Physica A 502 (2018) 345–355

Contents lists available at ScienceDirect

Physica A

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

GDPC: Gravitation-based Density Peaks Clustering algorithm☆

Jianhua Jiang [a], Dehao Hao [a], Yujun Chen [a], Milan Parmar [a], Keqin Li [b,*]

[a] School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, China
[b] Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

ARTICLE INFO

Article history:
Received 24 November 2017
Available online 26 February 2018

MSC:
00-01
99-00

Keywords:
Clustering analysis
Density peaks clustering
Gravitation theory
Anomaly detection

ABSTRACT

The Density Peaks Clustering algorithm, which we refer to as DPC, is a novel and efficient density-based clustering approach, and it is published in *Science* in 2014. The DPC has advantages of discovering clusters with varying sizes and varying densities, but has some limitations of detecting the number of clusters and identifying anomalies. We develop an enhanced algorithm with an alternative decision graph based on gravitation theory and nearby distance to identify centroids and anomalies accurately. We apply our method to some UCI and synthetic data sets. We report comparative clustering performances using *F*-Measure and 2-dimensional vision. We also compare our method to other clustering algorithms, such as K-Means, Affinity Propagation (AP) and DPC. We present *F*-Measure scores and clustering accuracies of our GDPC algorithm compared to *K*-Means, AP and DPC on different data sets. We show that the GDPC has the superior performance in its capability of: (1) detecting the number of clusters obviously; (2) aggregating clusters with varying sizes, varying densities efficiently; (3) identifying anomalies accurately.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Clustering is known as an unsupervised classification in pattern recognition, or nonparametric density estimation in statistics [1–4]. The goal of clustering is to separate finite unlabeled objects into different clusters with characteristics of internal homogeneity and external separation [5,6]. Clustering has been applied in a wide varieties of fields, ranging from engineering (machine learning, artificial intelligence, pattern recognition, mechanical engineering, electrical engineering) [7], computer sciences (web mining, spatial database analysis, textual document collection, image segmentation, complex networks) [8,9], life and medical science (genetics, biology, microbiology, psychiatry, clinic, pathology), to earth sciences (geography, geology, remote sensing), social sciences (sociology, psychology, education), and economics (marketing, business) [1,3,10–13].

Traditional approaches in clustering can be broadly categorized into hierarchical, partitioning, density-based, model-based, grid-based, and soft-computing methods [1]. Until now, Sander etc. [14] and Ertoz etc. [15] have been proposed numerous density-based clustering methods that inspired by the classical DBSCAN algorithm [16] that having capability of extracting clusters with arbitrary shapes with an overall average runtime complexity of $O(n \times \lg n)$. In 2014, there was a big breakthrough in density-based clustering approaches. A novel clustering algorithm based on *density peaks*, named as DPC, was proposed by Rodriguez and Laio [17] in the journal *Science*. Compared with various classic density-based approaches

☆ The authors are grateful to the financial support by the National Natural Science Foundation of China (no. 61572225), the Foundation of the Education Department of Jilin Province, China (no. JJKH20170119KJ) and the Natural Science Foundation of the Science and Technology Department of Jilin Province, China (no. 20180101044JC).

* Corresponding author.
E-mail address: lik@newpaltz.edu (K. Li).

https://doi.org/10.1016/j.physa.2018.02.084
0378-4371/© 2018 Elsevier B.V. All rights reserved.

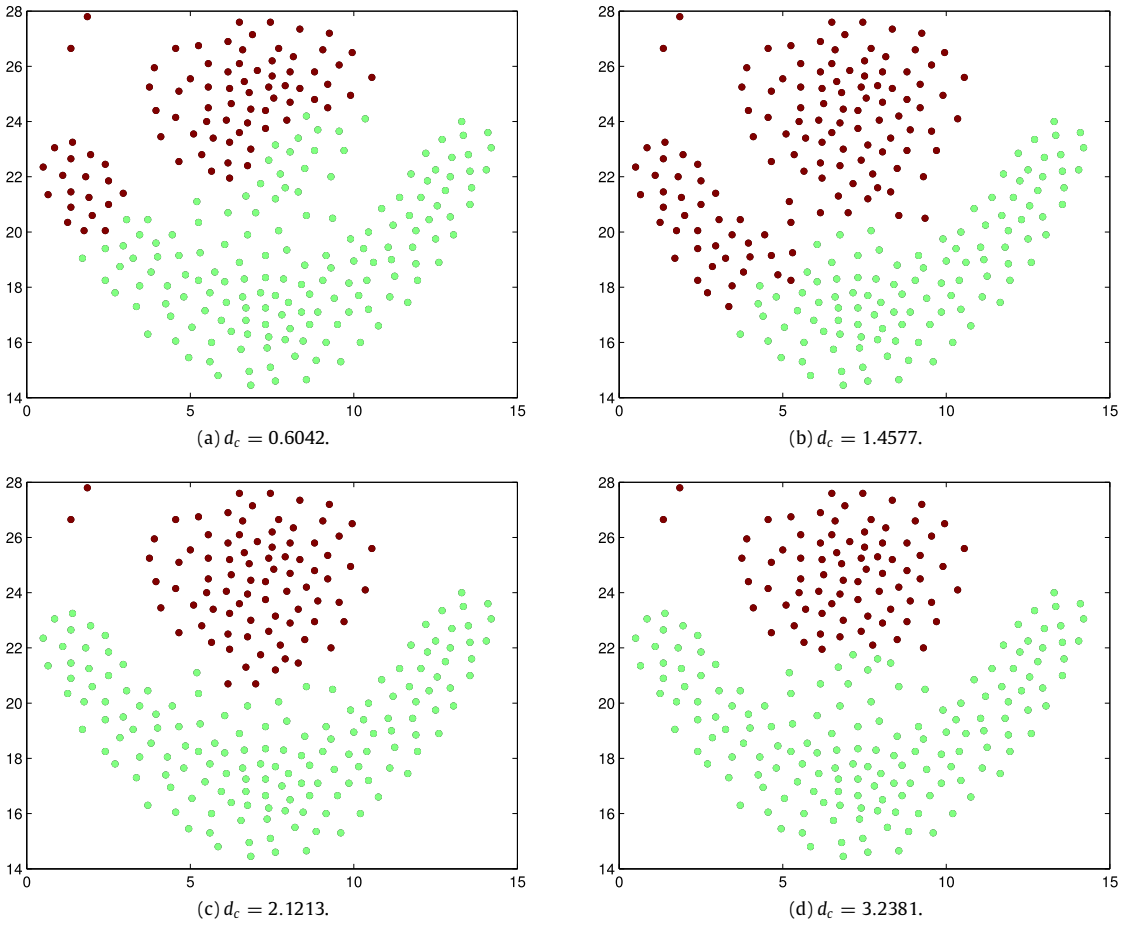


Fig. 1. Failures in anomaly detection by the DPC in the *Flame* data set with different d_c values.

[14–16], the DPC [17] has relative advantages of identifying centroids, aggregating arbitrary shaped clusters with varying sizes and densities. The DPC is based on the concept that cluster centers are characterized by a higher density in comparison with their neighbors and by a relatively larger distance from points with higher densities. The DPC is effective with two assumptions [17]:

(1) cluster centers are surrounded by neighbors with lower local density;
(2) they are at relatively larger distance from any points with a higher local density.

There are some limitations [18–22] in the DPC algorithm, and anomaly detection is one of the major limitations. *Anomaly detection* refers to the problem of finding patterns in data that do not conform to expected behavior [23]. The importance of *anomaly detection* is due to the fact that anomalies in data translate to significant (and often critical) actionable information in a wide variety of application domains [23]. As illustrated in Fig. 1, the two anomalies in the top left corner cannot be identified with the DPC algorithm even with different d_c values in the *Flame* data set. Henceforth, the capability of anomaly detection should be enhanced.

The proposed Gravitation theory based Density Peaks Clustering algorithm, which we refer to as GDPC, is inspired by the gravitational clustering algorithm [24]. When compared with algorithms such as *k*-Means [25], AP [2] and DPC [17] illustrated in Table 1, the proposed novel GDPC algorithm has the capability of detecting anomalies accurately, and aggregating clusters with varying sizes and densities efficiently.

We have tested the novel GDPC algorithm in the most popular clustering benchmarks to demonstrate its feasibility and correctness. The proposed GDPC has overcome the limitation of anomaly detection problem mentioned above with satisfactory results when compared with the DPC and other algorithms on synthetic data sets and some UCI data sets. The major contributions of this paper can be highlighted as follows:

(1) Gravitation theory is applied into the DPC algorithm to enhance its capability of detecting anomalies.

Table 1
Advantages of GDPC when compared with k-Means, DBSCAN and DPC, where '×' refers to disable, '√' means able and '∂' is partial.

Algorithms	N of clusters	Varying sizes	Varying densities	Anomaly detection
k-Means	∂	∂	×	×
AP	∂	∂	∂	×
DPC	∂	√	√	∂
GDPC	√	√	√	√

(2) The decision graph is optimized by the universal gravitation force with capability of identifying centroids and anomalies accurately.

The rest of this paper is organized as follows. Section 2 proposes the novel universal gravitation theory based density peaks clustering algorithm. Section 3 presents experimental results on some synthetic data sets and UCI data sets. Section 4 makes discussions to explain the major reasons of the GDPC's strengths in contrast with the DPC. Finally, we have derived the conclusions and relevant remarks are given in last section along with the expected future works.

2. Methods

The proposed GDPC inherits the strengths of the DPC [17] and gravitational clustering algorithm [24]. The GDPC assumes that: (1) a cluster is formed by a centroid and surrounded by density decreasing nodes; (2) a node can be assigned to the cluster where there is a higher density node with relatively higher gravitation force. As illustrated in Algorithm 1, the GDPC includes three major steps:

- Step 1: calculate and sort node density;
- Step 2: generate universal gravitation decision graph;
- Step 3: aggregate clusters with universal gravitation force.

2.1. Calculate and sort node density

Similar to the DPC [17], a scientific cutoff distance d_c is adopted to calculate their local densities ρ for sorting these density values in the descending order as follows:

$$d_{ij} = distance(node_i, node_j), \tag{1}$$

$$\rho_i = \sum_j \chi(x) \times (d_{ij} - d_c), \tag{2}$$

where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise, and d_c is the only user-defined parameter of the cutoff distance. As a rule of thumb, one can choose d_c so that the average number of neighbors is around 1% to 2% of the total number of points in a data set [17]. Another local density of a point $node_i$ presented in the DPC is in Eq. (3):

$$\rho_i = \sum_j exp\left(-\frac{d_{ij}^2}{d_c^2}\right), \tag{3}$$

$$\delta_i = \min_{j:\rho_j > \rho_i} d_{ij}. \tag{4}$$

2.2. Generate universal gravitation decision graph

Newton's law of universal gravitation states that a particle attracts every other particle in the universe using a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them as depicted in Eq. (5):

$$F = G \times \frac{m_1 \times m_2}{r^2}, \tag{5}$$

where:

- F is the force between the masses;
- G is the gravitational constant;
- m_1 is the first mass;
- m_2 is the second mass;
- r is the distance between the centers of the masses.

Table 2
Mapping between Newton's law of gravitation and parameters from DPC.

Parameters from gravitation	Parameters from DPC	Equations
m_i	ρ_i	Eq. (2) or (3)
m_j	ρ_j	Eq. (2) or (3)
r	δ_i	Eq. (4)

Table 3
Seven different data sets.

Data sets	Nodes	Dimensions	Clusters
Iris	150	4	3
Seeds	210	7	3
Wine	178	13	3
Glass	214	9	6
Flame	240	2	2
Aggregation	788	2	7
Spiral	312	2	3

Newton's law of universal gravitation gives inspiration that the distance can be replaced by the gravitation force F to have a better metric to detect centroids and anomalies. The mapping between Newton's law of universal gravitation and parameters from the DPC is depicted in detail in Table 2.

Based on Table 2, the Newton's law of gravitation can be described as Eq. (6).

$$F = G \times \frac{\rho_i \times \rho_j}{{\delta_i}^2}. \tag{6}$$

We adopt the reciprocal of gravitation F as the vertical axis of the decision graph, while the density ρ as its horizontal axis as illustrated in Fig. 3.

2.3. Aggregate clusters with universal gravitation force

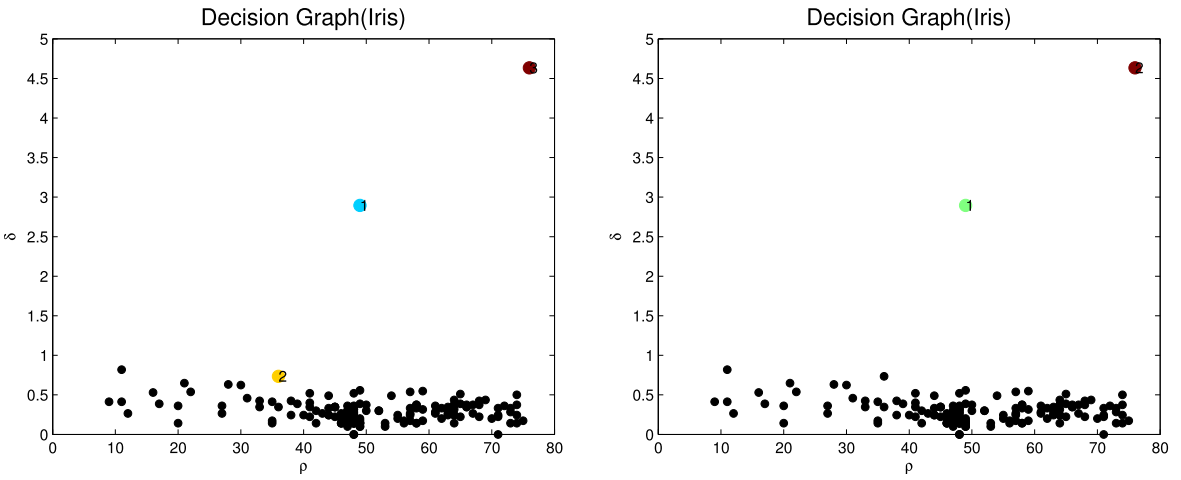
Universal gravitation based aggregation is processed by detecting centroids based on density peaks and clustering nodes based on universal gravitation force. Firstly, each centroid of a cluster is found by its relatively low universal gravitation force and its relatively high density. Secondly, any node can be assigned to a cluster where there is a node with a relatively high universal gravitation force. Finally, anomalies can be identified by their relatively low universal gravitation forces and low densities. The proposed gravitation theory based density peaks clustering approach is depicted in Algorithm 1.

Algorithm 1 The GDPC algorithm.

Require: Initial nodes $X \in R_{N\times M}$, d_c
Ensure: All clusters and anomalies are identified accurately
STEP 1. Calculate and sort node density
1.1 Calculate $d_{i,j}$ from X by distance formula
1.2 Sort $d_{i,j}$ in the ascending order
1.3 Determine d_c value with principles of DPC
1.4 Calculate ρ_i based on Equation (2) or (3)
1.5 Put all ρ_i in the descending order
STEP 2. Generate universal gravitation decision graph
2.1 Compute universal gravitation force based on Equation (6)
2.2 Form the decision graph with density ρ and the reciprocal of universal gravitation force F
2.3 Choose centroids and anomalies from the decision graph
STEP 3. Aggregate clusters with universal gravitation force
3.1 Put anomalies to the special cluster
3.2 Assign each node with cluster Id by its value of the reciprocal of universal gravitation force F
3.3 Iterate until all nodes are clustered

3. Experimental results

To test its feasibility and effectiveness of the GDPC algorithm, we compare it with k-Means [25], AP [2] and DPC [17] in the above four UCI data sets and the next three synthetic data sets listed in Table 3.



(a) 3 centroids. (b) 2 centroids.

Fig. 2. Difficult to determine the number of centroids in the decision graph of DPC with the *Iris* data set.

Table 4
F-Measure with seven different data sets.

Data sets	*k*-Means	AP	DPC	GDPC
Iris	0.8208	0.4851	0.7715	0.7715
Seeds	0.8068	0.3877	0.8026	0.8169
Wine	0.5835	0.3142	0.5892	0.6494
Glass	0.5052	0.2874	0.5418	0.5427
Flame	0.7364	0.2874	1	1
Aggregation	0.7725	0.3429	1	1
Spiral	0.3277	0.2853	0.7795	0.7795

3.1. F-Measure evaluation

F-Measure can be defined in Eq. (10), which is an index to evaluate the performance of clustering results. Seven different data sets (*Iris*, *Seeds*, *Wine*, *Glass*, *Flame*, *Aggregation* and *Spiral*) are selected to evaluate the clustering performance of *k*-Means, AP, DPC and GDPC in Table 4.

F-Measure index measures the accuracy of clustering result. It considers both the precision P and the recall R of algorithms: P is the ratio of the number of correct results to the number of all returned results, and R is the ratio of the number of correct results to the number of results. P, R and FM are defined as the following Eqs. (7), (8), (9) and (10):

$$P = (M_j, C_i) = \frac{|\, M_j \cap C_i \,|}{|\, C_i \,|}, \tag{7}$$

$$R = (M_j, C_i) = \frac{|\, M_j \cap C_i \,|}{|\, M_j \,|}, \tag{8}$$

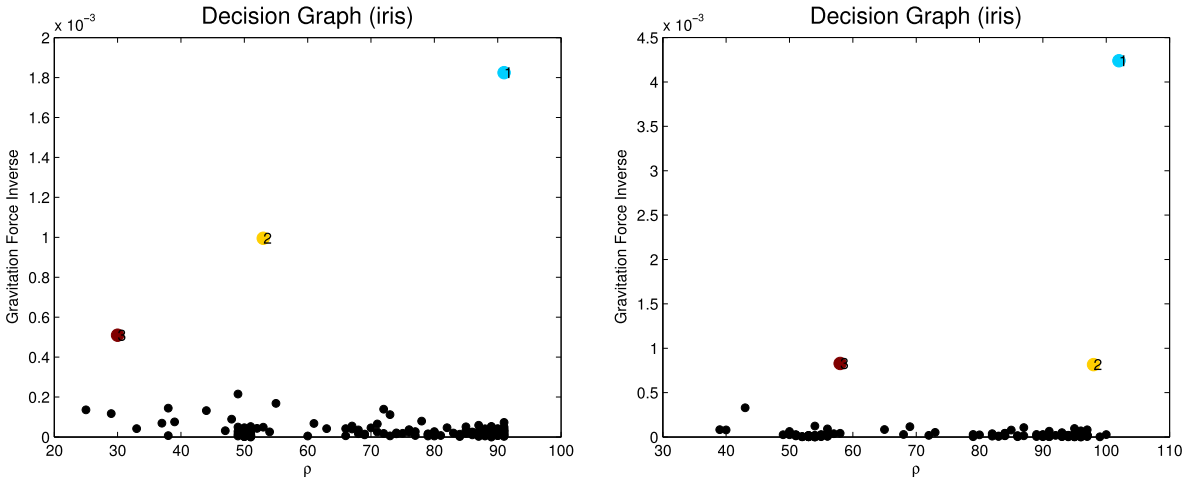
$$F\left(M_j, C_i\right) = \frac{2 \times P\left(M_j, C_i\right) \times R\left(M_j, C_i\right)}{P\left(M_j, C_i\right) + R\left(M_j, C_i\right)}, \tag{9}$$

$$FM = \sum_j \frac{|\, M_j \,|}{N} \times \max_i F\left(M_j, C_i\right). \tag{10}$$

Besides the clustering performance depicted in Table 4, numerous experiments have been done to evaluate the proposed GDPC capability with different clusters of varying sizes, varying densities, and identifying number of clusters and anomalies.

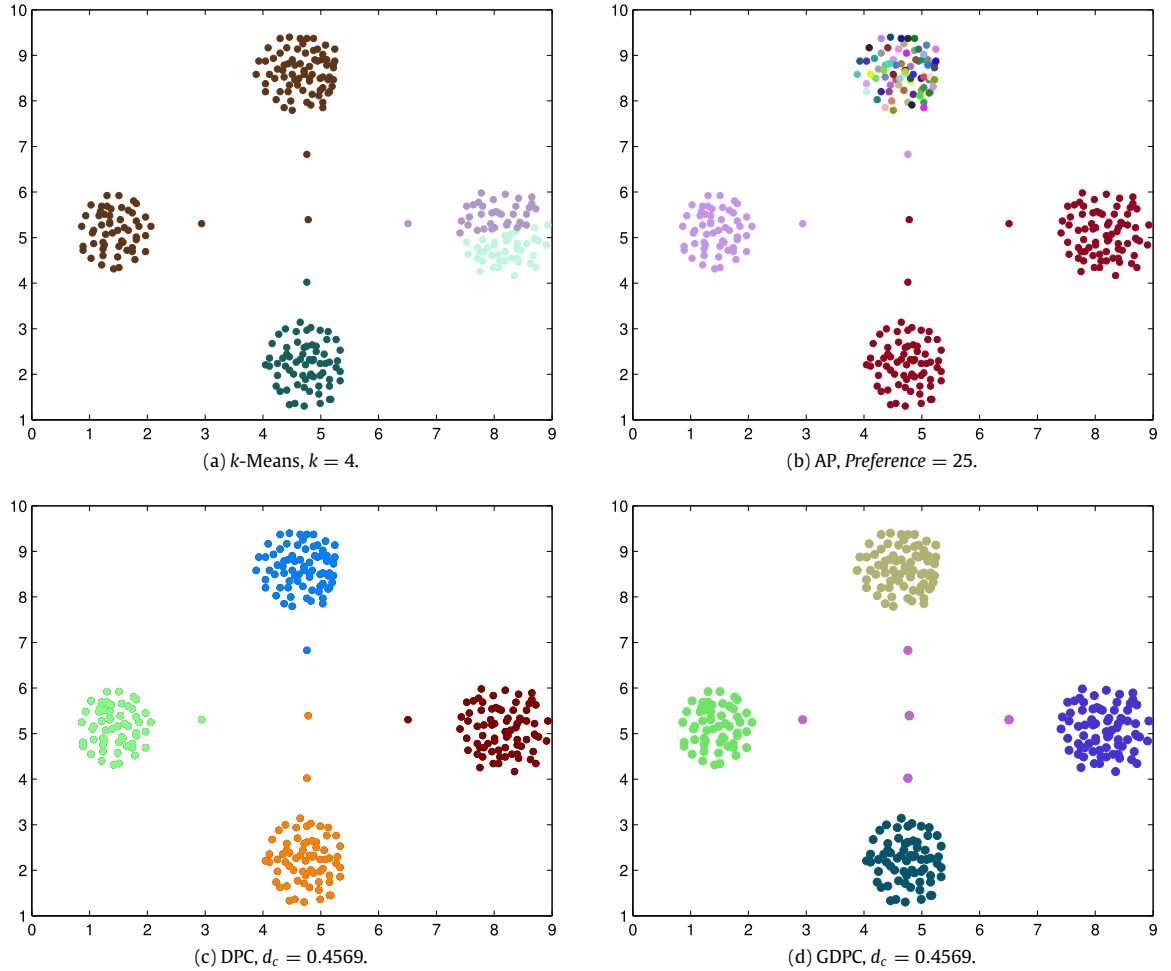
3.2. Detecting the number of clusters automatically

The proposed GDPC algorithm can identify centroids much easier than the DPC algorithm. As depicted in Fig. 3, the proposed GDPC algorithm can make more judicious decision graph. However, the DPC algorithm does not work well in the *Iris* data set, because it is difficult to select cluster centroids from the decision graph shown in Fig. 2. Nevertheless, it is easy to identify cluster centroids through GDPC illustrated in Fig. 3.



(a) $d_c = 1.9875$. (b) $d_c = 2.4372$.

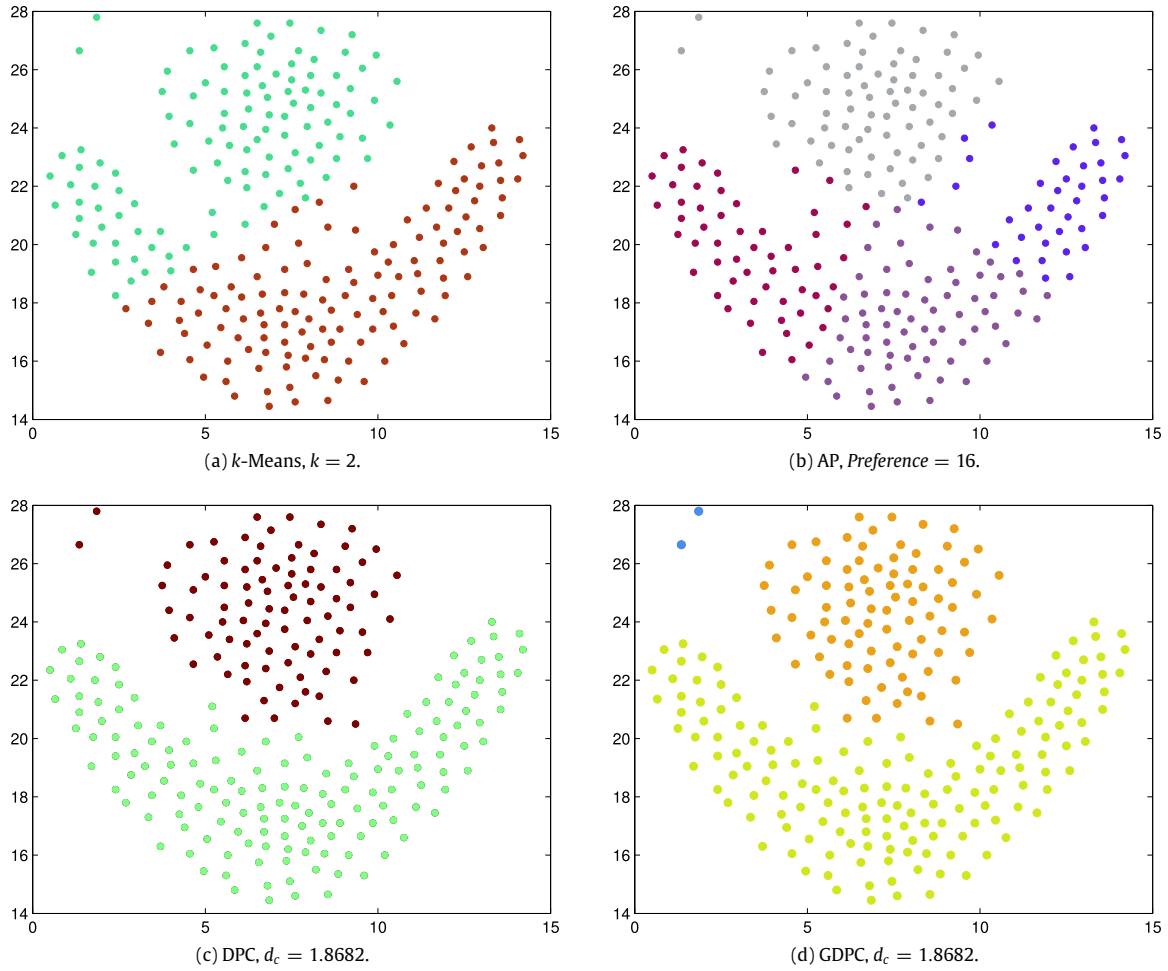
Fig. 3. Easy to determine the number of centroids in the decision graph of GDPC with the *Iris* data set.



(a) k-Means, $k = 4$. (b) AP, *Preference* $= 25$.

(c) DPC, $d_c = 0.4569$. (d) GDPC, $d_c = 0.4569$.

Fig. 4. Anomaly detection in the D_1 data set.



(a) k-Means, $k = 2$.

(b) AP, *Preference* $= 16$.

(c) DPC, $d_c = 1.8682$.

(d) GDPC, $d_c = 1.8682$.

Fig. 5. Anomaly detection in the *Flame* data set.

3.3. Anomaly detection accurately

The proposed GDPC has overcome the anomaly detection limitation of the DPC as shown in Figs. 1 and 4. In the DPC, anomalies are assigned to their nearby clusters incorrectly as illustrated in Figs. 1 and 4(c). However, our proposed GDPC algorithm can detect anomalies accurately as shown in Figs. 4(d) and 5.

3.4. Processing clusters of varying sizes

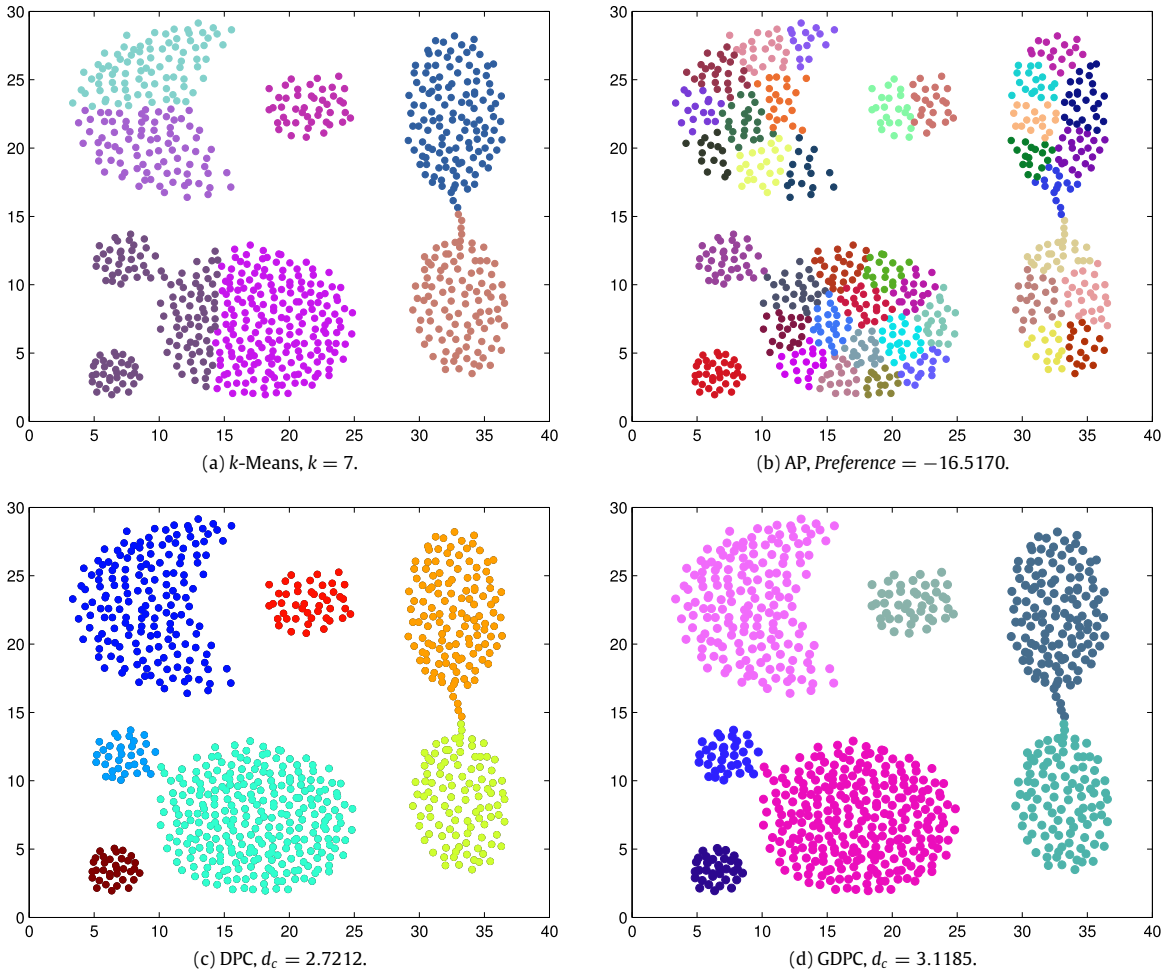
Both GDPC and DPC can aggregate clusters with varying sizes. The UCI data sets of *Aggregation* and *Flame*, and the user-defined data set of D_2 are selected to evaluate this competence. As illustrated in Figs. 5 and 6, both GDPC and DPC can find clusters accurately while k-Means and AP cannot cluster in an intuitive way.

3.5. Processing clusters of varying densities

Both GDPC and DPC can identify clusters of varying densities accurately. As comparison, Fig. 7 exemplifies the cluster assignations obtained by both k-Means [25] and AP [2] for this test case cannot attain a good performance. Even if both k-Means and AP optimization are performed with use of the correct parameters, the assignations are, in most of the cases, not compliant with visual intuition. Nevertheless, both the novel DPC and the proposed GDPC can handle clusters of varying densities in a good performance.

4. Discussion

In this paper, we propose a novel method of the enhanced capability of anomaly detection for the DPC algorithm on the basis of gravitation theory. GDPC employs universal gravitational force to identify centroids and anomalies. We have



(a) k-Means, $k = 7$.
(b) AP, *Preference* $= -16.5170$.
(c) DPC, $d_c = 2.7212$.
(d) GDPC, $d_c = 3.1185$.

Fig. 6. Clustering the *Aggregation* data set with varying sizes.

assessed our GDPC method on both synthetic data sets and some special purposed test data sets. When compared with clustering methods, experimental results demonstrate that the GDPC has the superior performance in its capability of:

- detecting number of clusters obviously;
- aggregating clusters with varying sizes, varying densities efficiently;
- finding anomalies accurately.

Exodus from the existing density-based clustering methods, our method espouses gravitation theory to improve the competence of the DPC method. We employed a different way, called alternative decision graph, to identify centroids and anomalies efficiently.

Espousing *gravitation force inverse* metric has comparative advantage in its decision graph. As illustrated in Fig. 8, all metrics of distance, gravitation force and *gravitation force inverse* can be adopted as an important decision factor in the decision graph. Amongst them, the range of *gravitation force* metric is too big, while both distance and *gravitation force inverse* have relative small range. Especially, the metric of *gravitation force inverse* can make nodes except centroids and anomalies have relative equal values, that is good for making decision when choosing centroids in its decision graph illustrated in Figs. 8 and 9. Consequently, the GDPC has relative advantage of detecting the number of clusters accurately.

The proposed gravitation theory based GDPC algorithm can identify anomalies accurately when compared with DPC, AP and k-Means as exemplified in Fig. 5. The DPC algorithm cannot detect anomalies when the distance between anomalies and relative higher density nodes is less than d_c. However, the metric of *gravitation force inverse* can reduce the influence from d_c value.

The GDPC has equal capability of processing clusters with vary sizes with the DPC. The aggregation rules in GDPC are similar to DPC. Analogous to the DPC algorithm, the proposed GDPC also has this capability by absorbing lower density nodes with gravitation force based on universal gravitation theory in Eq. (6) while using δ in the DPC.

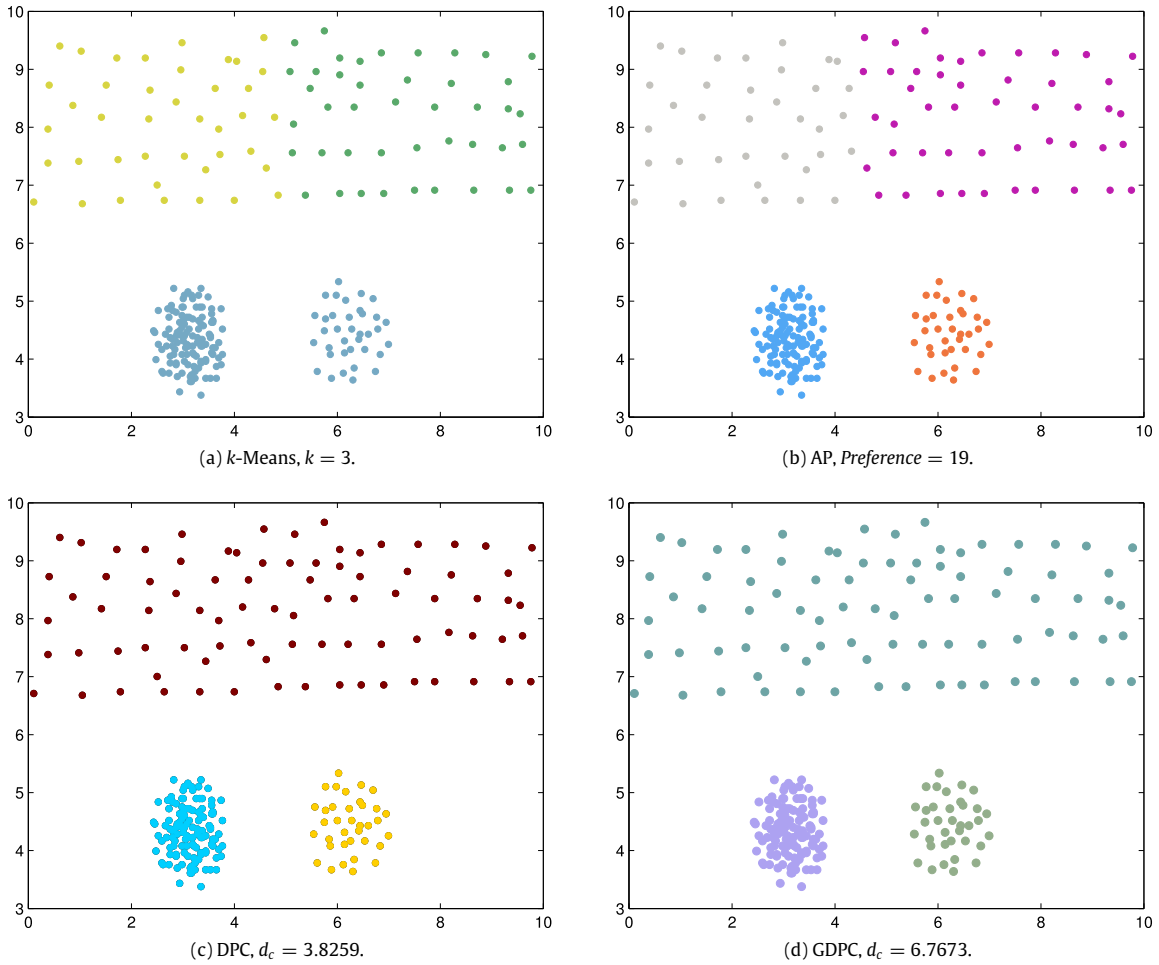


Fig. 7. Clustering the D_2 data set with varying densities.

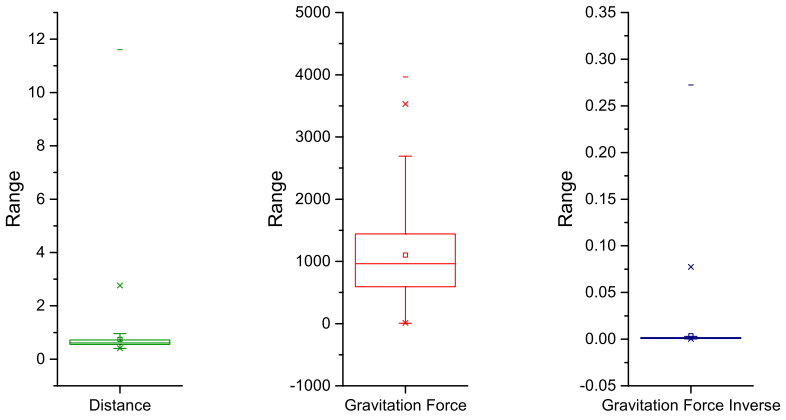
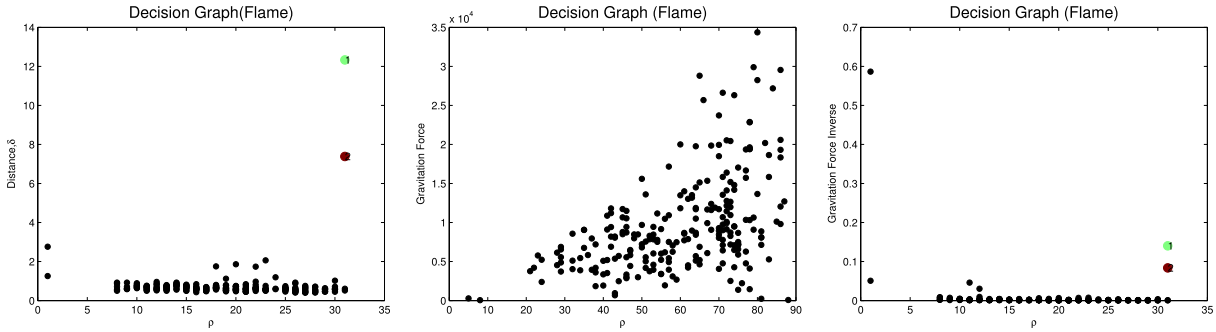


Fig. 8. Distribution of adopting different Y-axis metrics in the *Flame* data set.

Both GDPC and DPC can aggregate clusters with varying densities. It is really a challenging task to aggregate an accurate cluster with different densities based on principles in DBSCAN [16]. Both GDPC and DPC possess this capability because both of them do not implement density principles in the process of generating clusters. A cluster is formed by finding its centroid to absorb decreasing density nodes one by one in both GDPC and DPC.



(a) Distance. (b) Gravitation force. (c) Gravitation force inverse.

Fig. 9. Different metrics in the decision graph of the *Flame* data set.

5. Conclusions

The current novel density peaks clustering algorithm, proposed by Laio [17] in the journal *Science*, has the major limitation of anomaly detection. It is incapable of detecting anomalies as illustrated in Fig. 1. The reason behind this limitation is that there is a problem in adopting δ to evaluate its relationship among all the nodes. In Fig. 1, we can see that anomalies are not obvious when adopting δ to be distinguished.

Gravitation force inverse is a good alternative for δ in the decision graph shown in Graph 9. In the proposed GDPC algorithm, we redefined the gravitation force inspired by the principles of gravitation theory. With experiments, the proposed GDPC algorithm has two relatively good characteristics: (1) to detect the number of clusters obviously; (2) to identify anomalies accurately. Besides these, the GDPC accomplishes better clustering performances when compared with various classical methods, such as *k*-Means, AP and DPC, in several data sets.

Future studies will involve irregular shaped data processing, in addition to anomaly detection for clustering. Specially, low density node processing will be taken into account the gravitation theory.

References

[1] R. Xu, D.C. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 645–678.
[2] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (5814) (2007) 972–976.
[3] E.R. Hruschka, R.J. Campello, A.A. Freitas, A.D. Carvalho, A survey of evolutionary algorithms for clustering, IEEE Trans. Syst. Man Cybern. 39 (2) (2009) 133–155.
[4] A.A. Abbasi, M. Younis, A survey on clustering algorithms for wireless sensor networks, Comput. Commun. 30 (14) (2007) 2826–2841.
[5] S. Guha, R. Rastogi, K. Shim, Cure: an efficient clustering algorithm for large databases, Inf. Syst. 26 (1) (2001) 35–58.
[6] A.L. Strehl, J. Ghosh, Relationship-based clustering and visualization for high-dimensional data mining, Informs J. Comput. 15 (2) (2003) 208–230.
[7] Y. Cheng, Mean shift, mode seeking, and clustering, Pattern Recognit. Mach. Intell. 17 (8) (1995) 790–799.
[8] G. Dong, L. Tian, M.F.R. Du, H.E. Stanley, Analysis of percolation behaviors of clustered networks with partial support-dependence relations, Physica A 394 (2) (2014) 370–378.
[9] S. Shao, X. Huang, H.E. Stanley, S. Havlin, Robustness of a partially interdependent network formed of clustered networks, Phys. Rev. E 89 (3) (2014) 032812.
[10] A.R. Benson, D.F. Gleich, J. Leskovec, Higher-order organization of complex networks, Science 353 (6295) (2016) 163–166.
[11] M.J. Brusco, H. Kohn, Comment on "Clustering by passing messages between data points", Science 319 (5864) (2008) 726.
[12] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.
[13] O. Seref, Y. Fan, W.A. Chaovalitwongse, Mathematical programming formulations and algorithms for discrete *k*-median clustering of time-series data, Informs J. Comput. 26 (1) (2013) 160–172.
[14] J. Sander, M. Ester, H. Kriegel, X. Xu, Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications, Data Min. Knowl. Discov. 2 (2) (1998) 169–194.
[15] L. Ertöz, M. Steinbach, V. Kumar, Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data, in: Proceedings of SIAM International Conference on Data Mining, San Francisco, CA, USA, 2003, pp. 47–58.
[16] M. Ester, H. Kriegel, J. Sander, X. Xu, A density based algorithm for discovering clusters in large spatial databases with noise, Data Min. Knowl. Discov. 1996 (1996) 226–231.
[17] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.
[18] M. Du, S. Ding, H. Jia, Study on density peaks clustering based on *k*-nearest neighbors and principal component analysis, Knowl. Based Systems 99 (2016) 135–145.
[19] M. Wang, W. Zuo, Y. Wang, An improved density peaks-based clustering method for social circle discovery in social networks, Neurocomputing 179 (2016) 219–227.
[20] Y. Chen, D. Lai, H. Qi, J. Wu, J. Du, A new method to estimate ages of facial image for large database, Multimedia Tools Appl. 75 (5) (2016) 2877–2895.
[21] C. Wiwie, J. Baumbach, R. Rottger, Comparing the performance of biomedical clustering methods, Nature Methods 12 (11) (2015) 1033–1038.

[22] K. Sun, X. Geng, L. Ji, Exemplar component analysis: A fast band selection method for hyperspectral imagery, IEEE Geosci. Remote Sens. Lett. 12 (5) (2015) 998–1002.
[23] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (3) (2009) 1–72.
[24] H.C. Yung, H. Lai, Segmentation of color images based on the gravitational clustering concept, Opt. Eng. 37 (3) (1998) 989–1000.
[25] J.A. Hartigan, M.A. Wong, Algorithm AS 136: A *k*-means clustering algorithm, J. R. Stat. Soc. Ser. C. Appl. Stat. 28 (1) (1979) 100–108.