ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Insight into Dynamic Granular Clustering

G.Dhivya
Assistant Professor

Department of Computer Applications
Dr Umayal Ramanathan College for Women, Karaikudi
dhivyagangadharan@gmail.com

Abstract - **Big data is the term for a collection of datasets so huge and complex that it becomes difficult to be processed using on-hand theoretical models and technique tools. Cluster analysis method can detect data in different granularity levels, and the clustering process is suitable for most of practical problems. This study offers an introduction to the concepts of information granules, elaborates on the main formal frameworks of Granular Computing (GrC). It is a domain of science aiming at modelling computations and reasoning that deals with imprecision and incompleteness of information. This paper outlines the current status of GrC with dynamic granular clustering combining with the particle swarm optimization and provides the general overview of the process of building granular solutions to challenges posed by various real-life problems involving granularity.**

Keywords - Clustering, Granular Computing, K means, PSO (Particle Swarm Optimization)

I. INTRODUCTION

With the rapid development of the information society, the information need to be processed is increasing; a high velocity of incoming data requires strategies to dynamically adapt the analytic system to cope with the changing behaviour of the respective data structures. Clustering is the process of dividing data to different classes or clusters, so objects in a cluster are very similar, and objects between different clusters have great dissimilarity. There are mainly two methods for clustering: hierarchical clustering and partitioned clustering.

The k-means algorithm is most widely used well known unsupervised partitioning method for data clustering and it coordinates local search. K-means clustering method grouped data based on their closeness to each other according to Euclidean distance.

Particle swarm optimization (PSO) has shown to be an efficient, robust and simple optimization algorithm. It adopts Population - based global search strategy, coordinates global search and local search through the inertia weight, to ensure optimal solution with greater the probability, moreover to overcome the shortcomings of local optimum of gradient descent method.

Granular computing is a paradigm of computing with information granules. These granules can be regarded as collections of objects that exhibit some similarity in terms of their properties or functional appearance. Granulation of information is a suitable way of abstraction that helps solve problems in a hierarchical fashion as well as convert the original problem in manageable subtasks.

Big Data is clustering. The goal of clustering is to group similar objects into a cluster while dissimilar objects should be separated by assigning them to different clusters. Putting these three keywords, dynamic, granular computing, and clustering, together we get dynamic granular clustering. The objective of the paper is to address these three keywords holistically by developing a framework for dynamic granular clustering: DCC—Dynamic Clustering Cube.

II. DYNAMIC DATA

Data to be analyzed can be dynamic in different ways, e.g., with respect to location or time, etc. Dynamic data as any kind of data that take only time-dependent aspects into account and refrain from any other dynamic phenomena, like geographical movements and others.

We can distinguish two cases regarding time-dependent aspects: (1) objects whose feature vectors contain just values at a certain moment of time, i.e., snapshots of feature values and (2) objects that contain functions of feature values over time, i.e., feature trajectories.

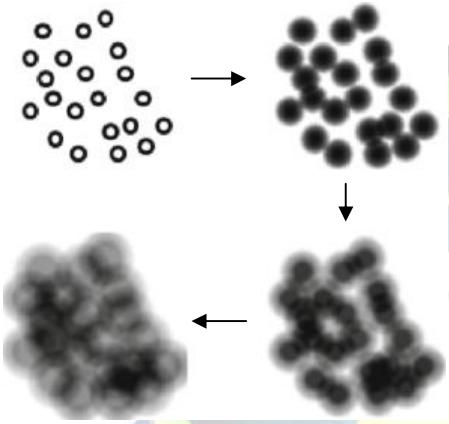
Another important issue is whether observations can be identified over time or not. If objects are identifiable over time, their respective profile constitutes dynamic data, e.g., customers' buying behaviour over time. In the opposite case, i.e., objects are not identifiable, dynamic data provides information on the behaviour of the entire set of analyzed

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

objects, e.g., changing buying behaviour of all customers from a customer base.

II. GRANULAR COMPUTING

A granule may be interpreted as one of the numerous small particles forming a larger unit. Granular computing is the use of groups, classes, or clusters of elements called granules. The process of formation and representation of granules is called granulation. The below fig represents the granulation of the objects.



Granular computing (GrC) is a domain of science aiming at modelling computations and reasoning that deals with imprecision and incompleteness of information. Computations in GrC are performed on granules that are obtained as a result of information granulation.

A) Concept granulation

The origins of the *granular computing* ideology are to be found in the rough sets and fuzzy sets

Fuzzy set theory

Fuzzy set theory starts with the definition of membership function and granulates the features; thereby producing the fuzzy granulation of feature space. The fuzziness in granules and their values characterize the ways in which human concepts of granulation are formed, organized and manipulated.

The number of concepts formed through fuzzy granulation determines the corresponding granulation being relatively fine or coarse, and choice of the number is an application specific optimization problem.

Rough set theory

The rough set theory provides an effective model to acquire knowledge in an information system with upper approximation and lower approximation as its core concepts and in making decisions according to the definition of indistinguishibility (indiscernibility) relation and attribute reduces.

The Rough set approach (RS) can be used to granulate a set of objects into information granules (IGs).

B) Granular computing: Components, characteristics and features

GrC by its name is not new, but the popularity in its use for various domains has gained recently. It is a computing paradigm of information processing that works with the process of information granulation/abstraction. In processing large-scale information, GrC plays an important role that finds simple approximate solution which is cost effective and provides improved description of real world intelligent systems.

Broadly, the prime motivation in the development of GrC-based methodologies is three-fold. These are: (i) GrC does not go for an excessive precision of solution, (ii) the structural representation of the problem in the work-flow of GrC makes the solution process more efficient, and (iii) the computing method provides transparency in the information processing steps.

The components of GrC that drive the complete process are: granules, granulation, granular relationship and computing with granules. A granule is considered as a building block, which plays a significant role in the process of GrC.

Granulation is performed in both ways, such as integrating and dispersing the granular structure. Integration involves the process of developing larger and/or higher level granules with smaller and/or lower level granules, whereas dispersion involves the process of decomposing larger and/or granules into smaller and/or lower level granules. These processes of integration and dispersion are also known as bottom-up and top-down approaches, respectively in the development of granules.

III. CLUSTERING

The goal of clustering is to group similar objects into the same cluster, while dissimilar objects should belong to different clusters. The degree of similarity of objects is calculated based on their respective feature values. In terms of granular computing, a cluster can be interpreted as an information granule that presents its objects on a coarser and more granular level. Two important areas in clustering are hierarchical and partitioned approaches.

Hierarchical clustering is divided into divisive and agglomerative methods. In hierarchical divisive clustering, one starts with one cluster for all data. By splitting clusters, the representation of the data gets finer and finer until each object

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

forms its own cluster. In hierarchical agglomerative clustering, one starts with each object forming its own cluster and move upwards merging clusters until all objects belong to the same cluster.

In partitioned clustering, the objects are assigned to a (predefined) number of clusters based on similarities. The most popular partitioned cluster approach is probably k-means and its extensions and derivatives.

A) K-MEAN

The k-means algorithm is most widely used well known unsupervised partitioning method for data clustering. K-means clustering method grouped data based on their closeness to each other according to Euclidean distance. In this clustering approach user decide that how many clusters should be and on the basis of closeness of data vector to the centroid, which is mean of the data vector of cluster is assigned to that cluster which shows minimum distance. The result depends on the number of cluster (k value) and the initial centroid chosen by the K-Mean algorithm.

K-Mean Algorithm

Step 1:

Initialize the cluster's centroid vectors randomly.

Step 2:

For each data vector calculate the distance between data vector and each cluster centroid which will minimum data vector will assign with that cluster and distance calculate using equation (1). Where d is the dimension.

$$d(Z_p, M_j) = \sqrt{\sum_{k=1}^{d} (Z_p, K - M_j, K)} \quad \text{..... (1)}$$

Where Zp is the pth data point, M_J is centroid of jth cluster.

Step 3:

Recalculate the centroid of cluster using equation (2)

$$Mj = \frac{1}{n}(\sum \nabla Z_p \in C_j \; Z_p) \quad \text{..... (2)}$$

Where nJ is the number of data point in cluster j.

Step 4:

Repeat step 3&4 until stopping criteria satisfied. The satisfying criteria can be either number of iteration or change of position of centroid in consecutive iterations.

B) PARTICLE SWARM OPTIMIZATION (PSO)

PSO is a population-based search algorithm which is initialized with a population of random solutions, called particles. As against the other evolutionary computation techniques, each particle in this algorithm, called PSO is also associated with a velocity. In PSO, swarm is composed of a set of particles $P = \{ p1 , p2 , p3 ,..... pn \}$. In PSO a number of simple entities "the particles" are placed in the search space of some problem or function, and each one of these evaluates the objective function at its current location.

The main advantage of Particle Swarm Optimization (PSO) technique is that it offers a globalized search methodology that can be used in K- Mean algorithm to find the optimal solution. Particle swarm optimization is an evolutionary computation technique which finds optimum solution in many applications. Using the PSO optimized clustering results in the components, in order to get a more precise clustering efficiency.

Clustering with swarm-based algorithms (PSO) is emerging as an alternative to more conventional clustering techniques. PSO is a population-based stochastic search

algorithm that mimics the capability of swarm (cognitive and social behaviour). Data clustering with PSO algorithms have recently been shown to produce good results in a wide variety of real-world data.

C) Benefits Of Using K-Mean And PSO In Combination

K-mean clustering is widely used to minimize squared distance between features values of two points reside in the same cluster. Particle swarm optimization is an evolutionary computation technique which finds optimum solution in many applications. Using the PSO optimized clustering results in the components, in order to get a more precise clustering efficiency.

PSO is a population-based stochastic search algorithm that mimics the capability of swarm (cognitive and social behaviour). PSO performs global search ability K-mean performs local search ability. One of the big issue with K-mean clustering algorithm was to define the number of clusters at the start of the clustering process by the user. To overcome such a problem, particle swarm optimization (PSO) and fuzzy theorem, automatically determines the appropriate number of clusters and their centres. PSO in sequence with K-Means algorithm for data clustering. When used overcome drawbacks of both algorithms, improves clustering and avoids being trapped in a local optimal solution. In this algorithm initial process starts by PSO due to its fast convergence and then the result of PSO algorithm is tuned by the K-Means near optimal solutions

IV. FOUNDATIONS OF THE DCC-FRAMEWORK

Static data characteristics are rather the exception than the rule in many real-life applications. Hence, dynamic approaches to clustering have become of rapidly increasing importance recently. They address the need to constantly adapt the clustering process to changes in the analyzed data domain. To categorize algorithms in the field of dynamic granular clustering, we propose the DCC-Framework that consists of the three crucial dimensions of dynamic clustering.
i.e.,

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

– Characteristics of change,
– Types of granulation,
– Clustering processes.

A) DCC: characteristics of change

DCC's Characteristics of Change dimension addresses kinds of change in the data domain to be analyzed. Characteristics of Change can be observed in several circumstances, e.g., in spatial environment or regarding time, beside others. They are:
– No change,
– Cluster movements, and
– Changes in cluster structures.

B) DCC: Types of granulation

Granulation dimension for the following three elements separately:

- input data,
- pre-processing
- Cluster approach.

C) DCC: clustering processes

In the Clustering Processes dimension, different types of algorithmic structures are identified and categorized.

The basic structure is similar for all k-means-like algorithms, basically the following four steps:
1. Initialization,
2. Calculation of the means,
3. Assignment of objects to clusters, and
4. Termination or going back to step 2.

In the dynamic case, the process of clustering is determined by several dimensions that may depend on each other. Here we propose

– Type of cluster algorithm,
– Flow of data, and
– Implemented dynamics.

V. SELECTED AREAS OF APPLICATION

A) Dynamic clustering of supermarket transactions

An ever increasing, tough competition in the retail industry calls for anticipation of customer's needs and requirements. Clustering the respective transactions provides important insights into consumer behaviour. If, e.g., in a supermarket customer behaviour changes during a day, data gathered at the point of- sales system (POS) reflect these drifts. An initial solution obtained, e.g., based on transactions during the morning hours could be updated as new transactions occur during the day.

B) Clustering messages in social media

Social media generate continuously new data, e.g., in short text messages which reflect what certain community members are concerned about: propose methods to analyze such messages in the context of marketing or crime detection, to name just a few. While these analyses are mostly static, it can be interesting to apply dynamic concepts as presented in the DCC-Framework.

VI. CONCLUSION

Dynamic aspects and granular information processing have received a lot of attention recently, both in the scientific community as well as in industry. Clustering is one of the most important tasks in data mining with a long and successful record of real-life applications. Today, clustering comprises of many different methods and extensions. In this paper a review of various researches done in the areas of K-mean and PSO is presented and it is concluded that when K-mean is used in combination with PSO it produces efficient results in terms of efficiency and accuracy because both the algorithms overcome the drawbacks of their own. Consequently, dynamic methods, granular computing, and clustering constitute important techniques in data mining. However, It was a structured presentation of existing approaches in the area that merges these three aspects of data mining, i.e., dynamic granular clustering. To fill this gap, we proposed the DCC-Framework. The Dynamic Clustering Cube can be regarded as an information granule itself that helps to make dynamic granular clustering more accessible and transparent by categorizing this field in an illustrative way. The analysis of the cubes three dimensions—(1) Characteristics of Change, (2) Types of Granulation, and (3) Clustering Processes—provides valuable insights into the corresponding phenomena. The DCC-Framework may support to further methodically enhance this field and may also motivate to use dynamic granular clustering in new application areas.

REFERENCES

[1] Xue-yong Li, Jia-xia Sun, Guo-hong Gao and Jun-hui Fu. Research of Hierarchical Clustering Based on Dynamic Granular Computing. Journal of computers, Vol. 6, No. 12, December 2011.

[2] Hsien-Tsung Chang, Nilamadhab Mishra, Chung-Chih Lin. IoT Big-Data Centred Knowledge Granule Analytic and Cluster Framework for BI Applications. journal.pone.0141980 November 24, 2015.

[3] Haiyan Li , Shen Yang, Hong Liu. Study of qualitative Data Cluster Model based on Granular Computing. 2013 AASRI Conference on Intelligent Systems and Control.

[4] Sankar K. Pal , Saroj K. Meher , Andrzej Skowron. Data Science, Big Data and Granular Mining 10.1016/j.patrec.2015.08.001 . *Pattern Recognition Letters.*

[5] Yao, Y.Y., Perspectives of Granular Computing Proceedings of 2005 IEEE International Conference on Granular Computing, Vol. 1, pp. 85-90, 2005

[6] Akansha Agrawal, Shreya Sharma. Optimizing k-means for Scalability . *International Journal of Computer Applications (0975 – 8887) Volume 120 – No.17, June 2015*

[7] A. K. Jain. Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31:651-666, 2010.

[8] Lorenzo Livi, Alireza Sadeghia. Granular computing, computational intelligence, and the analysis of non-geometric input spaces. Granul. Comput. (2016)

[9] Yiyu Yao. Granular Computing for Data Mining. *Proceedings of IEEE International Conference on Granular Computing*.

[10] Guoyin Wang • Ji Xu. Granular computing with multiple granular layers for brain big data processing. Brain Informatics (2014).

[11] Yao YY, (2008a) Granular computing: past, present, and future. In: Rough set and knowledge technology (RSKT 2008). LNAI, vol5009. Springer, Berlin, pp 27–28

[12] Yao YY (2008b) A unified framework of granular computing. In: Pedrycz W, Skowron A, Kreinovich V (eds) Handbook of granular computing. Wiley, Chichester, pp 401–410.

[13] Pritesh Vora, Bhavesh Oza February 2013 A Survey on K-mean Clustering and Particle Swarm Optimization, International Journal of Science and Modern Engineering, (IJISME) ISSN: 2319-6386, Volume-1, Issue-3.

[14] Pallavi Purohit and Ritesh Joshi March 2013 A New Efficient Approach towards k-means Clustering Algorithm ,International Journal of Computer Applications (0975-8887) Volume 65-No.11.