iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 Issue: I Month of publication: January 2018

DOI: http://doi.org/10.22214/ijraset.2018.1002

Survey on Clustering Over Categorical Streaming Data

Ms. P. B. Bachhav[1], Prof. S. S. Banait[2]
([1]M.E. Student [2]Asistant Professor) Department of computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik Savitribai Phule Pune University, Maharashtra, India.

***Abstract**: Clustering can be done for static as well as dynamic dataset. Clustering is performed on the basis of underlying data structure and data distribution. The data distribution varies with respect to time. Hence there need to be a smooth change in underlying cluster model with respect to time. The clustering techniques of numerical dataset are different than Categorical dataset. This works aim to study various clustering techniques for streaming data with respect to the numerical and Categorical dataset.*

Keywords: Clustering, data streaming, data drift, categorical data, optimization model

I. INTRODUCTION

Clustering of data or cluster analysis is emerging field in exploratory data analysis. This technique is applied in variety of domains of engineering as scientific disciplines such as biology, medicines, marketing, computer vision, etc. The clustering technique extracts the underlying data structure based on data distribution and creates individual groups or hierarchy of groups. The group includes similar objects in one set. Unlike other statistical methods these techniques do not require any pre-assumptions or training to describe underlying data structure. This is unsupervised method. Cluster analysis and its various methods are discussed in[2].

In recent work, clustering of continuous streaming data is important aspect. In variety of domain continuous data is get generated such as stock market, credit/debit card transaction, online selling web clicks, etc. The conventional models are unable to generate appropriate clusters over streaming data as streaming data appears continuously. In traditional techniques all data is get collected first and then clustering is performed. This is not the case of continuous data stream. The Underlying data structure varies with upcoming new stream entries. Generally in real time applications data varies with respect to time.

The existing techniques can be applied to such streaming data. And upcoming new points can be merged in existing clusters. But this clustering solution is not meet several challenges.

Based on this analysis, clustering is again partitioned in 2 categories: static Data clustering and dynamic data clustering.

In Static data clustering whole dataset is available in the beginning phase. The clusters are generated based on the analysis of complete daa distribution. This technique is based on the assumption that cluster structure do not changes with respect to time.

In dynamic data clustering , continuous data stream is available and nature if data changes with time. As data model changes with respect to time underlying cluster structure also changes. This is called as data drifting.

Paper is organized as follows: Section I introduction about clustering of continuous streaming data. Section II gives the literature review. Section III concludes the paper.

II. LITERATURE WORK

For dynamic data clustering conversional clustering approaches are not suitable. In[5][6][7][8] various clustering problem of streaming data are discussed.

CluStream clustering algorithm[3] is proposed for data stream. This technique tries to cluster whole data points in the stream at ones irrespective to the knowledge of Time-Changing Data Streams. For evolving environment it applies clustering after certain timestamp. This technique follow the clustering procedure in 2 phases : online and offline. In online phase data collection task is performed whereas in offline phase data clustering is performed.

C. Aggarwal, J. Han, J. Wang, and P. Yu,[5] faced problem by one pass techniques for stream data clustering is discussed. The one pass technique faces the cluster scalability issue. The cluster quality degrades as per the evolving data.

Density based clustering technique [6] is proposed for streaming data. It follows the constraint that, unlike existing strategy there should not be any limit on cluster count for dynamic data clustering. Along with cluster count constraint, limited memory, arbitrary cluster shape and outlier detection over streaming data are also applied. DenStream algorithm is proposed in[4]. core-micro-cluster and outlier micro-cluster are generated to distinguish the clusters and outliers. Pruning strategy is applied to satisfy the memory constraint.

Temporal smoothness[7] is introduced to measure cluster quality. This technique works on the principle as: good clustering should be generated to fit the data in appropriate clusters and there should not be a dramatic change in clustering over evolving data. Two frameworks are proposed in this paper for evolutionary spectral clustering.

Three algorithms are proposed : a drifting-concept detection, data-labeling and cluster-relationship-analysis.[8] These algorithms uses sliding window technique. Based on sliding window and distance between two concepts data drifts are detected. The concept is nothing but a data arrived at one sliding window event.

Clustering over categorical data [9] is proposed along with drifting concept. Maximal Resemblance Data Labeling- MARDL is proposed in this paper. N-Nodeset Importance Representative –NNR technique is used to represent the cluster characteristics, members in cluster and outlier points. This technique also applies Drifting Concept Detection-DCD algorithm to compare cluster results for last generated cluster and temporal current clustering result. Using DCD data drifting is identified over time evolving data. The techniques proposed in[9][10] do not optimize the clusters due to lack of optimization and validity criteria. The second important aspect is there is lack of relevance between clustering and drifting strategies.

A framework is proposed named as: clustering Over Multiple Evolving streams by correlations.[10] This technique handles multiple data streams at a time and manages there distribution based correlation function.

The above techniques are applicable only for numerical data streams. But data stream may contain Boolean or categorical data. The numerical calculations used in clustering technique such as mean, correlation, deviation, etc cannot be directly applied to the categorical data streams. The categorical data imposes several difficulties in such calculations and hence numerical clustering techniques cannot be directly applied to the categorical datasets.

A. Cluster Optimization For Static Categorical Dataset

Following are 3 objective functions to optimize cluster containing categorical value dataset.

1) *K-modes Objective Function:* This is the extension of K-means algorithm proposed in[3]. It replaces the cluster mean calculation with cluster mode. Apart from distance function it applies new dissimilarity function. For clustering occurrence count of categorical data attribute is considered.
2) *The Category Utility Function:* This function is used to measure the 'category goodness'. This technique evaluates the probability of same categorical attribute value belonging to the same cluster as well as different categorical attribute value belonging to the different cluster. This is the attribute value scoring function stated in [1].
3) *Information Entropy Function:* This function is applied to the attribute of categorical dataset to measure how much information is retrieved from it. It calculates the uncertainty information. COOLCAT algorithm is proposed in [4] based on an incremental heuristic technique and entropy function. Generalized objective function is proposed [13], [14]. This technique implements 3 objective functions and analyses differences and generality. It stated that the category utility function and Information entropy function generate the equivalent results. The optimal results generated using .k-modes objective functions are better than the category utility function. The relationship between within-cluster and between-cluster information [14] is discussed by considering the validity functions.

MKM_NOF and MKM_NDM, algorithm are proposed in [13]. These algorithms use weighted cluster prototypes. These algorithms outperforms with respect to k-modes algorithms.

k-modes objective function based solution is proposed in [12]. This technique evaluates the density of categorical dat. It is calculated based on the distance between two objects. Using density evaluation and distance calculation initial cluster centers are identified.

Simultaneous procedure is proposed in[11] to identify good initial partitions i.e. initial cluster centers and candidates for number of clusters. This method generates better results for initial cluster generation for categorical streaming data.

Cluster validity function in [15] is used as a objective function to evaluate cluster over categorical data streams. An iterative algorithm is proposed to find optimized cluster after every incoming data stream points. This mechanism observes the data changes and evolution in trend. Based on the trends data drift are identified and clusters are refined accordingly.

III. CONCLUSION

A bulk amount of data is generated in various real life applications. Some application generates continuous streaming data. By applying Traditional clustering algorithm to the DataStream do not generates the appropriate results. To resolve this problem dynamic clustering algorithm are proposed. The underlying structure of data changes with respect to time called as data drift. There is need to detect change is situation over evolving data to upgrade new cluster model.

Lot of techniques are applicable only for numerical data. For categorical dataset different techniques are proposed as due to the limitations in numerical calculation over categorical data. Some techniques are proposed for categorical data clustering with data drift analysis. There is need to develop a system that support hybrid data clustering i.e. data containing numerical as well as the categorical data attributes.

IV. ACKNOWLEDGMENT

Authors would like to thanks Prof. Dr. K. N. Nandurkar, Principal and Prof. Dr. S. S. Sane, Head of Department of Computer Engineering, K.K.W.I.E.E.R., Nashik for their kind support and suggestions. We would also like to extend our sincere thanks to all the faculty members of the department of computer engineering and colleagues for their help.

REFERENCES

[1] M. A. Gluck and J. E. Corter, “Information uncertainty and the utility categories,” in Proceedings of the Seventh Annual Conference of Cognitive Science Society, 1985, pp. 283–287.

[2] Z. Huang, “A fast clustering algorithm to cluster very large categorical data sets in data mining,” in Proceedings of SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery, 1997, pp. 1–8.

[3] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.

[4] Barbara, Y. Li, and J. Couto, “Coolcat: An entropy-based algorithm for categorical clustering,” in Proceedings of the eleventh international conference on Information and knowledge management, 2002, pp. 582–589.

[5] C. Aggarwal, J. Han, J. Wang, and P. Yu, “A framework for clustering evolving data streams,” in in Proc. Very Large Data Bases Conf., 2003, pp. 81–92.

[6] Cao, M. Ester, Q. Qian, and A. Zhou, “Density-based clustering over an evolving data streams with noise,” in in Proc. SIAM Conf., 2006, pp. 328–339.

[7] Y. Chi, X.-D. Song, D.-Y. Zhou, K. Hino, and B. Tseng, “Evolutionary spectral clustering by incorporating temporal smoothness,” IEEE Trans. Knowledge and Data Engineering, vol. 18, no. 9, pp. 1166–1180, 2006.

[8] M. Yeh, B. Dai, and M. Chen, “Clustering over multiple evolving streams by events and correlations,” IEEE Trans. Knowledge and Data Engineering, vol. 19, no. 10, pp. 1349–1362, 2007.

[9] H. Chen, M. Chen, and S. Lin, “Catching the trend: A framework for clustering concept-drifting categorical data,” IEEE Trans. Knowledge and Data Engineering, vol. 21, no. 5, pp. 652–665, 2009.

[10] Cao, J. Liang, and L. Bai, “A framework for clustering categorical time-evolving data,” IEEE Trans. Fuzzy Systems, vol. 18, no. 5, pp. 872–885, 2010.

[11] L. Bai, J. Liang, and C. Dang, “An initialization method simultaneously find initial cluster centers and the number of clusters for clustering categorical data,” Knowledge-Based Systems, vol. 24, no. 6, pp. 785–795, 2011.

[12] L. Bai, D. C. Liang, J.Y., and F. Cao, “A cluster centers initialization method for clustering categorical data,” Expert Systems with Applications, vol. 39, no. 9, pp. 8022–8029, 2012.

[13] L. Bai, J. Liang, C. Dang, and F. Cao, “The impact of cluster representatives on the convergence of the k-modes type clustering,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, no. 6, pp. 1509–1522, 2013.

[14] L. Bai and J. Liang, “Cluster validity functions for categorical data: A solution-space perspective,” Data Mining and Knowledge Discovery, no. doi: 10.1007/s10618-014-0387-5, 2014.

[15] Liang Bai, Xueqi Cheng, Jiye Liang and Huawei Shen, "An Optimization Model for Clustering Categorical Data Streams with Drifting Concepts," IEEE Transactions on Knowledge and Data Engineering Vol. 28, Issue: 11,pp. 2871 - 2883, Nov. 2016