



ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijareeie.com

Vol. 7, Issue 2, February 2018

Appraise on Clustering Methods of Clustering Data Streams Based on Shared Density between Micro-Clusters and Macro Clusters

Prof. Dr. K. Naveen Kumar M.Tech., Ph.D, Prof. Dr. R. Prakash Rao M.Tech., Ph.D, Nagajyothi M.Tech,
B. Ramesh M.Tech

Professor, Department of Electronics & Communication Engineering, St. PETER'S ENGINEERING COLLEGE,
Dhoolapally, Hyderabad, India

Professor, Department of Electronics & Communication Engineering, St. PETER'S ENGINEERING COLLEGE,
Dhoolapally, Hyderabad, India

Asst. Professor, Department of Electronics & Communication Engineering, St. PETER'S ENGINEERING COLLEGE,
Dhoolapally, Hyderabad, India

Asst. Professor, Department of Electronics & Communication Engineering, St. PETER'S ENGINEERING COLLEGE,
Dhoolapally, Hyderabad, India

ABSTRACT: As increasingly applications produce streaming data, clustering information streams has come to be an critical method for statistics and statistics engineering. A normal technique is to summarize the records motion in actual-time with an internet way right proper right proper right into a big quantity of so known as micro-clusters. Micro-clusters constitute community density estimates via way of way of using aggregating the information of many facts elements in a defined place. On call for, a (modified) traditional clustering set of pointers is finished in a second offline step to re-cluster the micro clusters into large very last clusters. For re-clustering, the facilities of the micro-clusters are used as pseudo elements with the density estimates used as their weights. However, records approximately density inside the location amongst micro-clusters isn't always preserved in the on line approach and re-clustering is based totally in fact in reality without a doubt totally on probable defective assumptions about the distribution of facts inner and amongst micro-clusters (e.g., uniform or Gaussian). This paper describes DBSTREAM, the number one micro-cluster-based completely simply absolutely in reality definitely on-line clustering element that explicitly captures the density amongst micro-clusters thru a shared density graph. The density statistics in this graph is then exploited for re-clustering primarily based definitely really genuinely in reality truly mostly on actual density among adjacent micro-clusters. We speak the gap and time complexity of preserving the shared density graph. Experiments on a excellent form of artificial and real facts devices highlight that the use of shared density improves clustering amazing over in truth considered in reality considered considered one in every of a kind famous data go with the go with the go with the flow clustering strategies which require the appearance of a larger shape of smaller micro clusters to gain similar outcomes.

KEYWORDS: Data mining, Density based clustering, Micro cells, data stream clustering Entropy and Divergence.

I. THE FOUNDATIONS OF DATA MINING

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining



International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijareeie.com

Vol. 7, Issue 2, February 2018

is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996.1 In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

II. THE SCOPE OF DATA MINING

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviors.** Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
- **Automated discovery of previously unknown patterns.** Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

The most commonly used techniques in data mining are:

- **Artificial neural networks:** Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Decision trees:** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .
- **Genetic algorithms:** Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbor method:** A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where $k \geq 1$). Sometimes called the k-nearest neighbor technique.
- **Rule induction:** The extraction of useful if-then rules from data based on statistical significance.

III. CLUSTERING METHODS

Clustering methods can be classified into the following categories –

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method



International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijareeie.com

Vol. 7, Issue 2, February 2018

- Model-Based Method
- Constraint-based Method

Partitioning Method

Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups, which satisfy the following requirements –

- Each group contains at least one object.
- Each object must belong to exactly one group.

Points to remember –

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here –

- Agglomerative Approach
- Divisive Approach

Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering

Here are the two approaches that are used to improve the quality of hierarchical clustering –

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantage

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.



ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijareeie.com

Vol. 7, Issue 2, February 2018

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

REFERENCES

- [1] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams," in Proceedings of the ACM Symposium on Foundations of Computer Science, 12-14 Nov. 2000, pp. 359–366.
- [2] C. Aggarwal, Data Streams: Models and Algorithms, ser. Advances in Database Systems, Springer, Ed., 2007.
- [3] J. Gama, Knowledge Discovery from Data Streams, 1st ed. Chapman & Hall/CRC, 2010.
- [4] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, and J. a. Gama, "Data stream clustering: A survey," ACM Computing Surveys, vol. 46, no. 1, pp. 13:1–13:31, Jul. 2013.
- [5] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proceedings of the International Conference on Very Large Data Bases (VLDB '03), 2003, pp. 81–92.
- [6] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in Proceedings of the 2006 SIAM International Conference on Data Mining. SIAM, 2006, pp. 328–339.
- [7] Y. Chen and L. Tu, "Density-based clustering for real-time stream data," in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2007, pp. 133–142.
- [8] L. Wan, W. K. Ng, X. H. Dang, P. S. Yu, and K. Zhang, "Densitybased clustering of data streams at multiple resolutions," ACM Transactions on Knowledge Discovery from Data, vol. 3, no. 3, pp. 1–28, 2009.
- [9] L. Tu and Y. Chen, "Stream data clustering based on grid density and attraction," ACM Transactions on Knowledge Discovery from Data, vol. 3, no. 3, pp. 1–27, 2009.
- [10] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the ACM .