Advanced Materials and Technologies for Environmental Applications - AdvMatTechEnv
Open Acces Journal
powered by SciEdTech

INCREMENTAL CLUSTERING BY FAST SEARCH AND FIND OF DENSITY PEAKS

Waqas AHMAD[1], Saeed EL-ASHRAM[2,*], Maged EL-KEMARY[2], Ibrahim AL NASR[3,4]

[1]College of information science and technology, Beijing normal university, Beijing, China;

[2]Faculty of Science, Kafr El-Sheikh University, Kafr El-Sheikh, Egypt (saeed_elashram@yahoo.com);

[3]College of Science and Arts in Unaizah, Qassim University, Unaizah, Saudi Arabia;

[4]College of Applied Health Sciences in Ar Rass, Qassim University, Ar Rass 51921, Saudi Arabia

Abstract

Clustering by fast search and find of density peaks (CFSFDP) is a new density based algorithm that discovers the centers of cluster by finding the density peaks efficiently. CFSFDP is applicable to a lot of clustering problems that deal with static data. Nowadays, more and more data, such as, social networks, blogs, web pages, Internet of things etc., is appearing in dynamic manner. However, CFSFDP is applicable only to organize the static data into different clusters. This paper considers the technique to be used with CFSFDP in the incremental clustering problem. In this paper, a novel approach ICFSFDP based on Nearest Neighbor Assignment (NNA) is proposed. ICFSFDP utilizes the CFSFDP mechanism for clustering the initial dataset and the remaining data-points are assigned to existing clusters based on NNA. Three standard clustering benchmark datasets are used to test the performance of the proposed method. The experimental results present that ICFSFDP based on NNA is efficient and effective to cluster the dynamic data and it is robust to noise as well.

Keywords: Density Peaks, Clusters, Incremental Clustering, Dynamic data

1. Introduction

Clustering is one of the most fundamental approach to organize data into appropriate groups. It is frequently applied in different fields such as social networks [1], pattern recognition [2], cyber security [3], bioinformatics [4], environmental data analysis [5], and health care [6] etc. It aims to cluster the data according to the estimated intrinsic characteristics or similarities.

Most of the clustering algorithms are designed to discover intrinsic hidden patterns in static data [7]. Nowadays, more and more data, for example, social networks, blogs, web pages, Internet of things etc., is appearing in dynamic manner and is leading towards the big data [8, 9]. The characteristics of big data make it an extreme challenge for extracting the useful patterns from the data. Therefore, incremental clustering, evolutionary clustering, and stream mining are becoming hot research topics in the field of data mining. The big data concerns with large volume, complex patterns, and growing datasets from autonomous hydrogenous data sources [8]. This overhead directs the fundamental clustering algorithms to interpret and discover the hidden pattern from the data deluge rapidly. It requires the ability to adapt the changes occurring within the data over time, ability to detect new emerging clusters, merging of old clusters, and to detect the outliers or noise from data.

Clustering by fast search and find of density peaks (CFSFDP) [10] is a density based clustering algorithm, which cluster data rapidly by finding of density peaks. CFSFDP is based on the concept that: 1) cluster centers are points that have high density among its neighbors, and 2) cluster center is situated at a large distance from other cluster centers. For each data-point x_i, CFSFDP computes local density (ρ_i) and a minimum distance (δ_i) to a nearest high density point. Cluster centers are attained by plotting calculated values of ρ_i and δ_i, which is referred as the decision graph. In cluster analysis, the key challenge is to discover correct cluster centers in the data sets [11]. However, CFSFDP uses decision graph to select the correct cluster centers with minimum human interaction, which makes it more worthy to analyze big data/ streaming data. However, many variant of CFSFDP has been proposed to overcome the limitations of CFSFDP, such as [12,13,14,15,16]

In this paper to cluster streaming data, incremental clustering by fast search and find of density peaks (ICFSFDP) is presented. ICFSFDP is an NNA based clustering algorithm, which is robust to cluster the dynamic data.

The rest of paper is organized as follows. The related work is presented in Section II. Section III describes the ICFSFDP based on NNA. Experimental results are presented and discussed in Section 1V, and finally, the concluding remarks are presented in Section V.

2. Related Work

CFSFDP has the ability to create arbitrary shaped clusters by fast search of cluster centers. The main objective of CFSFDP is to identify clusters centers by fast search and finding of density peaks. The notion of decision graph is utilized to separate noise from cluster core points and to detect cluster centers, efficiently. CFSFDP is based on two assumptions that: 1) cluster centers are surrounded by neighbors with lower local densities, and 2) cluster center is relatively positioned at a large distance from other higher local densities. For each data point i, CFSFDP calculates its local density (ρ_i) and a minimum distance (δ_i) to its nearest high dense data point. Both ρ_i and δ_i depend only on the distance d_{ij}, which is assumed to satisfy the triangular inequality. The ρ_i of a point i is computed as follows:

Definition-1:

$$\rho_i = \sum_j^{n-1} X(d_{ij} - d_c),$$

such that

$$X(x) = \begin{cases} 1 & x < 0 \\ 0 & otherwise \end{cases}$$

where the distance of point i to j is denoted by d_{ij} and d_c is cutoff distance. ρ_i counts the numbers of data-points that are closer than d_c to i. d_c is selected through heuristic approach i.e. in average there exist 1 to 2 % of neighbors in a dataset. According to definition-1, d_c is an essential parameter used for the estimation of ρ_i. Therefore, the effectiveness of CFSFDP is a subject for an appropriate selection of d_c. However, in case of small datasets, ρ_i could be affected by a large statistical errors [10]. To illuminate such statistical errors, it is useful to use [17, 18] methods to estimate the densities in a more compact manner [10].

For each data point i, the distance δ_i is calculated for assigning i to its nearest high local dense point max_{ρ_i}. δ_i is computed as follows:

Definition-2:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} (d_{ij}) & if\ \exists\ j\ s.t: \rho_j > \rho_i \\ \max\limits_{j:\rho_j > \rho_i} (d_{ij}) & otherwise \end{cases}$$

For local or global maximum dense data-points, δ_i is much larger than typical nearest neighbor distance. Thus, cluster centers are considered as points for which the value of ρ_i and δ_i are unusually large. After the calculation of ρ_i and δ_i for entire dataset, these statistics are plotted on decision graph, as presented in Figure 1.

Figure 1 is a simple demonstration of CFSFDP.

Figure 1 (a) presents 28 sample data-points plotted in a decreasing density order. Figure 1(b) demonstrate the decision graph of sample points. Decision graph in Figure 1 (b) illustrates the data point 1 and 10 are the most highly dense with maximum value of δ, which is a characteristic of cluster centers. However, the data point 26, 27, 28 have the highest value of δ and low value of ρ therefore, can be considered as outliers or isolated clusters.

With a minimum human interaction, cluster centers can be selected successfully with the notion of decision graph. After successful identification of cluster centers, the remaining data-points are assigned to the nearest cluster centers in a single round.

Furthermore, to refine the clusters and to separate the noise, a border region of each cluster is identified. The border region is defined as the set of data-points assigned to a cluster but being at d_c from other cluster data-points. For each cluster, CFSFDP finds higher dense points in the border region denoted as ρ_b. The data-points of cluster whose density is higher than ρ_b are considered as cluster core, while the rest of others are identified as clusters halo and suitable to be considered as noise or outliers.

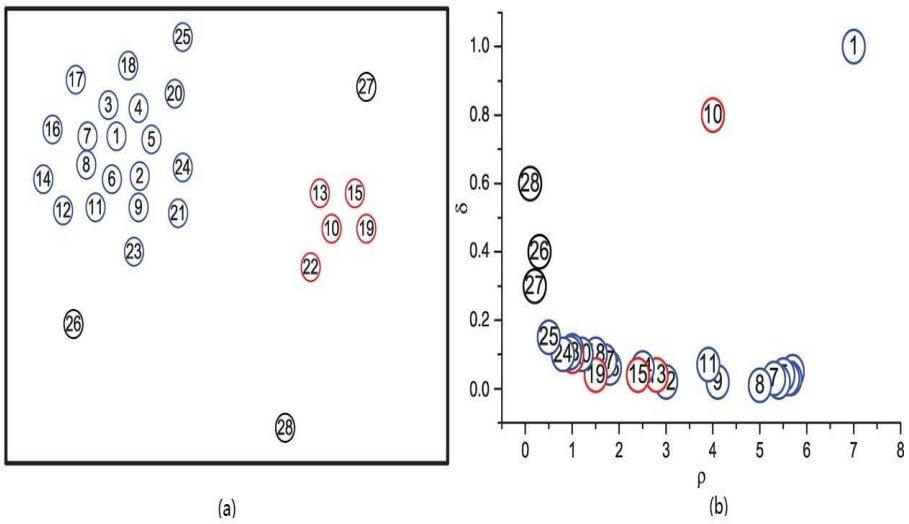

Figure 1: Decision graph representation of CFSFDP. (a) 28 data-points are plotted in a decreasing density order. (b) Decision graph representation of (a) [1].

3. Incremental CFSFDP Clustering

CFSFDP can be utilized in different fields such as pattern recognition, face detections, medical, document clustering etc. [10]. In the era of big data, most of data is geographically distributed and it is continuously increasing [8]. In various fields, such as banking, health and security etc., the value of analysis is highly subject to the freshness of the data, which emphasis on the concept of stream processing. CFSFDP is designed to work on static data. To cluster the dynamic data using CFSFDP is still considered as a big problem. In incremental clustering problems, the data depends upon the time at which it is available to cluster. Like in the case of the static data, data is arrived at different time stamps. The idea of incremental clustering is to cluster big partition of data by using CFSFDP and then new data-points are arranged according to the existing statistics of discovered clusters.

3.1. Incremental CFSFDP based on NNA

In this subsection, we describe Nearest Neighbor Assignment (NNA) as a subsequent algorithm to assign new arrival of data-points at time t_i to previously created clusters. Firstly, CFSFDP is used to cluster initial batch of data at t_0. For new arrival of data is assigned to existing clusters based on NNA relationship. NNA is based on the fact that if two objects are similar to each other they must be organize in a same cluster [9]. Definition-3 explains the notion of NNA.

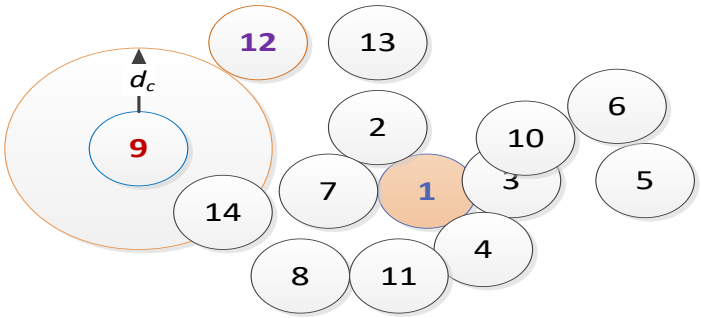

Figure 2: Density reachability. Black color points are clustered using CFSFDP at t_0. Data-point 9 is arrived at time t_1 and assigned to existing cluster because it is at a d_c distance to cluster core point 14.

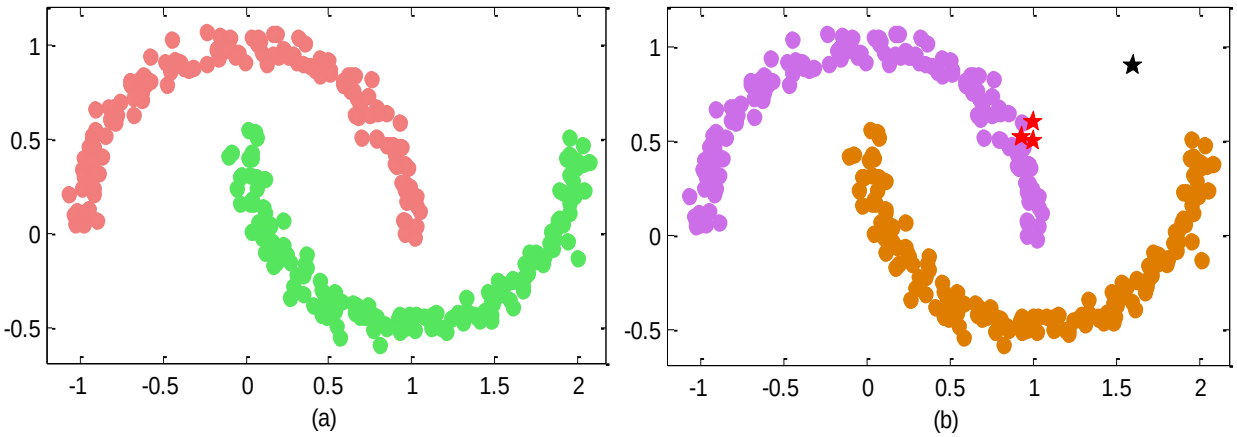

Figure 3: Toys problem dataset. (a) Toys problem dataset cluster by CFSFDP. (b) Presents that four new points are assigned to clusters core points based on nearest neighbor assignment configuration.

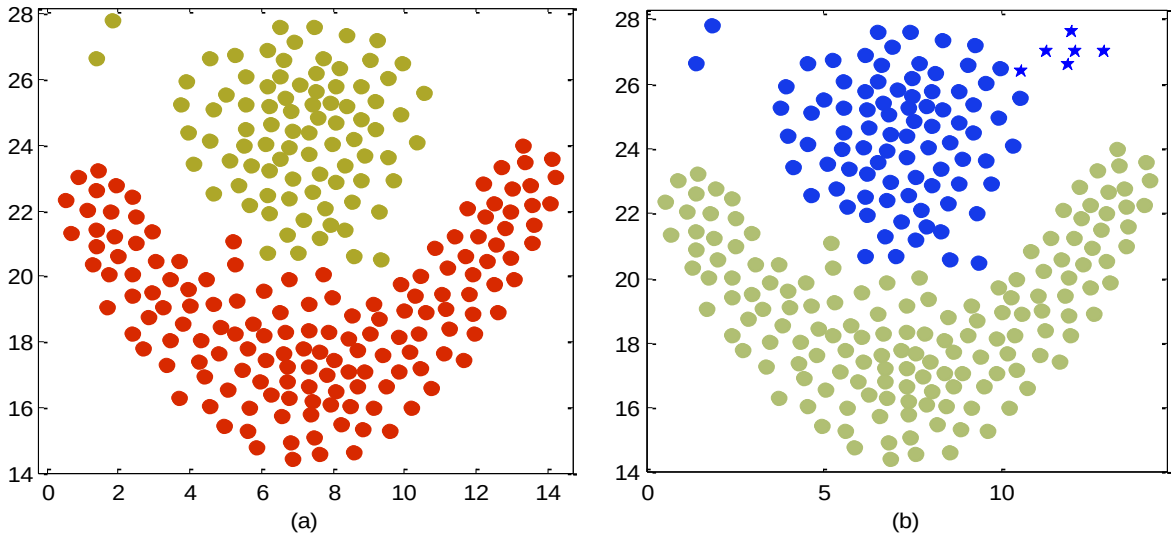

Figure 4: Toys problem dataset. (a) Toys problem dataset cluster by CFSFDP. (b) Presents that four new points are assigned to clusters core points based on nearest neighbor assignment configuration.

Definition-3: (Density Reachability)

A point i is density reachable from a cluster core point if it is in d_c radius of any cluster core point $Core_j$ such that if $\rho_i >= \rho_{Core_j}$. Figure 2 explains the density reachability fact. The point 9 arrives at t_1 and it falls in d_c radius from a core point 14 and its density is greater than the minimum dense point in underlying cluster. If a point is density reachable from more than two different cluster core points then it is assigned to the cluster, which have maximum resemblance of density to that data-point.

Definition-4: (Halo clusters)

If a point i is not density reachable according to definition-3, it can be considered as halo cluster.

Algorithm

Input: $existing\ cluster\ center\ \boldsymbol{C_i}, \forall\ \boldsymbol{i} \in \{\boldsymbol{1}, \boldsymbol{2}, \boldsymbol{3}, \dots, \boldsymbol{n}\}$,

, new data $Ndata$

Output: organized clusters

1 Calculate density of new data according to **definition-1.**

2 Calculate Distance delta from nearest cluster core point.

3 $for\ i = 1{:}\ size(Ndata), begin$

$if\ density_{reachibity} = True, begin$

$Ndata_i = core\_point$

$Endif$

$otherwise$

$Ndata_i = halo$

$endfor$

In this above algorithm, $\boldsymbol{C_i}$ are cluster's core points and $Ndata$ is new partition of dataset.

4. Results and Experiments

We use three clustering benchmark datasets with a small addition of some points to evaluate the robustness of proposed ICFSFDP method.

We use toys problem dataset by adding four new data-points including a noise data-point. The initial set of 300 data-points are clustered by standard CFSFDP at time (t_0). For remaining data-points, we calculate the densities of each data-point and find minimum distance to cluster's core point. Then data-points are assigned to existing clusters on the basis of definition-3. Figure 3(a) presents clusters at t_0 while Figure 3(b) presents clusters at t_1. Blue stars represent new cluster core points while black star represents a noise point because its density is lower than any core point and it is far away from d_c radius of any core point.

To validate our method over jeans dataset, we use flame dataset and add six new data-points in the dataset. The initial partition of the flame dataset is clustered by using CFSFDP as presented in the Figure 4(a). Furthermore, for time t_1 ICFSFDP successfully assigned the newly arrived data-points into existing clusters as presented in Figure 2(b). In Figure 4 (b) stars reprint the data-points that arrive at t_1. ICFSFDP finds their densities and then finds the minimum distance from neighboring core points. The new points are assigned to nearest neighbor clusters core points based on definition-3.

We also test proposed ICFSFDP on path-based spiral dataset. We partitioned the spiral dataset into three portions D1, D2, and D3. At t_0 ICFSFDP successfully discovers three clusters from D1 and then at t_1 and t_2 ICFSFDP assigns D2 and D3 data-points to existing core point's cluster based on the NNS respectively. Only two experimental results are presented because of shortage of space.

Conclusion

In this paper, we consider the case to apply CFSFDP in incremental and streaming clustering task. In ICFSFDP, the initial partition of dataset in clusters is made by using CFSFDP and the newly arrived data-points are assigned to existing clusters based on NNA. We test ICFSFDP on three synthetic datasets with addition of some data-points. The experiments validate ICFSFDP robustness in incremental clustering task.

References:

[1] Chang M-S, Chen L-H, Hung L-J, Rossmanith P, and Wu G-H, Exact algorithms for problems related to the densest k-set problem. Inform Process Lett. 2014;114(9):510-513

[2] Ahn C-S, and Oh S-Y. Robust vocabulary recognition clustering model using an average estimator least mean square filter in noisy environments. Personal and Ubiquitous Computing 2014;18(6):1295-1301.

[3] Yan Y, Qian Y, Sharif H and Tipper D, A survey on cyber security for smart grid communications, IEEE Commun Surv Tut. 2012; 14(4): 998-1010

[4] Fu L and Medico E, FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data, BMC Bioinformatics. 2007:8(1):3.

[5] Charreire H, Weber C, Chaix B, Salze P, Casey R, *et al.*, Identifying built environmental patterns using cluster analysis and GIS: Relationships with walking, cycling and body mass index in French adults, Int J Behav Nutr Phys Act. 2012; **9**:59 (11pg)

[6] Lefèvre T, Rondet C, Parizot I, and Chauvin P, Applying Multivariate Clustering Techniques to Health Data: The 4 Types of Healthcare Utilization in the Paris Metropolitan Area. PloS One. 2014; 9(12): e115064.

[7] Liao TW, Clustering of Time Series Data: A Survey, Pattern Recognit, 2015; 38(11):1857-74

[8] Wu X, Zhu X, Wu G-Q, and Ding W. Data mining with big data. IEEE Trans. Knowl. Data Eng. 2014;26(1): 97-107.

[9] Sun L and Guo C. Incremental affinity propagation clustering based on message passing. IEEE Trans. Knowl. Data Eng. 2014;26(11):2731-2744.

[10] Rodriguez A, and Laio A. Clustering by fast search and find of density peaks. Science; 2014:344(6191):1492-1496.

[11] Jain AK. Data clustering: 50 years beyond K-means. Pattern Recogn Lett. 2008;31(8):651-666.

[12] Mehmood R, Zhang G, Bie R, Dawood H, Ahmad H. Clustering by fast search and find of density peaks via heat-diffusion. J Neurocomputing, 2016;208:210–217.

[13] Bie R, Mehmood R, Ruan R, Sun Y, Dawood H. Adaptive fuzzy clustering by fast search and find of density peaks. J Personal and Ubiquitous Computing, 2016;20(5):785–793

[14] Shanshan R, Mehmood R, Alowibdi J, Dawood H, Daud A. International World Wide Web Conference. An adaptive method for clustering by fast search-and-find of density peaks. ACM publisher, WWW' 17 Companion, 2017: 119-127.

[15] Mehmood R, El-Ashram S, Bie R, Dawood H, Kos A. Clustering by fast search and merge of local density peaks for gene expression microarray data. Sci Rep 2017:7: Article number: 45602.

[16] Mehmood R; Bie R, Jiao L, Dawood H, Sun Y. Adaptive cutoff distance: clustering by fast search and find of density peaks. J Intell Fuzzy Syst. 2016;31(5):2619-2628

[17] Marinelli F, Pietrucci F, Laio A and Piana S. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. PLoS Comput. Biol. 2009;5(8): e1000452.

[18] Horenko I, Dittmer E, Fischer A, Schütte C, Automated model reduction for complex systems exhibiting metastability. Multiscale Model Simul. 2006;5:802–827