A novel hierarchical clustering approach based on data gravitation model

Peng Zhang¹,², Xiwang Xiang², Kun She¹*¹

¹ School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 610054, China
² School of Science, Southwest University of Science and Technology, Mianyang, Sichuan, 621010, China
*Corresponding author’s e-mail: kun@uestc.edu.cn

Abstract. The purpose of clustering is to group the data points into several clusters according to their similarity. Inspired by law of gravitation, each data point considered as an object is exposed in data space and influenced by the data gravitation of the other data points. In this paper, a novel hierarchical clustering approach (NGHC) is proposed based on local data gravitation. Firstly, the dataset is partitioned into some clusters as intermediate result by a simple gravity-based clustering approach. Then the intermediate clusters are merged by a new linkage measure which combines the data gravitations between two clusters until the satisfactory result is obtained. Two real-world datasets are used to validate the clustering performance of NGHC compared with five representative clustering algorithms. The results exhibit that the NGHC algorithm obtains the best performance on low- and high-dimensional datasets.

1. Introduction
Clustering is one of the most challenging tasks in Machine Learning. Numerous clustering algorithms have been proposed to group the data points into some clusters based on similarity between them, and widely applied in many fields such as data mining, pattern recognition and so on[1]. K-means and K-means++ are representative partition-based clustering approaches which need to specify the number of clusters in advance[2]. DBSCAN is a typical density-based clustering algorithm which can detect arbitrary shaped clusters and discard the points with lower density as noises[3]. BIRCH is one of the most effective hierarchical clustering methods based on tree data structure[3]. There are other classical and state-of-the-art clustering algorithms such as Meanshift, DP, Chameleon and so on. However, the similarities used by these algorithms are usually a scalar, such as Euler distance, cosine similarity, density and so on. Inspired by universal gravitational law, the gravity between objects can be considered as a new similarity which is a vector and contains more information[7]. On the one hand, gravitation force may be used to describe the similarity of two data points more exactly. The similarity gravitation force is determined by not only the mass (density) information of data point but also the distance information between two data points. On the other hand, the linkage measure of hierarchical clustering algorithm can be defined with the principle of combining the gravity force, which is different from traditional linkage measures such as complete-linkage, single-linkage and so on. Therefore, we proposed a novel hierarchical clustering algorithm which sufficiently utilizes the vector characteristics of gravitation force.
The remainder of this paper is organized as follows. In section 2, we introduce the data gravitation model and the proposed novel gravity-based hierarchical clustering algorithm. In section 3, two real world datasets are utilized to validate the performance of the proposed approach. And the results show that the proposed approach is superior to the contrast clustering algorithms. Finally, the conclusions are drawn in section 4.

2. Model and Algorithm

2.1. Data gravitation model and simple gravity-based clustering approach

Inspired by the Newton’s law of universal gravitation, we assume that all data points of datasets locate in a data space. Each data point is exerted the data gravitation by the other data points. The data gravitation is similar to the gravity between any two objects in the real world. The data gravitation is directly proportional to the product of two data points’ masses and inversely proportional to the square of the distance between the two data points. Mathematically, the data gravitation can be calculated as follow:

\[
F_{ij} = m_i m_j g(d_{ij}) \mathbf{u}_{ij} \quad d_{ij} > 0
\]

where \(m_i\) and \(m_j\) denote the mass of data point \(i\) and \(j\) respectively, \(d_{ij}\) denotes the distance between the two data points, \(\mathbf{u}_{ij}\) is the unit vector from data point \(i\) to data point \(j\), \(g(.)\) denotes an inverse proportion function of the distance between two points (e.g. \(g(x) = x^{-2}\)). If the distance between two data points equals to zero, the data gravitation is also equal to zero because the two points can be considered as the same data point. In order to calculate the mass of data point, we consider the mass can be replaced by density of which the definition is similar to the density used in DBSCAN. Mathematically, the mass of data point can be defined as:

\[
m_i = \sum_{\rho} \Gamma(d_{\rho i} - \gamma)
\]

where \(\Gamma(d) = 1\) if \(d = 0\) and \(\Gamma(d) = 0\) otherwise, \(\gamma\) is a cutoff threshold. Essentially, a data point’s mass equals to the number of points within a radius of the cutoff threshold and centered at the point. In nature, an object isn’t always significantly affected by the gravity exerted from other objects. So, we assume that data gravitation exerted on one data point is determined by \(k\) neighbor data points which bring the top \(k\)-largest-gravitation in data space. The resultant force of each data point can be computed by combining the data gravitation between it and its neighbors. Mathematically, the formula is derived as follows:

\[
\vec{F}_i = \sum_{j \in K_i} \vec{F}_{ij}
\]

where the set \(K_i\) consists of the neighbors of data point \(i\). The neighbors with larger masses provide more influence on the data point, and the neighbors with smaller distance to it also provide more influence. Obviously, there are more information between the gravitation forces and the resultant force which may be used to realize the data clustering.

**Algorithm 1** The simple gravity-based clustering algorithm (SGC).

**Input** dataset, \(\gamma\), \(\theta\), \(k\).

**Output** labels of clusters.

**Initialize** the number of clusters \(N\).

**for** \(i = 1; i \leq n; i = i + 1\) **do**

**for** \(j = 1; j \leq n; j = j + 1\) **do**

| Calculate the gravitation force \(M[i][j]\) of data points \(i\) and \(j\) by Eq. (1). |

**end for**

Select \(k\) data points with top \(k\)-largest gravitation force of data point \(i\).

Calculate the resultant force of data point \(i\) by Eq. (3).

**for** each data point \(d_p\) with top \(k\)-largest gravitation force **do**

| Calculate the GIC between data point \(i\) and data point \(d_p\) by Eq. (4). |

|
if \( GIC > \theta \) do
    Assigned the same label to data point \( i \) and \( d_p \).
end if
end for
end for
\[ \text{return labels of clusters.} \]

In order to estimate the correlation between data point and its adjacent point, we introduce a new coefficient called gravitational influence coefficient (GIC). The definition of GIC can be described as:

\[
GIC_{ij} = \frac{\vec{F}_i \cdot \vec{F}_j}{\|\vec{F}_i\| \|\vec{F}_j\|}
\]

where \( \vec{F}_i \) is the resultant force of data point \( i \), \( \vec{F}_j \) is the gravitation between data point \( i \) and its neighbor \( j \). The value of GIC ranges from -1 to 1. The bigger the GIC of two points is, the more similar they are. Intuitively, the GIC can be adopted into clustering approach to realize the data cluster analysis. The two data points with bigger GIC value are grouped into a cluster whereas two data points with smaller GIC value are grouped into different clusters. We bring in a threshold \( \theta \) here. If the GIC of two data points is greater than \( \theta \), they are grouped into the same sub-cluster. Then the dataset is partitioned into many sub-clusters after some iterations. However, these sub-clusters are only the intermediate result of the clustering model because they are not the genuine clusters of dataset. Above all, the clustering approach which utilizes gravitational influence coefficient to group the data points into different sub-clusters is called as simple gravity-based clustering approach named SGC for short.

For the sake of clarity, the algorithm of SGC is summarized in Algorithm 1.

**Algorithm 2** The gravity-based hierarchical clustering algorithm (NGHC).

**Input** dataset, \( \gamma \), \( \theta \), \( k \).

**Output** labels of clusters.

**Initialize** the number of clusters \( N \).

**Obtain** the intermediate clusters \( C = \{c_1, c_2, ..., c_m\} \) by Algorithm 1.

for each element \( c_i \) in \( C \) do
    for each element \( c_j \) in \( C \) do
        Calculate the GMC between cluster \( c_i \) and cluster by \( c_j \) Eq. (5).
    end for
    Merge the two clusters with largest GMC.
    Update the clusters set \( C \).
    if stop condition is satisfied do
        break
    end if
end for

Obtain the label of each cluster.
\[ \text{return labels of clusters.} \]

2.2. The proposed gravity-based hierarchical clustering algorithm

From the previous section, the performance of SGC is unsatisfactory though it can be used to cluster the data points. In order to boost the clustering performance, the clusters produced by SGC can be merged repeatedly until the satisfactory results have been obtained. The key is to define the linkage metric between two clusters to obtain the satisfactory clustering result. We adopt a novel linkage measure to determine the distance of two clusters by using the vector property of forces. The linkage criterion is named as gravitational merging coefficient (GMC). Mathematically, the GMC is formulated as follow:
\[ GMC(C_i, C_j) = \frac{1}{N_i \cdot N_j} \sum_{s \in C_i, t \in C_j} \tilde{F}_{st} \]  

where \( C_i \) and \( C_j \) are the \( i \)th and \( j \)th sub-graph, \( N_i \) is the number of data points in \( C_i \), \( N_j \) is the number of data points in \( C_j \). The linkage measure is computed by combining the gravitational forces from data points in \( C_i \) to those in \( C_j \). It each iteration, the two clusters with biggest GMC are merged to form a new cluster. The clustering process is terminated until the stop conditions are met. We called the new clustering algorithm as NGHC. The overall procedure of NGHC is described in Algorithm 2.

3. Experiments

To validate the superiority of the proposed approach, we conduct experiment on real world datasets compared with five representative clustering algorithms including K-means, K-means++, Spectral clustering(SC)[6], DBSCAN, and Birch. Meanwhile, there are four clustering performance criteria used to evaluate the performance of the proposed clustering approach and the contrast clustering algorithms in this paper. The performance metrics include the Purity, Rand Index (RI), F-measure, and Normalized Mutual Information (NMI)[5].

| Iris   | K-means | K-means++ | SC  | DBSCAN | Birch | NGHC |
|--------|---------|-----------|-----|--------|-------|------|
| RI     | 0.88    | 0.88      | 0.88| 0.77   | 0.82  | 0.97 |
| Purity | 0.89    | 0.89      | 0.89| 0.68   | 0.86  | 0.97 |
| F-measure | 0.82  | 0.82      | 0.82| 0.73   | 0.72  | 0.95 |
| NMI    | 0.76    | 0.76      | 0.75| 0.66   | 0.67  | 0.90 |

| SControl | K-means | K-means++ | SC  | DBSCAN | Birch | NGHC |
|----------|---------|-----------|-----|--------|-------|------|
| RI       | 0.84    | 0.85      | 0.70| 0.17   | 0.87  | 0.89 |
| Purity   | 0.64    | 0.64      | 0.23| 0.17   | 0.67  | 0.67 |
| F-measure | 0.61  | 0.63      | 0.19| 0.28   | 0.71  | 0.75 |
| NMI      | 0.69    | 0.71      | 0.01| 0.00   | 0.82  | 0.85 |

Two real world datasets Iris and SControl from the UCI Machine Learning Repository are adopted to validate the clustering performance of the proposed algorithm. The dataset Iris consists of 150 samples belonged to 3 classes with 4 features, which each class refers to a kind of iris plant. SControl contains 6 classes of 100 samples each, in which each sample includes 60 features. To obtain the best performance, the optimal threshold is sought for parameters of each different clustering approach. For K-means, K-means++ and SC, the parameter \( k \) is assigned to the number of classes in each dataset. The other optimal parameters with best Rand Index value for all above-mentioned algorithms are sought out by Grid Search method. For the parameter \( \sigma^2 \) of SC, it is set to 150 for dataset SControl and 0.5 for Iris. The parameter MinPts of DBSCAN is set to 10 for SControl and 4 for Iris. The parameters (threshold, branching_factor) of Birch are set to (0.5, 30) for SControl and (0.5, 20) for Iris. For the parameters (\( \gamma \), \( \theta \), \( k \)) of NGHC, they are set to (0, 0.3, 5) for SControl and (0.4, 0, 4) for Iris. The performance results of different clustering approaches on dataset Iris and SControl are tabulated in Table 1 and Table 2 respectively. It can be clearly seen that the performance metrics of proposed approach are much better than those of the other representative algorithms on dataset Iris. The proposed algorithm has the best performance except Purity metrics on dataset SControl. Especially, the dimension of SControl reaches 60. Above all, the proposed algorithm can obtain more excellent performance than the benchmark algorithms on these real world datasets.

4. Conclusions

In this paper, we proposed a novel hierarchical clustering algorithm (NGHC) which derives from the universal gravitation law. Firstly, we designed a coarse clustering approach (SGC) which utilizes the
relation between the resultant force and the gravitation force of data points to partition data points into many sub-clusters. However, the performance of SGC is not good enough to obtain the genuine clustering results. Even so, the result of SGC can be considered as intermediate result of NGHC. Then SGC algorithm is improved as a novel gravity-based hierarchical clustering approach which merges the intermediate clusters repeatedly until the satisfactory results are obtained. Finally, two real world datasets are adopted to validate the performance of NGHC compared with five representative clustering algorithms. The results show that NGHC algorithm is superior to the other contrast clustering algorithms. But the principle of the NGHC algorithm is not studied theoretically. Meanwhile, the time complexity of NGHC is still too high. The above-mentioned problems of NGHC can be solved in the future.

Acknowledgments
This work was supported by National Natural Science Foundation of China (11872323). The authors are very grateful to the editor and anonymous reviewers for their insightful comments.

References
[1] Adolfsson, A., Ackerman, M., & Brownstein, N.C. (2018) To cluster, or not to cluster: an analysis of clusterability methods. Pattern Recognition. 88:13-26.
[2] Zhao, W.L., Deng, C.H., & Ngo, C.W. (2018) K-means: a revisit. Neurocomputing, 291:195-206.
[3] Mahesh Kumar, K., & Rama, M.R.A. (2016) A fast dbscan clustering algorithm by accelerating neighbor searching using groups method. Pattern Recognition, 58:39-48.
[4] Zhang, T., Ramakrishnan, R., & Livny, M. (1997) Birch: a new data clustering algorithm and its applications. Data Mining and Knowledge Discovery, 1(2), 141-182.
[5] Wong, K. C., Peng, C., Li, Y., & Chan, T.M. (2014) Herd clustering: a synergistic data clustering approach using collective intelligence. Applied Soft Computing, 23(Complete), 61-75.
[6] Luxburg, U.V. (2007) A tutorial on spectral clustering. Statistics and Computing, 17(4), 395-416.
[7] Wang, Z., Yu, Z., Chen, C.L.P., You, J., Gu, T., & Wong, H. S., et al. (2017) Clustering by local gravitation. IEEE Transactions on Cybernetics, 1-14.