idea of conceptual clustering [9], which creates a conceptual structure to represent a concept (cluster) during clustering. However, NIR only analyzes the conceptual structure and does not perform clustering, i.e., there is no objective function such as category utility (CU) [11] in conceptual clustering to lead the clustering procedure. In this aspect our method can provide in better manner for the clustering of data points on time based.

Cluster-1

A	A	A
M	M	M
C	D	C

Cluster-2

X	Y	X
M	M	M
P	P	P

Fig. 1: Sample data points of categorical data and initially cluster performed for the sliding window 1

Example 1: consider the data set in fig 1.cluster c_{11} contains three data points. The node $\{[A_1=A]\}$ occurs three times in $c_{11}$and does not occurs in c_{12}.

The importance of node $\{[A_1=A]\}$ in c_{11} and in c_{21} is calculated as follows. The weight of the node $d(\{A_1=A\}) = ((3/3)2+(0/3)2)/2=0.5$ and therefore an importance of the node $\{A_1=A\}$ in cluster c_{11} is $w(c_1, \{A_1=A\}) = (3/3)*0.5=0.5$ and in cluster c_{21} it is zero. Similarly the remaining nodes as follows : Weight of the node $d(\{A_2=M\})= ((3/5)2+(2/5)2)/2=0.26$ and therefore node

importance in cluster c_{11} is $w(c_1,\{A_2=M\})=(3/3)*0.26=0.26$ and also $w(c_2,\{A_2=M\})=(2/2)*0.26=0.26$, Weight of the node $d(\{A_3=C\})=((2/2)2+(0/2)2)/2=0.5$ and node importance in cluster in cluster c_{11} is $w(c_1,\{A_3=C\})=(2/3)*0.5=0.33$, weight of the node $d(\{A_3=D\})=((1/1)2+(0/1)2)/2=0.5$ and node importance in cluster in cluster c_{11} is $w(c_1,\{A_3=D\})=(1/3)*0.5=0.166$, Finally, the cluster Our-NIR values are represented by a table respectively cluster c_1 in fig. 2 (a) and c2 in fig. 2(b)

Cluster C1		
Node	CNIR	Our NIR
A1=A	1	0.5
A2=M	0.029	0.26
A3=C	0.67	0.33
A3=D	0.33	0.166

Cluster C2		
Node	CNIR	Our NIR
A1=X	0.5	0.25
A1=Y	0.5	0.25
A2=M	0.029	0.26
A3=P	1	0.5

Fig. 2: Node importance values of cluster 1 and cluster 2 from the fig. 1

A. Comparison of Our-NIR and CNIR

Fig 2 shows the importance value of attributes with size of window is 5. This study fixes for the two slides of data points with the time evolving that is from t_1 to t_2. As we said the importance of the node in that way we comparing the node importance values. The node importance values of our method is provide in a different way that increase the purity of the cluster which is impact on the accuracy of clustering. In fig. 3 (a) and (b), we present the importance values of the each system maintained and over the 4 attributes of each cluster with sliding window size of given on importance of attributes of CNIR method maintained sudden drift and where as in our method (Our-NIR)that was not occurred but here it is satisfying the graduality.

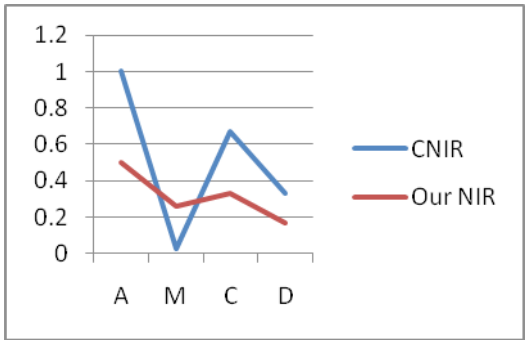

Fig. 3 (a): NIR values of nodes for Cluster C1

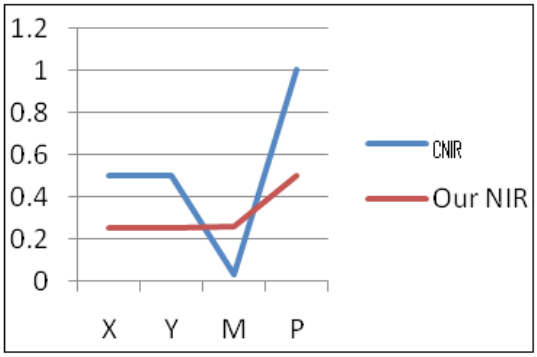

Fig. 3 (b): NIR Values of nodes for Cluster C2

V. Conclusion

In this paper, a frame work proposed by Ming-Syan Chen in 2009 [8] which is modified by new method that is to find node importance. We analyzed by taking same example in this find the differences in the node importance values of attributes in same cluster which plays an important role in clustering. The future work deciding the class label of unclustered data point and therefore the result demonstrates that our method is accurate as said in section 4, than by CNIR and also it improves the performance of precision and recall of DCD.

References

[1] Aggarwal, C.; Han, J.; Wang, J.; Yu, P. "A Framework for Clustering Evolving Data Streams," Proc. 29th Int'l Conf. Very Large Data Bases (VLDB), 2003.

[2] Aggarwal, C.C.; Wolf, J.L.; Yu, P.S.; Procopiuc, C.; Park, J.S. "Fast Algorithms for Projected Clustering," Proc. ACM SIGMOD '99, pp. 61-72, 1999.

[3] Andritsos, P.; Tsaparas, P.; Miller, R.J.; Sevcik, K.C. "Limbo: Scalable Clustering of Categorical Data," Proc. Ninth Int'l Conf. Extending Database Technology (EDBT), 2004.

[4] Barbará, D.; Li, Y.; Couto, J. "Coolcat: An Entropy-Based Algorithm for Categorical Clustering," Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM), 2002.

[5] Cao, F.; Ester, M.; Qian, W.;Zhou, A. "Density-Based Clustering over an Evolving Data Stream with Noise," Proc. Sixth SIAM Int'l Conf. Data Mining (SDM), 2006.

[6] Chakrabarti, D.; Kumar, R.; Tomkins, A. "Evolutionary Clustering," Proc. ACM SIGKDD '06, pp. 554-560, 2006.

[7] Chen, H.L.; Chuang, K.T.; Chen, M.S. "Labeling Unclustered Categorical Data into Clusters Based on the Important Attribute Values," Proc. Fifth IEEE Int'l Conf. Data Mining (ICDM), 2005.

[8] Chen, H.L.; Chen, M.S.; Chen Lin, SU. "Frame work for clustering Concept –Drifting categorical data," IEEE Transaction Knowledge and Data Engineering Vol. 21 no 5, 2009.

[9] Fisher, D.H. "Knowledge Acquisition via Incremental Conceptual Clustering," Machine Learning, 1987.

[10] Gaber, M.M.;Yu, P.S. "Detection and Classification of Changes in Evolving Data Streams," International Journal .Information Technology and Decision Making, Vol. 5 no 4, 2006.

[11] Gluck, M.A.; Corter, J.E. "Information Uncertainty and the Utility of Categories," Proc. Seventh Ann. Conf. Cognitive Science Soc., pp. 283-287, 1985.

[12] Hulton, G.; Spencer. "Mining Time-Changing Data Streams" Proc. ACM SIGKDD 2001.

[13] Jain, A.K.; Murthy, M.N.; Flyn, P.J. "Data Clustering: A Review," ACM Computing Survey,1999.

[14] Narsoui,O.; Rojas, C. "Robust Clustering for Tracking Noisy Evolving Data Streams" SIAM Int. Conference Data Mining, 2006.

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

[16] Viswanadha Raju, S.; Venkateswara Reddy,H.; Sudhakar Reddy, N. "A Threshold for Clustering Concept – Drifting Categorical Data", IEEE Computer Society, ICMLC 2011.