Global Journal of Computational Intelligence Research.
ISSN 2249-0000 Volume 3, Number 1 (2013), pp. 21-28

https://dx.doi.org/10.37622/GJCIR/3.1.2013.21-28

A GUI Based Tool for Clustering By Synchronization

P. Venkata Kishore[1], S. Chiranjeevi[2], A. Srinivas[3]

[1]Dept. of Computer Science and Engineering
[2]Assistant Professor, Dept. of CSE, GVPCE (A), Visakhapatnam
[3]Assoc.Professor&Dept. of CSE, CITM, Vizianagaram, A.P., India

Abstract

Synchronization is the one of the most captivating collective phenomena in nature. In every field like physics, surveys, research this concept plays a huge volume of attention. In nature, Every Data Object is an individual phase oscillator and synchronization is a powerful and inherently hierarchical concept regulating a large variety of complex processes. In the dynamical synchronization process, each object interacts with its local neighborhood only. As time evolves, small distinct groups of similar objects emerge which are gradually combined to form larger groups. Synchronized cluster as a phase oscillator and the dynamical behaviors of the objects over time are simulated. By the communication with similar objects, the phase of an object gradually aligns with its neighborhood data object, resulting in a non-linear object movement naturally driven by the local cluster structure. The inherently hierarchical nature of object movement allows exploring the hierarchical cluster structure at several levels of abstraction. In this paper, we present a GUI Based Tool which is designed to enhance the concept of clustering by synchronization using hierarchical clustering synchronization algorithm for identifying meaningful levels of the cluster hierarchy corresponding to high-quality synchronized clusters. We used K-means partitioned clustering algorithm to form clusters. For the effective synchronized clusters, we used hierarchical synchronization algorithm with the Minimum Description Length principle. We experimented the GUI based tool on synthetic and real time datasets obtained from UCI.
Keywords: Synchronization, clustering, hierarchical model.

Introduction

A cluster is a group of data objects which are similar characteristics to each other within cluster and dissimilar characteristics to other data objects belonging to other cluster [5].Data clustering is a young scientific discipline under vigorous development. There are large number of research papers scattered in many conference proceedings and periodicals, mostly in the fields of data mining, statistics, machine learning, spatial database, biology, marketing, and so on, with different emphases and different techniques. Owing to large amounts of data which is collected from different data resources, cluster analysis has recently become highly active topic in data mining research.

Partitioning clustering is efficient and conceptually simple, but is often difficult to unveil the natural cluster structure of complex hierarchical data. Hierarchical clustering provides output in the form of hierarchical structure which is more informative than the unstructured set of clusters formed by the partitioning clustering. During the last decades, hierarchical clustering has become very popular in various scientific disciplines, such as molecular biology, medicine or economy. However, well-known hierarchical clustering algorithms like Single Link [9] often fail to detect the true clusters for a real time data set. The dendrogram generated by Single Link is usually very complex and difficult to interpret. Moreover, the performance of many hierarchical clustering algorithms are very sensitive to noise and outliers. How can we find meaningful representation hierarchical clusters of a given dataset? In this paper, we present a GUI Based Tool which is designed to enhance the concept of clustering by synchronization using hierarchical clustering algorithm for identifying meaningful levels of the cluster hierarchy corresponding to high-quality clusters. For the effective clustering, we used synchronization algorithm sync which is combined with the Minimum Description Length principle [15]. The key idea is to regard each data object as a coupled phase oscillator and each object interacts dynamically with similar objects on different levels. The process of the hierarchical clustering inspired by synchronization involves the following stages: Starting from initial conditions, each object runs independently with its own phase. As time evolves, those objects with highest density forming local clusters will synchronize together with a small interaction range. Then, in a sequential process, more and more objects synchronize together and clusters are produced with a larger interaction range. Finally, the whole population will synchronize together and have a common phase. Thus, with different interaction ranges, the dynamical process of synchronization reveals the whole structure of the data set on all scales, from the micro-scale at an early stage up to the macro-scale. At each scale, outliers are effectively detected since they exhibit differently and hardly synchronize with any of the cluster objects. The principle of synchronization thus allows detecting a natural clustering of complex multi-scale data sets with outliers.

Rest of the paper is organized as follows: Section 2 gives the related work in the clustering process. The details of the GUI Based Tool are given in the section 3. Section 4 gives the experimental results. Conclusions are drawn in the section 5.

Related Work

K-Means clustering algorithm

K-means is one of the simplest partitioned clustering algorithm that solve the well known clustering problem. The procedure follows a simple and easy way to partition a given dataset into a certain number of clusters (assume k clusters) a fixed priori. The main idea is to define k centroids, one for each cluster. The next step is to take data object belonging to a given dataset and assign it to the nearest centroid. When no new data object is pending, the first step is completed and an early grouping is done. At this point we need to re-calculate k new centroids as centers of the clusters formed from the first step. After we have these k new centroids, a new binding has to be done between the same dataset points and the nearest new centroid. The procedure is repeated for all the data objects. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done.

Hierarchical synchronization

Hierarchical clustering algorithms [8,10] decompose a data set into several levels of partitions, representing by a dendrogram. One of the most well-known hierarchical clustering approaches is Single-Link [9].Initially the clusters are obtained by placing every data object in a unique cluster, in every step two closest clusters are merged until all objects are in a whole cluster [10].For the merging criterion several alternatives have been proposed, such as Average-Link and Complete-Link. The hierarchy obtained by the merging order is visualized as a dendrogram. For a real data set, the dendrogram is often very complex. If a large data set has N objects, the generated dendrogram contains N -1 layer and thus it is difficult to find optimal splitting levels that correspond to meaningful clusters. Outliers may also cause the so-called single-link effect that two clusters are difficult to be separated if there is a chain between the two clusters. The CURE [6] technique uses several representative points to evaluate the similarity measure between the clusters to form clusters of the arbitrary shape and avoid the so-called single-link effect. However, for a given data set, it is still difficult to define appropriate splitting levels which correspond to meaningful clusters. Furthermore, the dendrogram is created to indicate the clustering process, and does not display the true hierarchical structure of a data set.The well known OPTICS algorithm[7] analyzes the hierarchical data from the perspective of density. It provides the reachability plot to give a more intuitive and transparent way to visualize the hierarchical cluster structure for large data sets. However, for many real data sets, the reachability plot is very smooth and cannot find the hierarchal clusters. In combination with MDL [15], the algorithm hierarchical synchronization generates an interpretable cluster tree only consisting of meaningful levels, each representing a clustering of high quality. Below figure shows the generic framework of hierarchical clustering [8,10].

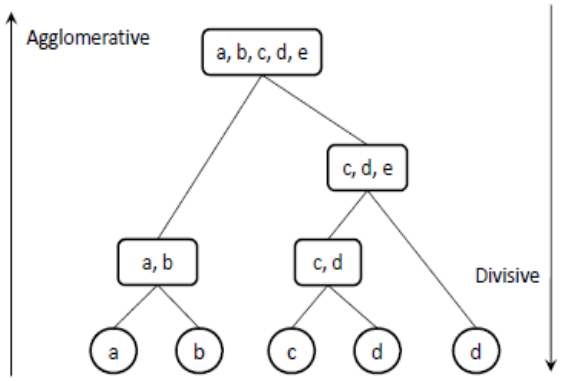

Figure 1: A Generic Framework showing working of hierarchical clustering.

Besides the cluster tree, the output of hierarchical synchronization includes the locality-quality diagram which allows the user to comprehensively assess the quality of the cluster hierarchy over all levels.

Key observation

The partitioning clustering algorithm Sync, it starts the dynamical interaction among objects with a small value of € and then increases it stepwise until all objects synchronize in a cluster. Minimum Description Length (MDL)[15] is used to find the best cluster structure: whenever the clusters are good representation of the data structure, they can be used for efficient coding(or compression) of the data set, which results in the minimal MDL value. Actually, the MDL principle can also be linked to hierarchical data analysis, not linking the global minimal MDL value, but all local stable minimal MDL values. The key observation is that if a data set exhibits a hierarchical cluster structure, the MDL values show several distinct stable local minima.

Specifically, when the interaction range € starts with a small value, objects with highest density will synchronize together and are regarded as a cluster. If the synchronized objects form reasonable clusters reacting the data structure at the micro-scale, the coding costs of the clusters will result in a local relatively low MDL value. By increasing of the € with the step size Δ€, if there exists a hierarchal structure of the data, a period of the interaction ranges € will result in the similar clustering results, which thus result in a period of stable MDL values. Then, in a sequential process, with the further increase of the €, more and more objects with less local density will tend to synchronize together. Equally, if these new synchronized clusters indicate a meaningful level of the hierarchical structure of data, it will result in a new period of relatively low and stable MDL values for a range of €. Finally, all objects may merge together with enough interaction range €.

Kuramoto Model

Kuramoto Model plays a very important role in synchronization concept.Kuramoto's theory for the synchronization transition of globally coupled phase oscillators to

populations where each oscillator has a different coupling strength. We show that, beyond the transition, even those oscillators with very small couplings may participate in the synchronized ensemble, provided that their natural frequencies are close enough to the synchronization frequency [11]. In finite systems, numerical realizations reveal that the transition is preceded by a regime of clustering where the population splits into internally synchronized groups of various sizes.

A GUI Based Tool

In this section we present the detail working of the GUI tool which uses synchronization concept and hierarchical clustering synchronization algorithm. Below figure shows the synchronized clusters from GUI tool.

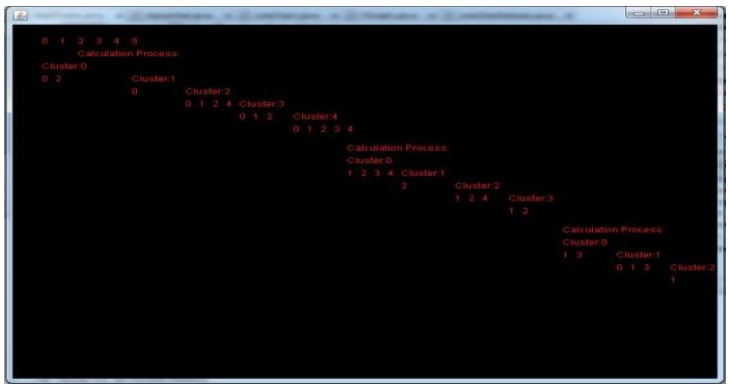

Figure 2: Synchronized clusters from GUI Tool with Hierarchical Clustering Algorithm.

Experimental Work

All the experiments were done on 2.50GHz Intel Core i5 machine with 4GB main memory running Window 7 operating system. We implemented the GUI Tool using Java. Below experiment results shows that clustering by synchronization using hierarchical clustering algorithm gives more accurate synchronized clusters when compared with clustering by synchronization with kuramoto model.

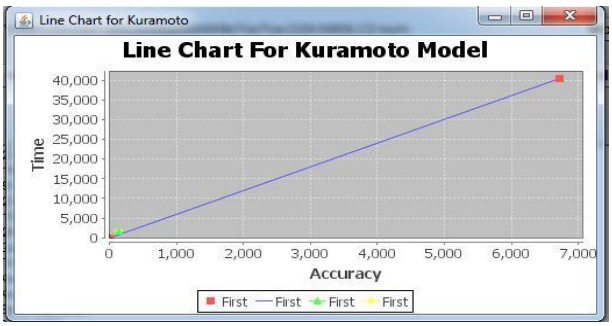

Figure 3: Performance of GUI Based Tool with kuramoto model.

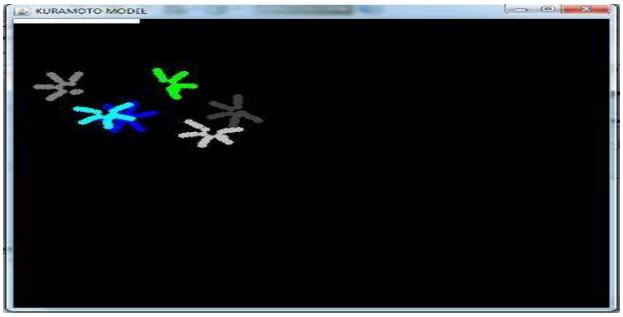

Figure 4: Figure shows Synchronized clusters from GUI tool with kuramoto model.

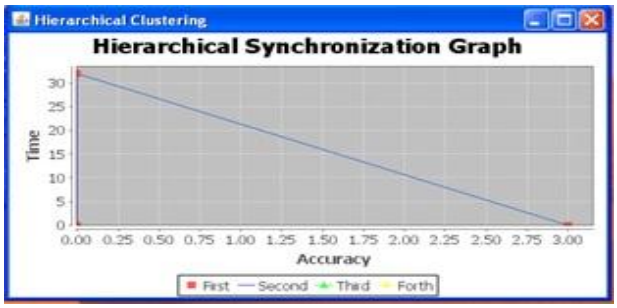

Figure 5: Performance of GUI Tool with Hierarchical Clustering Algorithm.

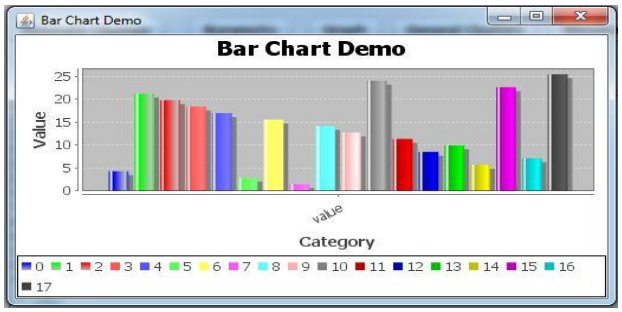

Figure 6: Performance of GUI Tool for synchronization level of hierarchical Clustering algorithm.

We experimented the GUI Tool on Tic-Tac-Toe Endgame dataset [16] that contains a set of board configurations possible at the end of game.It includes 958 instances (legal tic-tac-toe endgame boards) and 9 attributes, each corresponding to one tic-tac-toe square.

Conclusion and Future Work

In this paper we present a GUI based tool for clustering by synchronization which uses hiearchical clustering synchronization algorithm with MDL principle for accurate synchronized clusters.We experimented the GUI based tool on real dataset

and evaluated the performance with kuramoto model. Finally we conclude that the GUI based tool with hierarchical clustering synchronization algorithm for clustering by synchronization performs more accurately and time efficiently when compare with clustering by synchronization using kuramoto model.Future work will focus on exploiting the powerful concept of synchronization for subspace clustering.In addition, we will investigate on data visualization techniques based on the simulated object movement. As a long term goal we want to closely integrate simulation into the datamining process to design robust algorithms.

References

[1] J.A. Acebron, L. L. Bonilla, C. J. P. Vicente, F. Ritort, and R. Spigler. The kuramoto model: Asimple paradigm for synchronization phenomena. *Rev.of Modern Physics*, 77(2):137-185, Jan. 2005.

[2] D. Pelleg and A. Moore. X-means: Extending k-meanswith efficient estimation of the number of clusters.*ICML Conf.*, pages 727-734, 2000

[3] C. BÄohm, C. Faloutsos, J.-Y. Pan, and C. Plant.Robust information-theoretic clustering. *KDD Conf.*, pages 65-75, 2006.

[4] D. Aeyels, and F. D. Smet. A mathematical model forthe dynamics of clustering. *Physica D: NonlinearPhenomena*, 273(19):2517-2530, 2008.

[5] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice-Hall, 1988

[6] S. Guha, R. Rastogi, and K. Shim. Cure: An e_cient clustering algorithm for large databases. In SIGMOD Conference, pages 73-84,1998.

[7] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics:Ordering points to identify the clustering structure. In SIGMOD Conf, pages 49-60, 1999

[8] C. Bohm and C. Plant. Hissclu:a hierarchical density-based method for semi-supervised clustering. In EDBT, pages 440 -451, 2008

[9] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice- Hall, 1988.

[10] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. Comput. J., 26(4):354-359, 1983.

[11] A. Arenas, A. Diaz-Guilera, J. Kurths, Y. Moreno and C. S. Zhou. Synchronization in complex networks. *Phys. Rep. 469*, pages 93-153, 2008.

[12] T. Zhang, R. Ramakrishnan, and M. Livny. An efficient data clustering method for very large databases. *SIGMOD Conf.*, pages 103-114, 1996.

[13] R.T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. *VLDB Conf.*, pages 144-155. 1994.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD Conf.*, pages 226-231 1996.

[15] P. GrÄunwald. A tutorial introduction to the minimum description length principle. *Advances in Minimum Description Length: Theory and Applications*, 2005.
[16] http://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame