Knowledge-Based Systems 000 (2016) 1–14

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Synchronization clustering based on central force optimization and its extension for large-scale datasets

Wenlong Hang[a,b], Kup-Sze Choi[c], Shitong Wang[a,b,*]

[a] School of Digital Media, Jiangnan University, Wuxi, Jiangsu, 212122, PR China
[b] Department of Computing, Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
[c] School of Nursing, Hong Kong Polytechnic University, Hung Hom, Hong Kong, China

ARTICLE INFO

Article history:
Received 18 May 2016
Revised 9 October 2016
Accepted 13 November 2016
Available online xxx

Keywords:
Gravitational kinematics
Central force optimization
Partial synchronization
Synchronization clustering
Fast kernel density estimation
Large-scale datasets

ABSTRACT

Although research on clustering methods has been active in recent years, not only must most current clustering methods pre-set the number of clusters or other user-specific parameters but they also perform on large-scale datasets inefficiently. In this paper, we study the clustering problem by exploring the metaphor of gravitational kinematics based on Central Force Optimization (CFO). However, different from the global synchronization of CFO, we propose a new algorithm G-Sync by simulating the partial synchronization phenomenon. Specifically, we view each data object as a probe and simulate the dynamic interaction behaviour of data objects in the gravitational field. As time evolves, similar data objects will naturally come into partial synchronization and form distinct clusters measured by the proposed degree of local synchronization, and the dynamic interaction behaviour of the data objects is continually simulated over time. By introducing the Davies–Bouldin (DB) index, G-Sync can determine clusters of arbitrary size, shape and density. Moreover, pre-setting the number of clusters to be found is not required. The algorithm is further extended for handling large-scale datasets with the scalable S-G-Sync algorithm, which is based on fast kernel density estimation (FastKDE). S-G-Sync initially condenses a large-scale dataset quickly into its reduced dataset, followed by adaptive clustering on the reduced dataset using G-Sync. Finally, the Clustering on Remaining Objects (CRO) algorithm is proposed to cluster the remaining objects in the large-scale dataset and to capture outlier and singleton clusters effectively. The effectiveness of the G-Sync and S-G-Sync algorithms is theoretically analysed and experimentally verified on synthetic and real-world datasets.

1. Introduction

Cluster analysis is one of the most widely used techniques for pattern recognition, with applications ranging from image segmentation, data mining and computer vision to bioinformatics. Existing clustering algorithms can be broadly categorized into hierarchical clustering [1], partition-based clustering [2], clustering based on probability density [3], clustering based on graph theory [4], clustering by synchronization [5,31] and evolutionary-based clustering [6]. Developed based on different notions, these algorithms share the same objective – to separate the dataset through its internal homogeneity and external separation, i.e., patterns in the same cluster should be similar to each other and different from those in other clusters.

Despite the success of the current clustering algorithms, most of them require pre-setting the number of clusters or other user-specific parameters. In many real-world applications, the choice of the aforementioned information in particular is very time consuming or impossible, making many clustering methods infeasible. To address this problem, a feasible scheme has been proposed by simulating the synchronization phenomena of objects based on their intrinsic structure to achieve automatic clustering [5]. Synchronization is a basic yet powerful concept that regulates a wide variety of complex processes in nature, e.g., in the contexts of physics, chemistry and biology. It reveals the basic phenomena that objects with a common rhythm could eventually come into co-occurrence, despite the different initial distribution of individual objects. Based on synchronization phenomena, Bohm et al. [5] recently proposed a novel synchronization clustering method (Sync) via implementing the extensive Kuramoto Model. Moreover, Wang et al. [31] extended Sync to its scalable version for clustering analysis of large datasets. Sync and its scalable version can detect the number of clusters according to the intrinsic structure of a dataset. However, note that the formulation of the Kuramoto Model derives from the behaviour of systems of chemical and biological oscillators. The model makes two major assumptions: 1) there must be weak cou-

* Corresponding author.
E-mail addresses: hwl881018@163.com (W. Hang), kschoi@ieee.org (K.-S. Choi), wxwangst@aliyun.com (S. Wang).

http://dx.doi.org/10.1016/j.knosys.2016.11.007
0950-7051/

pling between each pair of objects, and 2) the interactions depend sinusoidally on the phase difference between each pair of objects. More critically, because different objects have their own intrinsic frequencies, the Kuramoto Model cannot theoretically guarantee its global convergence under the natural coupling strength.

To circumvent this shortcoming, we introduce an alternative model based on gravitational kinematics to simulate the phenomenon of partial synchronization. More specifically, in the modelling process, each data object is considered a probe when simulating the dynamic interaction behaviour of data objects according to gravitational kinematics over time. As time evolves, similar data objects will naturally come into partial synchronization and form distinct clusters measured by the proposed degree of local synchronization, and the dynamic interaction behaviour of the data objects is continually simulated over time. Different from the previous model, the metaphor of the gravitational kinematics model can guarantee that objects will eventually converge to global synchronization in view of Central Force Optimization (CFO) [7].

Inspired by the metaphor of gravitational kinematics, we propose a new clustering method, named G-Sync. Unlike existing clustering algorithms, G-Sync considers that objects interact dynamically and gradually synchronize to form local clusters accordingly to the intrinsic structure of the dataset as time evolves. G-Sync can determine clusters of arbitrary size, shape and density. It is not necessary to pre-set the number of clusters to be found. G-Sync can also adaptively specify the ε-neighbourhood of each object according to the Davies–Bouldin (DB) index; thus, an optimal clustering result can be selected. Additionally, the cluster order parameter O_{KDE} is proposed to measure the level of local synchronization, which can also act as the local cluster stop criterion in the process of clustering. Note that a relationship between O_{KDE} and kernel density estimation (KDE) is established in this study which therefore can facilitate the observation and evaluation of the local synchronization process from the perspective of KDE.

The G-Sync algorithm is further extended to the scalable version S-G-Sync for handling large-scale datasets, which condenses a large-scale dataset using the fast KDE algorithm (FastKDE) [11] based on the entropy-based integrated squared error (ISE) criterion [13]. By a large-scale dataset, we mean that it follows the definition in [40], i.e., it requires storage of between 10^6 and 10^8 bytes. The computational complexity of FastKDE is only related to the size of the reduced dataset rather than to that of the original large-scale dataset. Clustering is then performed on the reduced dataset using G-Sync, followed by the use of the Clustering on Remaining Objects (CRO) algorithm, which is proposed to cluster the remaining objects in the large-scale dataset and find the outliers and singleton clusters therein.

The main contributions of this paper are summarized as follows.

(1) Based on the metaphor of gravitational kinematics, the new partial synchronization-clustering algorithm G-Sync is proposed to discover clusters of arbitrary size, shape and density, without the need to pre-set the number of clusters to be found. Additionally, the clustering result of G-Sync can be automatically determined with a combination of the DB index.
(2) The cluster order parameter O_{KDE} is proposed to measure the level of local synchronization in the process of clustering. The relationship between O_{KDE} and the KDE-based density estimator is established to facilitate the observation and evaluation of the partial synchronization phenomenon.
(3) The scalable version S-G-Sync is further proposed for large-scale datasets. The FastKDE is adopted to realize fast condensation according to the entropy-based ISE criterion. Furthermore, the CRO algorithm is proposed to complete the clustering on the remaining objects in the original large-scale dataset and can identify the outliers and singleton clusters for these remaining objects.

The rest of the paper is organized as follows. In Section 2, we introduce the metaphors of gravitational kinematics and CFO and then propose the novel synchronization-clustering algorithm G-Sync. The scalable version of G-Sync, i.e., S-G-Sync, developed for large-scale datasets, is discussed in Section 3. In Section 4, we present the thorough experimental studies conducted on synthetic and real-world datasets and the performance of the proposed G-Sync and S-G-Sync to demonstrate their effectiveness. Finally, Section 5 concludes the paper and poses several open issues worthy of further study.

2. Basic concepts and the G-Sync algorithm

We initially introduce the concepts of gravitational kinematics and CFO, followed by a detailed description of the proposed synchronization-clustering algorithm G-Sync, which is based on the metaphor of gravitational kinematics. The major notations used in this paper are listed in Table 1.

2.1. Metaphor of gravitational kinematics and CFO

In this subsection, the metaphor of gravitational kinematics is initially presented, followed by the deterministic multi-dimensional search metaheuristic, CFO.

Newton's universal law of gravitation describes the motion of two bodies under mutual gravitational attraction. According to Newton's universal law of gravitation, the gravitational force of attraction F between masses m_1 and m_2 can be formulated as [7,12]

$$F = G\frac{m_1 m_2}{r^2}, \tag{1}$$

where the parameter r indicates the distance between masses m_1 and m_2. The coefficient G denotes the gravitational constant.

Due to the gravitational attraction, all of the masses mutually attract and move towards each other. The closer the masses, the more rapid the speed of motion because of the gravitational force.

Here, we give the acceleration $\mathbf{a}$ of mass m_1 caused by mass m_2:

$$\mathbf{a} = -G\frac{m_2\hat{\mathbf{r}}}{r^2}, \tag{2}$$

where $\hat{\mathbf{r}}$ denotes the unit direction vector. It provides the direction of acceleration $\mathbf{a}$, i.e., from mass m_2 to mass m_1.

During the interval t to $t+\Delta t$, with the above constant acceleration, the position vector of mass m_1 can be expressed as [12]:

$$\mathbf{x}(t+\Delta t) = \mathbf{x}(t) + \mathbf{v}(t)\Delta t + (1/2)\cdot\mathbf{a}\Delta t^2. \tag{3}$$

where $\mathbf{v}(t)$ and $\mathbf{x}(t)$ are the velocity vector and position vector at time t, respectively. In addition, $\mathbf{x}(t+\Delta t)$ indicates the position of mass at time $t+\Delta t$.

Because gravitational force acts along the line between the centres of the two masses, the deterministic multi-dimensional search metaheuristic developed based on the metaphor of gravitational kinematics is thus called "Central Force Optimization" [7,12].

Heuristic algorithms are generally problem-dependent and usually become trapped in a local optimum due to the usage of greedy strategy [32–35]. However, metaheuristic algorithms such as Particle Swarm Optimization (PSO) [36,37] and Ant Colony Optimization (ACO) [38,39] are problem-independent algorithmic frameworks that can provide a set of guidelines to form different heuristic algorithms. The major problem of current metaheuristic algorithms is that they can hardly offer the mathematical proof of global convergence [7,34]. Different from current metaheuristic algorithms,

Please cite this article as: W. Hang et al., Synchronization clustering based on central force optimization and its extension for large-scale datasets, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.11.007