ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

Outward Statistical Algorithm In Clustering On Density Metrics

1.R.Kiruthika, M.Phil-Research Scholar, Department Of Computer Science, Govt Arts College, Ariyalur, Tamilnadu,India.
2.Dr.V.Vijayalakshmi, Research Supervisor,Head Department of Computer Science, Govt Arts College, Ariyalur, Tamilnadu,India.

Abstract—Clustering is one of the research hotspots in the field of data mining and has extensive applications in practice. Recently, Rodriguez and Laio [1] published a clustering algorithm on Science that identifies the clustering centers in an intuitive way and clusters objects efficiently and effectively. However, the algorithm is sensitive to a preassigned parameter and suffers from the identification of the "ideal" number of clusters. To overcome these shortages, this paper proposes a new clustering algorithm that can detect the clustering centers automatically via statistical testing. Specifically, the proposed algorithm first defines a new metric to measure the density of an object that is more robust to the preassigned parameter, further generates a metric to evaluate the centrality of each object. Afterwards, it identifies the objects with extremely large centrality metrics as the clustering centers via an outward statistical testing method. Finally, it groups the remaining objects into clusters containing their nearest neighbors with higher density. Extensive experiments are conducted over different kinds of clustering data sets to evaluate the performance of the proposed algorithm and compare with the algorithm in Science. The results show the effectiveness and robustness of the proposed algorithm.

Index Terms—Clustering, clustering center identification, long-tailed distribution, outward statistical testing

INTRODUCTION

CLUSTERING is an important technique of exploratory data mining, which divides a set of objects (instances or patterns) into several groups (also called clusters) in such a way that objects in same group are more similar with each other in some sense than with the objects in other groups. It has been widely used in different disciplines and applica-tions, such as machine learning, pattern recognition [2], data compression [3], image segmentation [4], [5], time series analysis [6], [7], information retrieval [8], [9], spatial data analysis [10], [11], [12] and biomedical research [13]. More-over, as data's variety and scale increase rapidly, and the prior knowledge (e.g., category or class label) about the data is usually limited, clustering has been a challenging task.

In this context, a number of clustering algorithms have been proposed based on different clustering mechanisms [14], [15], [16], [17], [18], such as i) the connectivity based clustering assumes that the objects close to each other are more possible to be in the same cluster than the objects far away from each other; this kind of clustering algorithms usually organizes the objects as a hierarchical structure but does not produce a unique partition, and still needs users to preassign a distance threshold to generate appropriate clus-ters. The representative algorithms include Single-Link [19] and Complete-Link [20]. ii) The centroid based clustering represents each cluster as a central vector (or named cluster-ing center), and the objects are assigned to the nearest clus-tering center, the famous examples are k-Means and its variants such as k-Medoids [21] and k-Means++ [22], where denotes the number preassigned by user of clusters. The requirement of the parameter k specified in advance is con-sidered as one of the critical drawbacks of this kind of algorithms

. Meanwhile, it is usually not able to detect the non-spherical clusters. iii) The distribution-based clustering assumes that the objects in a given cluster are most likely to be derived from the same distribution. The most famous example is EM (Expectation maximization) algorithm [23] which employs a fixed number of Gaussian distributions to approach the distribution of the objects. However, for most real world data sets, the real distribution of the objects is usually difficult to define in advance and cannot be con-cisely defined as Gaussian distribution. Moreover, this kind of clustering algorithms still needs to preassign the number of clusters (or different distributions). iv) The density based clustering defines the clusters as areas with higher density, and can detect the clusters in any arbitrary shape. The most popular example of density-based clustering is DBSCAN [24] in which only the objects whose density is greater than the given thresholds are connected together to form a clus-ter. However, the proper threshold setting varies with different data sets, there is still no effective method to preas-sign these thresholds. v) The spectral clustering based algo-rithm does not make assumptions on the forms of the clusters; it utilizes the spectrum (i.e., eigenvalues) of the similarity matrix of the data to map the data into a lower-dimensional space in which the objects can be easily clus-tered by traditional clustering techniques [18], [25], [26]. Comparing to the traditional algorithms, such as k-Means

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

and single-linkage, this kind of clustering algorithm is use-ful in non- convex boundaries and performs empirically very well [27]. And the first few eigenvalues can be used to determine the number of clusters and reduce the dimension of data. Yet, it has stated in
[28] these first eigenvectors can-not successfully cluster objects that contain structures with different sizes and densities.

Recently, Rodriguez and Laio [1] proposed a novel Clus-tering algorithm (denoted as RLClu for convenience in this paper) that integrates the merits of the above mentioned algorithms. First, similar to the connectivity and centroid based clustering, RLClu is only based on the distance (or similarity) between objects. Second, as the density based clustering, it defines the clustering centers as the objects with maximum local density, and can detect the non-spheri-cal clusters. Moreover, in contrast with the other well-known clustering algorithms (e.g., k-means, EM) where an objective function needs to be optimized iteratively, RLClu assigns the clustering label for each object in a single step.

The algorithm RLClu first defines two metrics (local den-sity and minimum density-based distance) for each object based on the distances among objects. Then, it constructs a two-dimensional plot (named as decision graph in RLClu) with these two metrics, and identifies the objects with both greater local density and minimum density-based distance as clustering centers via the decision graph. Finally, each of the remaining objects is assigned into the cluster including its nearest neighbor with a higher local density. However, there is still room for improving RLClu. First, the local den-sity plays a critical role in RLClu but is sensitive to a preas-signed parameter, cutoff distance, when the data set is small. Second, for clustering center identification, it still needs users to preassign two minimum thresholds of the local density and the minimum density-based distance. Dif-ferent threshold settings would result in different clustering results. The proper setting of these thresholds will vary with different clustering data sets. Consequently, as the other existing representative clustering algorithms (e.g., k-means, EM and DBSCAN), RLClu is also sensitive to some preassigned parameters and suffers from the parame-ter setting problem.

In order to address the shortages in RLClu, we propose a new clustering algorithm STClu (Statistical Test based Clus-tering)[1] in this paper. At first, we define a new metric to evaluate the local density of each object, which shows better performance in distinguishing different objects than the metric used in RLClu and is not so sensitive to the preas-signed parameter. Then, we employ an outward statistical test method to identify the clustering centers automatically on a centrality metric constructed based on the new local density and new minimum density-based distance. The experimental results on the synthetic and real world data sets show the proposed algorithm is more effective and robust than RLClu. In a nutshell, the proposed algorithm STClu obtains the object representation in a low-dimen-sional (specifically two dimensional) space in which the objects can be easily clustered. This idea is quite similar with that of spectral clustering in which the spectrum of the similarity matrix of the data is used for dimension reduction and the reduced space is not necessarily two-dimensional.

The rest of this paper is organized as follows. Section 2 reviews the related work of clustering. Section 3 presents the details of our clustering algorithm STClu. Section 4 gives the experimental results comparing our clustering algo-rithm to RLClu. Section 5 concludes our work.

2 Related Work

Traditionally, many researchers have proposed a number of clustering algorithms to divide objects into different cate-gories

on the basis of their similarity. Yet, there is still no unified definition of a cluster [1] since that we could get dif-ferent clusters with different clustering mechanisms. For centroid based clustering (such as k-Means), the objects are always grouped into the nearest clustering center. So this kind of algorithm works well on the data set with spherical clusters but is not able to detect the non-spherical clusters. The spectral clustering based algorithms first make use the spectrum of the similarity matrix to reduce the dimension of data, then perform clustering on the reduced data by tradi-tional clustering algorithms (e.g., k-Means). The distribution based clustering algorithm aims at reproducing the data with a set of predefined probability distribution functions; its performance depends on the number of distribution func-tions and the quality of these functions to approximate the implied distributions. The density based clustering algo-rithm usually can be used to identify the clusters in arbitrary shape. It defines clusters as connected dense regions in the data space. The well-known density based clustering algo-rithm is DBSCAN [24] which can not only detect non-spheri-cal clusters but also discard the noise in the data set.

Although the above algorithms can be used to explore the structures implied in a given data set, one challenge for these algorithms is that they need some proper parameter settings in advance. Otherwise, they might fail to find the true structures. Such as, the number of clusters and the ini-tial clustering centers for k-Means, the number of clusters for EM [23] and spectral clustering [18], the radius of epsi-lon-range-queries and the minimum number of objects required in an epsilon-range- query for DBSCAN [24], etc. That is, the performance of these clustering algorithms depends on the parameter

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

settings. Nevertheless, the proper settings will vary with the clustering data sets. In order to overcome the parameter setting problem, researchers have attempted to resort to some automatic (or parameter-free) clustering algorithms. These algorithms can automatically search for the proper parameters in a specific way or do not require users to specify the parameters in advance. Such as, for the problem of determining the "ideal" number of clus-ters which has been discussed for a while [29], [30] and is attracting ever growing interest recently [31], [32], [33], the researchers put forward different kinds methods includ-ing information-theoretic based [34], structure complexity based
[32] and recently quantization error based [33], the eigengap heuristic based for determining the number of clusters for spectral clustering [18], [35]. Meanwhile, some of these automatic clustering algorithms view the process of clustering as an optimization problem, and utilize differ-ent optimization strategies to search the optimal (or sub-optimal) partitions. In practice, the commonly-used optimi-zation strategy is stochastic search, such as evolutionary
algorithms (EA) [36], [37] and Simulated Annealing (SA) Algorithm [38] or their improvements [39], [40]. However, the performance of these search methods is related to choice of the fitness or energy function and proper parameter set-ting for optimization. For instance, the probabilities of cross-over and mutation.

he size of population for Generic Algorithm (GA), and the state space, the candidate genera-tor procedure, the acceptance probability function, and the annealing schedule temperature, and initial temperature for Simulated Annealing.

The clustering algorithm proposed by Rodriguez and Laio [1] gives an alternative approach which can detect the clustering centers from irregular shapes of clusters in an intuitive way. They construct a two-dimensional decision graph with two metrics (i.e., local density and minimum density-based distance), and the points located in the top right corner of this graph are more possible to be the cluster-ing centers (See details in Section 3.1). Once the clustering centers have been found, each one of the rest objects is grouped into the same cluster as its nearest neighbor with a higher density. This is completed in a single step and quite effective compared with other clustering algorithms (e.g., k-Means and EM) where an objective function needs to be optimized iteratively [23], [41].

However, for different data sets, the decision graphs are different as well. The local density used for decision graph construction is sensitive to a preassigned parameter (named cutoff distance) especially for small data sets. Moreover, although the algorithm can map the clustering centers into the top right corner of the decision graph, it still needs users to pick up proper number of objects from the decision graph artificially or set proper thresholds to determine the exact number of clustering centers in advance. There is no any straightforward method to handle the threshold setting problem (either for local density or the decision graph). Consequently, RLClu also suffers from the problem of how to determine the "ideal" number of clusters.

In this paper, we propose a novel clustering algorithm in which we first redefine the metrics of local density and mini- mum density-based distance with good robustness; then, instead of identifying the clustering centers by observing the decision graph artificially in RLClu, we detect the clustering centers by an outward statistical test method automatically on the basis of the redefined metrics. Extensive experiments demonstrate the effectiveness of the proposed algorithm.

EXISTING BASED CLUSTRING

1) Connectivity based clustering that assumes the objects close to each other possible to be in the same cluster than the objects far away from each other

2) Centroid based clustering assume the some object or a class is the center the nearest objects are formed using k- Medoids and k-Means++.

3) The distribution based a clustering assumes that objects in cluster are most likely to be derived from the same distribution using Expectation Maximization algorithm

4) The density based clustering used in higher density areas can detect cluster in any arbitrary shape.The spectral clustering based algorithm does not make assumptions on the forms of the cluster

PROPOSED BASED CLUSTERING

1) The proposed model using statistical test based clustering (STclu).At first we define a new metric to evaluate the local density of each object.

2) Then, we employ an outward statistical test method to identify the clustering centers automatically on a centrality metric constructed based on the new local density and new minimum density-based distance.

3) The proposed algorithm STClu obtains the object representation in a low-dimensional (specifically two dimensional) space in which the objects can be easily clustered.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

4) This idea is quite similar with that of spectral clustering in which the spectrum of the similarity matrix of the data is used for dimension reduction and the reduced space is not necessarily two-dimensional

3 OUTWARD STATISTICAL TESTING BASED CLUSTERING ALGORITHM

In this section, we first review the original clustering algorithm RLClu proposed in [1], and then discuss the shortages in RLClu yet to be resolved. Furthermore, we propose an outward statistical testing based clustering algorithm to relieve these shortages.

Review of the Clustering Algorithm RLClu

The clustering algorithm RLClu is proposed based on the assumption of "Cluster centers usually have a higher local density and a relative larger distance from objects with higher local densities" [1]. It consists of three steps: metric extraction, clustering center identification, and object clustering.

1) Metric extraction. For each of the n objects {O1, O2, . . ., On} being clustered, RLClu defines two metrics r and d to evaluate the local density of the given object and the minimum density-based distance between the given object and the other objects.

2) Clustering center identification. RLClu constructs a two-dimensional point (r_i, d_i) for each object and maps all these objects into a two-dimensional space, where the two-dimensional plot is referred to as a decision graph. In the decision graph, only points which are far away from both of the r-axis and d-axis are identified as the clustering centers, i.e., the objects with both high r_i and d_i. RLClu defines two minimum thresholds of r_{min} and d_{min} to identify the clustering centers.

3) Object clustering. This part is straightforward once the clustering centers are picked up. That is, for all the objects except for the clustering centers, each one is assigned to a cluster which contains its nearest neighbor with higher local density r.

According to the brief introduction of RLClu, we can get that the metrics r and d play important roles in RLClu. In order to further understand RLClu and analyze its drawbacks, we briefly introduce the metrics r and d in advance.

In RLClu, the local density of a given object O_i is defined by Definition 1.

Definition 1. Local density r,

$$r_i = \sum_{j} D(d_{i,j} - d_c). \quad (1)$$

Where $d_{i,j}$ denotes the distance between objects O_i and O_j. The distance can be Euclidean distance or any measure which can evaluate the difference between two objects, d_c is the cutoff distance preassigned by users. And $D(x) = 1$ if $x < 0$ and $D(x) = 0$ otherwise. From Definition 1, we can get that the local density of object O_i is the number of objects appearing in the hypersphere whose center is O_i and radius is d_c, i.e., the number of neighbors with distance to O_i being smaller than the cutoff distance d_c.

Based on the local density r, the minimum density-based distance d_i of O_i to any other object with higher density is defined as follows.

Definition 2. Minimum density-based distance d,

$$d_i = \min_{j: r_j > r_i} (d_{i,j}). \quad (2)$$

According to Definition 2, for a given object O_i, we can get a distance d_i which is the minimum distance between O_i and any other object with higher local density.

The Proposed Clustering Algorithm

In this section, we propose a novel clustering algorithm aiming at overcoming the shortages of RLClu in Section 3.2. In the proposed algorithm, we first put forward a new met-ric $\hat{r}$ to measure the density of an object, which shows better performance in terms of the ability to distinguish different objects and is more robust to the preassigned parameter than the local density r in RLClu. Meanwhile, on the basis of this new density metric, we redefine the minimum den-

Input: O {O1; O2; . . . ; On}: A set of n objects K: the number of nearest neighbors in K-density $\hat{r}$;
Output CLU;//A set of clusters
RhoSet f, DeltaSet f, NNSet f, GamaSet f;
//Part 1: Metric extraction
distanceMatrix DistanceFunction(O);//Calculate distance
RhoSet $F_{\hat{r}}$(distanceMatrix, k);//Calculate $\hat{r}$
[DeltaSet, NNSet] $F_{\hat{d}}$(distanceMatrix, RhoSet);//Calculate $\hat{d}$ and identify the nearest neighbor for each object
GamaSet RhoSet DeltaSet;//$\hat{g} = \hat{r}\hat{d}$
//Part 2: Clustering center identification
X sort(GamaSet, "descend");//Sort GamaSet in descending order to get a set of ordered statistics X
R {R_i $X_{i;n}/X_{i+1;n}$} (1 i n 1);
m 0; CLU f;//Start at the mth hypothesis
//Identify the number of clusters k by Outward

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

{g^1; g^2; . . . ; g^n} to evaluate the centrality of each object. The second part (lines 6-17) employs the outward statisti-cal testing method presented in Proposition 1 to identify

k clustering centers. First, by sorting the metrics of Gama- Set in descending order, STClu generates a set of ordered statistics X and further constructs a set of statistics R for

statistical testing. Then, starting at the mth hypothesis $H_{0;m}$, STClu identifies the first hypothesis $H_{0;k}$ rejected by

comparing the statistic R_i with the estimated critical value r_i. Finally, the number of clustering centers is set as

k and the objects corresponding to the first k hypotheses $\{H_{0;1}; H_{0;2}; \ldots; H_{0;k}\}$ are detected as the clustering centers $\{c_1; c_2; \ldots; c_k\}$. In the third part (lines 18-24), for each object being not the clustering centers, STClu clusters it

into the group containing its nearest neighbor with higher sity-based distance of an object as a new version . With d

these two new metrics rho and d, we weigh the possibility of an object being a clustering center by a new centrality

metric * which is the product of * and ^ due to the fact that

g r d

the clustering centers usually have both of higher density (measured by r^) and larger distance from each other (mea-

sured by). The objects with extremely large gamma^ are rec-d ognized as the clustering centers. With this in mind, afterwards, by analyzing the distribution of this product metric g^, we transform the problem of clustering center identification into a problem of extreme-value detection from a long-tailed distribution, and employ an outward sta-tistical testing method to identify the clustering centers automatically. Finally, we accomplish the clustering process by assigning proper clustering labels to the remaining objects based on these identified clustering centers.

3.2.1 Statistical Test Based Clustering Algorithm

In this section, we present the proposed algorithm STClu (Statistical Test based Clustering), which is implemented on the basis of the metrics defined in Section 3.3.1 and the clustering center identification method introduced in Sec-tion 3.3.2. Algorithm 1 shows the pseudo-code descrip-tion of STClu.

The pseudo-code consists of three parts: i) metric extraction; ii) clustering center identification; and iii) object clustering. In the first part (lines 1-5), by calculating

r c

and now minimum distance

i i each object O_i, STClu gets a set of metrics GamaSet =

K- density. After that, CLU ¼ fClui; 1 i kg records the

k clusters found by STClu, where each cluster Clu_i (1 i k) is non-empty and contains at least one object (e.g., clustering

center c_i) and each object belongs to exactly one cluster. Therefore, the proposed algorithm STClu is a kind of partitional clustering.

In algorithm STClu, the number of nearest neighbors K is associated with the calculation of K-density, it is set to be a

pfffffffi

default value d ne which is usually adequate in most of the situations according to the sensitivity analysis of K on the performance of STClu in Section 4.3.

Algorithm 1. Outward Statistical Testing based Cluster-ing

Algorithm STClu

statistical testing
9 while m > 2 do
10 Calculate the critical value r_m according to Eq. (7);
11 if $R_m > r_m$ then
12 k m;
13 break;
14 end
15 m m 1;
16 end
17 Identify the objects corresponding to $\{R_1; R_2; \ldots; R_k\}$ as the clustering centers $\{c_1; c_2; \ldots; c_k\}$, and label c_i as i; //Part 3: Object clustering
18 for i 1 to n do
19 if O_i is unlabeled then

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

20 Mark O_i the label of its nearest neighbor with higher r^ according to NNSet;
21 end
22 end
23 CLU fClu$_i$; 1 i kg, where Clu$_i$ denotes the set of objects with label i;
24 return;

4 Experimental Study

In this section, we experimentally evaluate the performance of the proposed clustering algorithm with representative clustering data. At first, we introduce the benchmark clus-tering data sets in Section 4.1, and then present the experi-mental results and analyses in Section 4.2. Finally, we conduct the sensitive analysis of STClu in Section 4.3.

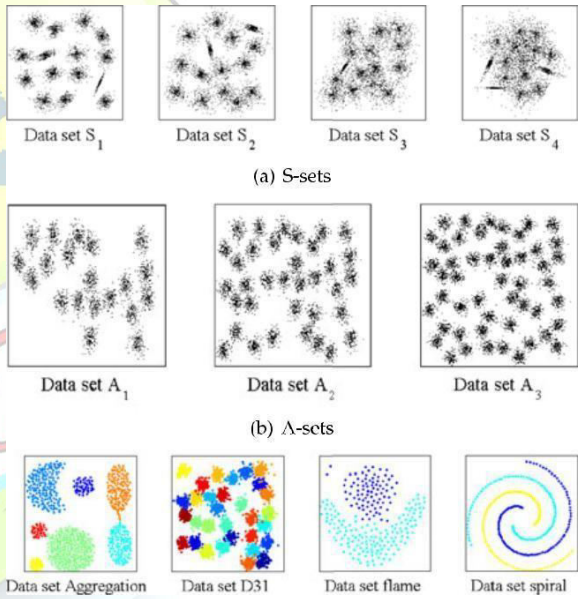
4.1 Benchmark Data Sets

Five groups of representative clustering data sets (e.g, includ-ing low and high dimensional data sets, synthetic and real world data sets) are employed as the benchmark to assess the performance of the proposed algorithm. These data sets are available on http://cs.joensuu.fi/sipu/datasets/and http:// people.sissa.it/ laio/Research/Res_clustering.php. The cor-rect clustering centers of the data sets are known in advance. The details of these data sets are introduced as follows.

1) S-sets: two-dimensional data sets with 5,000 objects and 15 Gaussian clusters with four different degree of clustering overlapping [47]. See Fig. 4a for details. This kind of data can be used to evaluate the robust-ness of the proposed algorithm. The degree of the clustering overlapping increases from data set S1 to S4. The greater the degree of overlapping, the more difficult to distinguish different clusters.
2) A-sets: two-dimensional data sets with varying num-ber of clusters (20, 35 and 50 for A1, A2 and A3), and there are 150 objects per cluster [48]. See Fig. 4b for details. This kind of data can be used to evaluate the scalability of the proposed algorithm in detecting different numbers of clusters.
3) Shape sets: two-dimensional data sets (named Aggre-gations, D31, flame and Spiral) represent some difficult clustering objects because they contain clus-ters of arbitrary shape, proximity, orientation and varying densities [49], [50], [51], [52]. The number of objects in these four data sets is 788 for Aggregations, 3,100 for D31, 240 for flame and 312 for Spiral, respectively. See Fig. 4c for details. This kind of data can be used to evaluate the proposed algorithm in detecting the clustering centers in complex cluster-ing data sets. Meanwhile, the four data sets in Fig. 4c have also been used to assess the algorithm RLClu.
4) High-dimensional data sets: six high-dimensional data sets with 1,000 objects and 16 Gaussian clusters in different dimensions [53]. The dimension of these six data sets is 32, 64, 128, 256, 512 and 1,024, respec-tively. Each data set with dimension x is named "Dimx". This kind of data can be used to assess the performance of the clustering algorithms when the dimension of the data increases.
5) Real world data sets: the Face detection database including 400 figures with 40 people. This data set proposes a serious challenge to the algorithm RLClu

since the real number of clusters is comparable with the number of objects in each cluster (10 different pictures for each people).

Fig. 4. Benchmark data sets



5 Conclusion

In this paper, we have proposed a statistical test based clus-tering algorithm (STClu) that can automatically identify the clustering centers and further cluster the objects in an effec-tive way. We first defined a new metric, K-density r^, to measure the local density of each object. Based on K-den-sity, we established a new metric ^ to evaluate the distance d of an object to its neighbors with higher density. Then, a product of these two metrics ^ ¼ ^ ^ was used to evaluate the centrality of each object. Afterwards, by analyzing the distribution of these metrics, we found that g^ could be rep-resented by a long-tailed distribution, and further trans-formed

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

the clustering center identification into a problem of extreme- value detection from a long-tailed distribution. Finally, we employed an outward statistical testing method to detect the clustering centers with g^ automatically and then completed the clustering process by assigning each of the rest objects to the cluster that contains its nearest neigh-bor with higher K-density. Extensive experiments have been conducted on both synthetic and real world data sets; the experimental results show the effectiveness and robust-ness of the proposed algorithm STClu.

6 Acknowledgments

The authors would like to thank the editors and the anony-mous reviewers for their insightful and helpful comments and suggestions, which resulted in substantial improve-ments to this work.

References

[1] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," Science, vol. 344, no. 6191, pp. 1492–1496, 2014.

[2] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Sil-verman, and A. Y. Wu, "An efficient k-means clustering algo-rithm: Analysis and implementation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 881–892, Jul. 2002.

[3] A. Gersho and R. M. Gray, Vector Quantization and Signal Compres-sion. New York, NY, USA: Springer, 1992.

[4] A. Shama and S. Phadikar, "Automatic color image segmentation using spatial constraint based clustering," in Emerging Trends in Computing and Communication. New York, NY, USA: Springer, 2014, pp. 113–121.

[5] G. Dong and M. Xie, "Color clustering and learning for image seg-mentation based on neural networks," IEEE Trans. Neural Netw., vol. 16, no. 4, pp. 925–936, Jul. 2005.

[6] T. W. Liao, "Clustering of time series data-a survey," Pattern Recog., vol. 38, no. 11, pp. 1857–1874, Nov. 2005.

[7] C.-P. Lai, P.-C. Chung, and V. S. Tseng, "A novel two-level clus-tering method for time series data analysis," Expert Syst. Appl., vol. 37, no. 9, pp. 6319–6326, 2010.

[8] N. Jardine and C. J. V. Rijsbergen, "The use of hierarchic cluster-ing in information retrieval," Inf. Storage Retrieval, vol. 7, pp. 217– 240, 1971.

[9] H. Xiong, W. Wu, and S. Shekhar, Clustering and Information Retrieval. Norwell, MA, USA: Kluwer, 2003.

[10] V. Estivill-Castro and I. Lee, "Argument free clustering for large spatial point-data sets via boundary extraction from delaunay dia-gram," Comput., Environ. Urban Syst., vol. 26, no. 4, pp. 315–334, 2002.

[11] W. Cui and X. Yang, "A novel spatial clustering algorithm based on delaunay triangulation," J. Softw. Eng. Appl., vol. 3, pp. 141– 149, 2010.

[12] D. Liu and O. Sourina, "Free-parameters clustering of spatial data with non-uniform density," in Proc. IEEE Conf. Cybern. Intell. Syst., 2004, pp. 387–392.

[13] R. Xu and D. C. Wunsch, "Clustering algorithms in biomedical research: A review," IEEE Rev. Biomed. Eng., vol. 3, pp. 120–154, 2010.

[14] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," ACM Comput. Surveys, vol. 31, pp. 264–323, 1999.

[15] P. Berkhin, "A survey of clustering data mining techniques," in Grouping Multidimensional Data. New York, NY, USA: Springer, 2006, pp. 25–71.

[16] A. K. Jain, "Data clustering: 50 years beyond k-means," Pattern Recog. Lett., vol. 31, no. 8, pp. 651–666, 2010.

[17] P. Rai and S. Singh, "A survey of clustering techniques," Int. J. Comput. Appl., vol. 7, no. 12, pp. 156–162, 2010.

[18] U. Von Luxburg, "A tutorial on spectral clustering," Statist. Com-put., vol. 17, no. 4, pp. 395–416, 2007.

[19] P. H. A. Sneath and R. R. Sokal, "Numerical Taxonomy," Nature, vol. 193, pp. 855–860, 1962.

[20] B. King, "Step-wise clustering procedures," J. The Am. Statist. Assoc., vol. 62, pp. 86–101, 1967.

[21] P. S. Bradley, O. L. Mangasarian, and W. N. Street, "Clustering via concave minimization," in Proc. Adv. Neural Inf. Process. Syst., 1997, pp. 368–374.

[22] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms, 2007, pp. 1027–1035.

[23] G. Mclanchan and T. Krishan, "The em algorithm and exten-sions," Series Probability Statist., vol. 15, no. 1, pp. 154–156, 1997.

[24] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. Int. Conf. Knowl. Discovery Data Mining, 1996, vol. 96, no. 34, pp. 226–231.

[25] W. E. Donath and A. J. Hoffman, "Lower bounds for the partition-ing of graphs," IBM J. Res. Develop., vol. 17, no. 5, pp. 420–425, 1973.

[26] L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 11, no. 9, pp. 1074–1085, Sep. 1992.

[27] A. Y. Ng, M. I. Jordan, Y. Weiss, et al., "On spectral clustering: Analysis and an algorithm," Adv. Neural Inf. Process. Syst., vol. 2, pp. 849–856, 2002.

[28] B. Nadler and M. Galun, "Fundamental limitations of spectral clustering," in Proc. Adv. Neural Inf. Process. Syst., 2006, pp. 1017–1024.

[29] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," Psychome-trika, vol. 50, no. 2, pp. 159–179, 1985.

[30] C. Biernacki and G. Govaert, "Using the classification likelihood to choose the number of clusters," Comput. Sci. Statist., vol. 29, pp. 451–457, 1997

[31] M. M.-T. Chiang and B. Mirkin, "Intelligent choice of the number of clusters in k-means clustering: an experimental study with dif-ferent cluster spreads," J. Classification, vol. 27, no. 1, pp. 3–40, 2010.

[32] B. Mirkin, "Choosing the number of clusters," Wiley Interdisciplin-ary Rev.: Data Mining Knowl. Discovery, vol. 1, no. 3, pp. 252–260, 2011.

[33] A. Kolesnikov, E. Trichina, and T. Kauranne, "Estimating the number of clusters in a numerical data set via quantization error modeling," Pattern Recog., vol. 48, no. 3, pp. 941–952, 2015.

[34] G. James and C. Sugar, "Finding the number of clusters in a data-set: An information-theoretic approach," J. Am. Statist. Assoc., vol. 98, no. 463, pp. 750–764, 2003.

[35] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in Proc. Adv. Neural Inf. Process. Syst., 2004, pp. 1601–1608.

[36] D. Doval, S. Mancoridis, and B. Mitchell, "Automatic clustering of software systems using a genetic algorithm," in Proc. Int. Workshop Softw. Technol. Eng. Practice, 1999, vol. 0, p. 73.

[37] H. He and Y. Tan, "A two-stage genetic algorithm for automatic clustering," Neurocomputing, vol. 81, no. 1, pp. 49–59, 2012.

[38] S. Saha and S. Bandyopadhyay, "A generalized automatic cluster-ing algorithm in a multiobjective framework," Appl. Soft Comput., vol. 13,

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

no. 1, pp. 89–108, 2013.

[39] K. K. Pavan, A. A. Rao, and A. Rao, "An automatic clustering technique for optimal clusters," Int. J. Compu. Sci. Appl., vol. 1, pp. 133–144, 2011.

[40] S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm," IEEE Trans. Syst., Man, Cybern., vol. 38, no. pp. 218–237, Jan. 2008.

[41] A. Vattani, "k-means requires exponentially many iterations even in the plane," Discrete Comput. Geom., vol. 45, no. 4, pp. 596–616, 2011.

[42] W. A. Trybulec, "Pigeon hole principle," J. Formalized Math., vol. 2, no. 199, pp. 1–5, 1990.

[43] Bauldry and C. William, Introduction to Real Analysis. Hoboken, NJ, USA: Wiley, 2011.

[44] W. H. Rogers and J. W. Tukey, "Understanding some long-tailed symmetrical distributions," Statistica Neerlandica, vol. 26, no. 3, pp. 211–226, 1972.

[45] S. Foss, D. Korshunov, and S. Zachary, An Introduction to Heavy-Tailed and Subexponential Distributions. New York, NY, USA: Springer,

2

011.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018