Expert Systems with Applications 42 (2015) 7746–7758

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A new clustering algorithm based on near neighbor influence

Xinquan Chen *

Web Sciences Center, University of Electronic Science & Technology of China, Chengdu 611731, China
School of Computer Science & Engineering, Chongqing Three Gorges University, Chongqing 404000, China

ARTICLE INFO

Article history:
Available online 11 May 2015

Keywords:
Near neighbor point set
Near neighbor influence
Dissimilarity measure
Clustering

ABSTRACT

Clustering has been used in many areas. It is an unsupervised learning method which tries to find some distributions and patterns in unlabeled data sets. Although clustering algorithms have been studied for decades, none of them is all purpose. This paper presents a new clustering algorithm, Clustering based on Near Neighbor Influence (CNNI), an improved version in time cost of CNNI algorithm (ICNNI), and a variation of CNNI algorithm (VCNNI). They are inspired by the idea of near neighbors and the superposition principle of influence. In order to clearly describe the three algorithms, it lists three basic concepts (near neighbor point set, grid cell, and near neighbor grid cell set) and introduces two important concepts (near neighbor influence and a kind of similarity measure). In the simulations, four famous clustering algorithms (K-Means, FCM, AP, and DBSCAN) are used as comparative algorithms. From the simulated experiments of some artificial data sets and some real data sets, we observed that CNNI, ICNNI, and VCNNI can find those obvious clusters and get better (or similar) clustering results than (or with) K-Means, FCM, and AP for some data sets. We also observed that ICNNI is faster than CNNI with the same clustering results, CNNI and ICNNI are faster than AP with better or similar clustering quality, CNNI needs less space than VCNNI and DBSCAN, and VCNNI gets similar clustering results with DBSCAN. Especially, CNNI, ICNNI, and VCNNI can easily find some noises or isolates. At last, it gives several solid and insightful future research suggestions.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Clustering is an unsupervised learning method which tries to find some distributions and patterns in unlabeled data sets. Usually, those points in the same cluster should have more similarity than other points in other clusters (Jain, Murty, & Flynn, 1999). Clustering has been used in many areas such as machine learning, pattern recognition, image processing, marketing and costumer analysis, agriculture, security and crime detection, information retrieval, and bioinformatics.

Clustering algorithms have been studied for decades. There have been hundreds of clustering algorithms until now, but none of them is all purpose. Almost all clustering algorithms have flaws. Some clustering algorithms are suitable for dealing with data with certain types, and some are suitable for handling data with special distribution structures. Many real data have complex distributions, diversiform types, great capacity, noises, or isolates. So there is a continuous demand for researching different kinds of clustering methods. In order to obtain better clustering results in real-world applications where the amount of data is often very large and the types of data are diversiform, some researchers try their best to develop new efficient and effective clustering algorithms.

Clustering algorithms can be categorized into partitioning methods (Bezdek, 1981; MacQueen, 1967), hierarchical methods (Guha, Rastogi, & Shim, 1998; Karypis, Han, & Kumar, 1999; Zhang et al., 1996), density-based methods (Ankerst, Breunig, Kriegel, & Sander, 1999; Ester, Kriegel, Sander, & Xu, 1996; Roy & Bhattacharyya, 2005), grid-based methods (Agrawal, Gehrke, Gunopolos, et al. 1998; Wang, Yang, & Muntz, 1997), and model-based methods (Theodoridis & Koutroumbas, 2006). Recently, quantum clustering (Horn & Gottlieb, 2002), spectral clustering (Luxburg, 2007; Schölkopf, Smola, & Müller, 1998), and synchronization clustering (Böhm, Plant, Shao, et al., 2010; Huang, Kang, Qi, & Sun, 2013; Shao, Ahmadi, & Kramer, 2014; Shao, He, Böhm, Yang, & Plant, 2013b; Shao et al., 2013a) have been presented and become popular.

Partitioning clustering methods divide the data into some non-overlapping groups by using an iterative process to minimize a cost function. Many partitioning clustering algorithms are efficient, but they need a predefined number of clusters that cannot be obtained easily. K-Means algorithm (MacQueen, 1967) and FCM algorithm (Bezdek, 1981) are two famous partitioning clustering algorithms.

* Address: School of Computer Science & Engineering, Chongqing Three Gorges University, Chongqing 404000, China. Tel.: +86 15123428097.
E-mail address: chenxqscut@126.com

http://dx.doi.org/10.1016/j.eswa.2015.05.007
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

Hierarchical clustering methods use a distance matrix as their input and obtain a hierarchy of clusters. They can be classified into agglomerative hierarchical clustering and divisive hierarchical clustering. Usually, many hierarchical clustering algorithms do not require any predefined parameter, although they have higher time cost than partitioning clustering methods. BIRCH algorithm (Zhang et al., 1996), CURE algorithm (Guha et al., 1998), and CHAMELEON algorithm (Karypis et al., 1999) are three famous hierarchical clustering algorithms.

Density-based clustering methods try to find obvious clusters and isolates by using sparse areas to separate areas of higher density. DBSCAN algorithm (Ester et al., 1996) is a famous density-based clustering method. In DBSCAN, "density-reachability" is used to connect points that satisfy a density criterion within certain distance thresholds. OPTICS algorithm (Ankerst et al., 1999), which is a generalization of DBSCAN, is said that it can handle different densities much better than DBSCAN. EnDBSCAN (Roy & Bhattacharyya, 2005) algorithm is an efficient variation of DBSCAN.

Grid-based clustering methods first divide the clustering space into a finite number of regular grids (hyper-rectangular cells) or flexible grids (arbitrary shaped polyhedra) that summarize the data, and then obtain obvious clusters by merging adjacent grids. We can see that almost all grid-based clustering methods are approximate and can deal with some massive data. STING algorithm (Wang et al. 1997) and CLIQUE algorithm (Agrawal et al. 1998) are two famous grid-based clustering algorithms.

Model-based clustering methods use some statistical models to obtain clusters of the data. They assume the data are generated from a finite mixture of probability distribution models. A mixture of multivariate Gaussian distributions is most widely used in this kind of clustering.

Quantum clustering methods, which are based on quantum mechanics, first create a scale-space probability function from the data, and then use analytic operations to obtain a potential function whose minima determine cluster centers. Finally, they search the cluster structure over scale that is determined by one parameter.

Spectral clustering methods use the eigenvalues of the similarity matrix of the data to obtain its clusters. Many spectral clustering algorithms that are successfully used in many applications can be implemented easily and often outperform traditional clustering algorithms such as K-Means. The Shi-Malik algorithm (Shi & Malik, 2000) for image segmentation was developed basing on spectral clustering.

Synchronization clustering methods, which are very novel in clustering field, try to find the intrinsic structure of the data without any distribution assumptions by using a dynamic synchronization process. SynC algorithm (Böhm et al., 2010) is a famous synchronization clustering algorithm.

Recently, several original clustering algorithms, such as Affinity Propagation (AP) algorithm (Frey & Dueck, 2007) and SynC algorithm (Böhm et al., 2010), were published. AP is a new type of clustering algorithm published on *Science* in 2007. After AP algorithm was published, clustering based on probability graph models grew a new research direction. As we know, SynC (Böhm et al., 2010) is the first synchronization clustering algorithm. After Böhm et al. (2010) presented SynC algorithm, synchronization clustering attracts some researchers, and some synchronization clustering methods (Huang et al., 2013; Shao et al., 2013a; Shao et al., 2013b; Shao et al., 2014) were published from different views.

Other many clustering papers concentrate on improving some famous clustering algorithms in time or space, extending some clustering algorithms, or doing some comparative experiments in new applications. For example, Bouguettaya, Yu, Liu, et al. (2015) present an efficient agglomerative hierarchical clustering algorithm by building a hierarchy based on a group of centroids, which represent a group of adjacent points in the data space. Ozturk, Hancer, and Dervis (2015) present IDisABC clustering algorithm in dynamic clustering. IDisABC algorithm cannot only automatically determine the optimal number of clusters but also get well clustering quality. Kolesnikov, Trichina, and Kauranne (2015) present a new method for determining an optimal number of clusters based on parametric modeling of the quantization error. Aliyari Ghassabeh (2015) find that the equilibrium points of Mean Shift (MS) algorithm are asymptotically stable, which means if the iterations in MS algorithm start in a neighborhood of an equilibrium point, the generated sequence will converge to that equilibrium point. We can see that it is a theoretical progress on the convergence of MS algorithm. Luz López García, García-Ródenas, González Gómez, (2015) present KK-Means clustering algorithm for functional data. Ritter, Nieves-Vázquez, and Urcid (2015) present a simple statistics-based nearest neighbor cluster detection algorithm that can eliminate background noise, outliers, and detect clusters with different densities from some data sets.

This paper presents a new clustering algorithm, Clustering based on Near Neighbor Influence (CNNI), an improved version in time cost of CNNI algorithm (ICNNI), and a variation of CNNI algorithm (VCNNI). They are inspired by the idea of near neighbors and the superposition principle of influence.

The remainder of the paper is organized as follows. Section 2 lists several related papers and introduces four comparative clustering algorithms simply. Section 3 gives some related concepts. Section 4 presents CNNI algorithm, ICNNI algorithm, and VCNNI algorithm. Section 5 validates our algorithms by some simulated experiments. Conclusions and future work are presented in Section 6.

2. Related work

This paper is related to several papers (Ertöz, Steinbach, & Kumar, 2003; Guha, Rastogi, & Shim, 1999; Hinneburg & Keim, 1998; Jarvis & Patrick, 1973; Strehl, Ghosh, & Mooney, 2000) in some aspects, although it is developed independently. Jarvis and Patrick (1973) presented a shared nearest neighbor approach to similarity first. In their work (Jarvis & Patrick, 1973), a shared nearest neighbor graph is constructed from a proximity matrix, in which a link is created between a pair of points X and Y if and only if point X and point Y have each other in their closest k nearest neighbor lists. A similar idea was later proposed in ROCK algorithm (Guha et al., 1999). Ertöz et al. (2003) presented an improved J-R clustering method by redefining the similarity between pairs of points in terms of how many nearest neighbors the two points share.

In the simulations of this paper, four famous clustering algorithms, K-Means, FCM, AP, and DBSCAN, are used as comparative algorithms. K-Means is a classical partitioning clustering algorithm presented in 1967. It has some inherent shortcomings. For example, it needs a predefined number of clusters, but the number of clusters cannot be obtained correctly before clustering. K-Means only adapts to find those globular clusters with near radius because of its assignment strategy of point to the nearest cluster. When the number of clusters is large, K-Means is very sensitive to initialization of means. FCM, which is a famous fuzzy partitioning clustering method, assigns a data set with a membership value between 0 and 1 to indicate its membership value to a cluster rather than assigning the data set to a unique cluster only. AP is a famous clustering algorithm based on probability graph models. It needs $\Theta(n^2)$ time cost and space cost when computing and storing a predefined similarity matrix corresponding to the data set. Finally, it produces a clustering result but can't present a hierarchical clustering structure in a form of multi-level tree. DBSCAN is a famous density-based clustering algorithm. It can find clusters of arbitrary shape based on the formal notion of density-reachability. K-Means and FCM have linear

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات