Finding Relationship between the POur-NIR Cluster Results

N. Sudhakar Reddy and K.V.N. Sunitha

Professor in CSE S.V.College of Engineering Tirupati, India
Professor in CSE G.Nayanamma Institute of Technology & Science Hyderabad, India

Abstract
Categorical data clustering is an interesting challenge for researchers in the data mining and machine learning, because of many practical aspects associated with efficient processing and concepts are often not stable but change with time. Typical examples of this are weather prediction rules and customer's preferences, intrusion detection in a network traffic stream. Another example is the case of text data points, such as that occurring in Twitter/search engines. In this paper we propose a generalized framework that detects drifting concepts and try to show the evolving clustering results in the categorical domain. This scheme is based on the cosine measure that analyzes relationship between clustering results at different time stamps using POur-NIR method.

Keywords-
Clustering, Weather Prediction, Drifting, POur-NIR method

1. 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, 11, 12]. 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 [10, 15]. The clustering time-evolving data in the numerical domain [1, 5, 6, 9] 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. We adopted sliding window technique and initial data (at time t=0) is used in initial clustering. These clusters are represented by using Chen NIR and POur-NIR [8, 19], 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 POur-NIR otherwise dump attribute value based on importance and then reclustering using clustering techniques [19]. In this paper mainly concentrating on the inter-similarity of adjacent clusters from time to time based similarity measure that is easy to find the drifts are occurred or not

The rest of the paper is organized as follows. In section 2 discussed related works, in section 3 vector representation provided, in section 4 cosine measure for relation analysis among the clusters discussed and also contains results with comparison of Ming-Syan Chen method and Pour-NIR method and finally concluded with section 5.

2. 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.

Manuscript received June 5, 2011
Manuscript revised June 20, 2011

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.

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.

The main reason is that in concept drifting scenarios, geometrically close items in the conventional vector space might belong to different classes. This is because of a concept change (drift) that occurred at some time point.

Our previous work [19, 20] addresses the node importance in the categorical data with the help of sliding window. That is new approach to the best of our knowledge that proposes these advanced techniques for concept drift detection and clustering of data points.

After scanning the literature, 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., pj=(pj1,pj2,...,pjq). And A = {A1,A2 ,..., Aq}, where Aa is the ath categorical attribute, 1 ≤ a ≤ q. The window size N is to be given so that the data set D is separated into several continuous subsets St, where the number of data points in each St is N shown in figure 1. 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 S1or S1. The 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. In the previous work we considered the sample data set for the clustering of concept drift categorical data in that paper initially clustering done by standard algorithm that result shown in figure 1 and finally concluded with the updated POur-NIR results respect to sliding window and clusters as shown in figure 2[20] .Based on the relationship analysis, the evolving clusters will provide clues for us to catch the time evolving trends in the data set. This can achieve by introducing vector model and cosine measure, the similarity measure is most efficient for the vector representation.

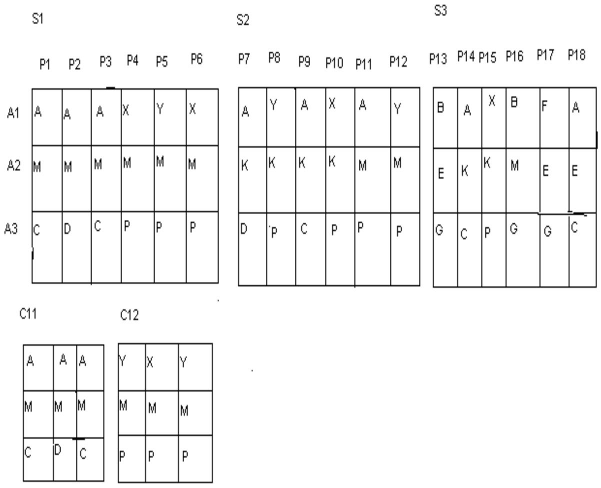
S1
S2
S3
P1 P2 P3 P4 P5 P6
P7 P8 P9 P10 P11 P12
P13 P14 P15 P16 P17 P18
A1 A A A X Y X | A Y A X A Y | B A X B F A
A2 M M M M M M | K K K K M M | E K K M E E
A3 C D C P P P | D P C P P P | G C P G G C
C11
C12
A A A | Y X Y
M M M | M M M
C D C | P P P

Fig. 1. Data set with sliding window size 6 where the initial clustering is performed

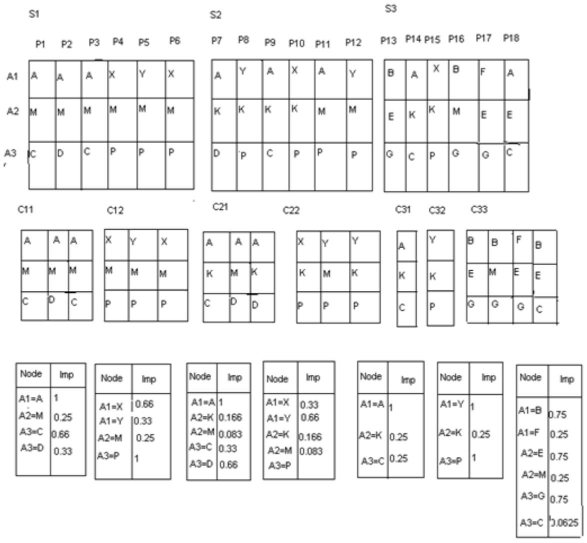

Fig 2: Final clustering results as per the data set of fig 1 and output POur-NIR Results

3. Vector representation

The vector model is a view of the representative to contain the domain of nodes. The size of vector is based on the total number of nodes in entire data set. A cluster in this space is a vector, and an each index of vector is the value of importance by POur-NIR method in that node domain. Based on the node vector representation, the node POur-NIR Vector of cluster Ci is shown as follows:

$$C_{i\,=}\ \ w_i(I_1),w_i(I_2),\ldots\ldots w_i(I_i),\ldots\ldots\ldots\ldots\ldots\ldots w_i(I_z)),$$

Where

$W_i(I_r)$=o, if I_r does not occur in C_i,
$Wi(I_i)=w(C_i,i_{ir})$, if I_r occur in C_i.

This model can work on all nodes in entire data set. The calculations needed for the vector model are:

1. The weight of each node across the entire data set needs to be calculated based on sliding window data set and POur-NIR method [20]. This gives how important the node is in the sliding window of data set.
2. The weight of every node within a given sliding window needs to be calculated for all slidings. This obtains how important the node is within a single sliding window.
3. Every two adjacent vectors of the sliding window clusters are compared

The value in the vector Ci on each node domain is the POur-NIR value of this node in cluster Ci, i.e., W (ci, N [i, r]). If the node does not occur in cluster Ci, the value in the vector Ci on this node domain is zero. Here contains all distinct nodes that occur in the entire data set, not just in cluster Ci based on the domain of attribute values. Therefore, the dimensions of all the vectors Ci are the same.

	A	B	C	D	E	F	G	K	M	P	X	Y
$\overline{C11}$	1	0	0.66	0.33	0	0	0	0	0.25	0	0	0
$\overline{C12}$	0	0	0	0	0	0	0	0	0.25	1	0.66	0.33
$\overline{C21}$	1	0	0.33	0.66	0	0	0	0.166	0.083	0	0	0
$\overline{C22}$	0	0	0	0	0	0	0	0.166	0.083	1	0.33	0.66
$\overline{C31}$	1	0	0.25	0	0	0	0	0.25	0	0	0	0
$\overline{C32}$	0	0	0	0	0	0	0	0.25	0	1	0	1
$\overline{C33}$	0	0.75	0.0625	0	0.75	0.25	0.75	0	0.25	0	0	0

Fig 3: POur-NIR Vectors $\overline{C1}$, $\overline{C2}$ and $\overline{C3}$ of the clustering results C1, C2 and C3 In fig 2

Example: In the example data set shown in fig 1, in that figure there are totally 12 distinct nodes in the entire data set and the POur-NIR results of C11 and C12 are shown in fig 3 based on this figure 2 the vector space defined as said above in this section the vector of cluster C11 and similarly for the remaining clusters as shown in fig 3

The clusters C_i and C_j are represented by the POur-NIR vectors $\overline{C_i}$and$\overline{C_j}$. We studied several similarity measures for the finding of similarity of clusters, finally concluded among them the cosine measure is often used to compare documents in text mining. In addition, it is used to measure cohesion within clusters in the field of Data Mining.

4. Cosine Measure

The cosine treats both vectors as unit vectors by normalizing them, it calculates the cosine of the angle between the two vectors. It does provide an accurate measure of similarity but with no regard to magnitude. But magnitude is an important factor while considering similarity. It is popular measure of similar in the vector representation [14] . The cosine measure between vectors $\overline{C_i}$and$\overline{C_j}$ are Ai and Bi respectively is calculated as the shown equation below.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Consider the clustering results C11 and C22 in fig 3. The POur-NIR vectors of the clustering results C11 and C12 are shown in fig 4. The similarity between vectors C11 and C21 is 0.8933 and similarly calculated for the other clusters.

	$\overline{C11}$	$\overline{C12}$	$\overline{C21}$	$\overline{C22}$	$\overline{C31}$	$\overline{C32}$	$\overline{C33}$
$\overline{C11}$			0.89	0.0129			
$\overline{C12}$			0.0129	0.906			
$\overline{C21}$					0.843	0.022	0.024
$\overline{C22}$					0.0311	0.943	0.012

Fig 4: cosine similarity table between the clustering results c1 and c2 and between the c2 and c3 by POur-NIR results in fig 3.

	C11	C12	C21	C22	C31	C32	C3
C11			0.9296	0			
C12			0	0.9296			
C21					0.178	0	0
C22					0	0.186	0

Fig 5: Cosine similarity table between the clustering results c1 and c2 and between the c2 and c3 By CNIR results

In figure 4 , the similarity of each pair of adjacent clustering results, where t^b is the time stamp that different concepts happens, is measured by the cosine measure. Based on this measure, it provides for us to catch the time-evolving trend in the data set and also it could help for how to link the clusters at different time stamps.

Comparison of CNIR and POur-NIR

The cosine similarity of each pair clustering results of both the CNIR and POur-NIR shown in figure 5. As per the observation in that figure some of the inter-clusters may get zero similarity by CNIR where as in POur-NIR getting different. That shows the relationship between the clustering results at different time stamps. At same time when we are looking into the sample data set in figure 1 there it could be different with the CNIR result that means POur-NIR showing the better performance.

5. Conclusion

In this paper, a frame work proposed by Ming-Syan Chen Node Importance Representative (CNIR) in 2009[8] is modified by new method that is POur-NIR to find node importance [20]. It is analyzed, finding the differences in the node importance values of attributes [19] in same cluster plays an important role in reclustering. The representatives of the clusters help improving the cluster accuracy and purity and hence the POur-NIR method performs better than the CNIR method [8].The pairing of each adjacent clusters similarity is based on POur-NIR method and is better than the CNIR in terms of cluster distribution. The future work improves the performance of precision and recall of DCD by introducing the leaders-subleaders algorithm for reclustering.

References

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

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

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

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

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

[6] D. Chakrabarti, R. Kumar, and A. Tomkins, “Evolutionary Clustering,”*Proc. ACM SIGKDD” 2006,* pp. 554-560..

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

[8] H.-L. Chen, 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.

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

[10] Fan, W. Systematic data selection to mine concept-drifting data streams. in Tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining. 2004. Seattle, WA, USA: ACM Press: p. 128-137.

[11] MM Gaber and PS Yu “Detection and Classification of Changes in Evolving Data Streams,” *International .Journal .Information Technology and Decision Making,* v5 no 4, 2006.

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

[13] G Hulton and Spencer, “Mining Time-Changing Data Streams” *Proc. ACM SIGKDD,* 2001.

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

[15] Klinkenberg, R., Learning Drifting Concepts: Example Selection vs. Exam- ple Weighting Intelligent Data Analysis, Special Issue on Incremental Learn- ing Systems Capable of Dealing with Concept Drift, 2004. 8(3): p. 281-200.

[16] O.Narsoui and C.Rojas,“Robust Clustering for Tracking Noisy Evolving Data Streams” SIAM I*nt. Conference Data Mining ,* 2006.

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

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

[19] S.Viswanadha Raju,H.Venkateswara Reddy and N.Sudhakar Reddy ” POur-NIR:Node Importance Representation of Clustering Categorical Data”, IJCSIS, May 2011.

[20] S.Viswanadha Raju, N.Sudhakar Reddy and H.Venkateswara Reddy,” Clustering of Concept Drift Categorical Data using POur-NIR Method, IJEE (accepted).

Dr K.V.N.Sunitha did her B.Tech ECE from Nagarjuna University, M.Tech Computer Science from REC Warangal and Ph.D from JNTU in 2006. She has 21 years of Teaching Experience, worked at various engineering colleges. She has been working as HOD, CSE Dept in G.Narayanamma Institute of Technology and Science from the inception of the CSE Dept since 2001. She is a recipient of Academic Excellence award by GNITS in 2004. She has received “Best computer engineering Teacher award “by ISTE (Indian society for Technical education) in Feb 2008.Her biography is included in Marquis who is who in the world ,28th edition 2011.She has published more than 55 papers in International & National Journals and conferences. She is guiding 12 PhDs. She is a reviewer for many International Journals . She is fellow of Institute of engineers, Sr member for International association CSIT, and life member of many technical associations like CSI, IEEE, and ACM. She has authored two books, “Programming in UNIX and Compiler design”- BS Publications & “Formal Languages & Automata Theory” -Tata Mc Graw Hill. Now she is currently doing research in the areas Data mining, Natural language processing, Speech Processing, Network & web Security.

Mr. N Sudhakar Reddy completed his B.E from College of Engineering, GITAM, Visakhapatnam and M.Tech(Computer Science) from College of Engineering JNTU, Anantapur. At present he is pursuing Ph.D in Data Mining from JNTUA. He has 13 years of Teaching Experience, worked at various engineering colleges. Now he is working as Head of the Department, Computer Science and Engineering at S V College of Engineering, Tirupati. He is currently doing research in the areas of Data mining, Pattern Recognition & Image Processing.