Analysis of Spatial Clustering Optimization

YANG Jianfeng  YAN Puliu  XIA Delin  GENG Qing

Abstract  Spatial clustering is widely used in many fields such as WSN (Wireless Sensor Networks), web clustering, remote sensing and so on for discovery groups and to identify interesting distributions in the underlying database. By discussing the relationships between the optimal clustering and the initial seeds, a clustering validity index and the principle of seeking initial seeds were proposed, and on this principle we recommend an initial seed-seeking strategy: SSPG (Single-Shortest-Path Graph). With SSPG strategy used in clustering algorithms, we find that the result of clustering is optimized with more probability. At the end of the paper, according to the combinational theory of optimization, a method is proposed to obtain optimal reference k value of cluster number, and is proven to be efficient.

Keywords  data mining; spatial clustering; SSPG; clustering optimization

CLC number  P208

Introduction

Clustering, in data mining (DM), is one of the efficient ways to discover the characteristics of an underlying spatial database and is widely used recently in many fields such as Information Retrieval (IR), web cluster, remote sensing (RS), wireless sensor network (WSN),[1, 2] etc. Cluster analysis is a set of methodologies for automatic classification of a given d-dimensional database into k groups, and the data in one group are similar while the data belonging to different groups are dissimilar.

In general, within-cluster (WC) and between-cluster (BC), as shown in Eq.(2) of subsection 2.2, are used to measure the similarity and dissimilarity of the clustering result, respectively. The procedure of clustering aims to find the numbers of (usually k, as an input variant of some clustering algorithms) optimal clusters which minimized WC and maximized BC simultaneously[3]. Note that: 1) for some clustering algorithms such as k-means or its ameliorations, it is necessary for the user to determine k; but sometimes it is difficult to do so; 2) even if a suitable k is given, some spatial clustering algorithms use a certain criterion function—the square-error criterion (SE) to judge whether the result is “good” or not, which is not accurate.

In our experiments, we find that different initial seeds sets will affect the clustering results dramatically. Therefore, optimal clustering can only be brought about with some of the initial seeds sets in all possible sets. For this reason we propose a $\theta_{\text{max}}$ principle to find “good” initial seeds and the SSPG is proposed subsequently to obtain these seeds. With SSPG strategy used for some data sets, we find it is possible to get an optimal reference k without any artificial influence.

Received on July 20, 2008.
Supported by the National Natural Science Foundation of China (No.60502028, No. 90204008).
YANG Jianfeng, School of Electronic Information, Wuhan University, 129 Luoyu Road, Wuhan 430079, China.
E-mail: yjf@whu.edu.cn


1 k-means review

The k-means clustering algorithm\cite{4} has been used widely and popularly in spatial clustering for its simplicity and efficiency. With a given data points set $\mathbf{D} = \{p_1, p_2, \cdots, p_n\}$ in $d$-dimensional metric space, and $k$ (the number of clusters), k-means algorithm attempts to minimize the square-error (SE) as shown in Eq.(1); here, $m_i$ is the mean of $i$th cluster, $\|p_j - m_i\|$ represents the Euclidean distance between those two points, and $[c_i]$ stands for the number of points in $i$th cluster. The SE minimization can compact all objects in one cluster or even more.

$$SE = \sum_{i=1}^{k} \sum_{p_j \in c_i} \|p_j - m_i\|$$

$$m_i = \frac{1}{|p_j \in c_i|} \sum_{p_j \in c_i} p_j$$

The k-means algorithm can generate good result with spherical or convex dataset, and some ameliorative algorithms, such as k-Medoids\cite{5}, BIRCH\cite{6}, R-tree\cite{7}, etc\cite{8}, are also proposed to make the procedure more efficient or to optimize the result.

Since k-means algorithm is not sensitive to the outliers (i.e. objects that are very far away from the rest of the objects), we dug the outliers out from the dataset used in this paper. Some algorithms to deal with outliers or noises are brought forward in References \cite{9,10}.

2 Clustering optimization

2.1 Discussions

In the procedure of k-means algorithm, $k$ must be an input variant, but sometimes, it is a big challenge for users to determine a suitable $k$ for lack of any prior knowledge of the given dataset. The larger the $k$ number is, the more groups the dataset will be divided into, and then the final clustering result will blur the data characteristics that users can extract from the original dataset with more complexity. Otherwise, if $k$ is too small, the information we can get from the dataset will be less.

Some papers try to estimate the optimal $k$ without any prior knowledge of the given dataset. Reference \cite{8} researched on the rationality of clustering model and used $s_k = BC_k + WC_k$ as cost function, exhausts all possible $k$ values and started with $k$ numbers of initial seeds sought evenly, then computed the $s_k$ to find which $k$ can minimize $s_k$, and drew a conclusion that the coincident $k$ is the optimal estimation of the numbers of clusters. Reference \cite{11} gave us a method of “gap statistic” for finding a suitable estimation of $k$ by exhausting all possible $k$ values, too. The two algorithms mentioned above both neglect that the different initial seeds set can “bring about” varied clustering results.

With a dataset DS1 illustrated in Fig.1, we execute k-means algorithm with $k$ ranging from 2 to 7, and with each $k$ the initial seeds set exhausted over the original data, e.g. when $k$ equals to 3, we execute k-means $C[21,3]=1330$ with all possible initial seeds sets. The experimental results are shown in Table1 ($N$: Numbers of S). We could not obtain the optimal $k$ for the reason that there exist $N$ results with a certain $k$ due to the various initial seeds sets, and $N$ increases when $k$ rises. We can get different curves of $s_k$ or $Gap_k$ for the same dataset, so that the optimal estimation of cluster numbers $k$ couldn’t be sought out, or a conclusion may be made that each integer $k$ ranging from 2 to 7 may be optimal, which is obviously not correct. At the same time, the view that the WC variance decreases monotonously in accordance with $k$ rises in \cite{8} and \cite{11} is not accurate without taking into consideration the influence of the initial seeds set.

| $k$ | $N$ | Range of $s_k$ |
|-----|-----|----------------|
| 2   | 1   | 16.1419        |
| 3   | 11  | [14.136 9, 17.126 8] |
| 4   | 34  | [12.965 2, 18.682 3] |
| 5   | 89  | [15.863 5, 20.571 3] |
| 6   | 236 | [13.013 8, 20.875 7] |
| 7   | 549 | [13.317 22,2968] |

![Fig.1 DS1, Scattered chart](image-url)
2.2 The clustering validity index

The optimal clustering must be in accordance with the purpose of minimizing $WC$ and maximizing $BC$. We propose a clustering validity index $\theta$ as Eq.(2) to judge which clustering is optimal with the restriction of $k$.

$$\theta_k = \frac{WC_k}{BC_k}$$

$\begin{align*}
WC_k &= \sum_{j=1}^{k} \sum_{i=1}^{c_j} \| p_j - m_i \| \\
BC_k &= \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \| m_i - m_j \| \\
m_i &= \frac{1}{[c_j]} \sum_{p \in c_j} p_j, i \subseteq [1,k]
\end{align*}$

Theorem The clustering evaluation indicator $\theta$ is a finite positive value, which is non-monotonous when integer $k$ is in the range $k \in [2,n)$. For a given dataset and cluster number $k$, the clustering result, from which we obtained the minimized $\theta_k$, is optimal.

Let us run $k$-means algorithm with $k=3$ for DS1 and DS2 (shown in Fig.2) with an exhaustive search of all possible initial seeds set. The clustering results of DS1 are 11 and 3 of DS2 as shown in Fig.2.

According to Fig.2, for DS1, the clustering result according with the minimum $\theta = 1.157$ is optimal as shown in Fig.3(a). Similar to DS1, the optimal result of DS2 shown in Fig.3(b) is obtained when $\theta = 0.844$.

As discussed above, it is impossible for a user to exhaust all potential initial seeds sets for clustering optimization whereas an optimal clustering must be obtained by $\theta_{\text{min}}$. In order to solve this contradiction, an initial seeds-seeking strategy based on the single-shortest-path graph is proposed below.

3 SSPG: an initial seeds seeking strategy

There are several familiar seeds-seeking methods, such as selecting $k$ points stochastically or evenly, selecting $k$ points which are farthest or nearest to the average value, etc. Obviously, the seeds with these algorithms will hit the seeds sets corresponding to $\theta_{\text{min}}$ with low probability sometimes. In our experiments, we find that more than seventy percent “good” initial seeds are distributed in the denser region, with which the algorithm can speed up the convergence. Consequently, we propose an initial seeds-seeking principle.

- **Density principle**: Initial seeds should be selected in the denser region as far as possible.

- **Scatteration principle**: The initial seeds should be dispersive as much as possible.

3.1 SSPG

With SSPG we can get the “good” initial seeds directly from the original data set, without any artificial influence in the process. For a given dataset $D \in \mathbb{R}^d$, the process is described below:

**Step 1** Compute the gap of each dimension of data set with Eq.(3). Let us suppose that $L$th gap is maximal. Then with the smallest point in $L$th dimension as source point, the biggest in $L$th dimension point as end point, iteratively, try to find the closest point and put this point in the end of set $D^*$ until the
end point is reached.

\[
D = \{D_1, D_2, \cdots, D_n\}
\]

\[
D_i = (D_{i1}, D_{i2}, \cdots, D_{in}), i \in [1, d]
\]

\[
\text{Gap}_L = \max(D^j) - \min(D^j), L \subset [1, d]
\]  

\[
(3)
\]

**Step 2** In \(D^\ast\), compute the Euclidean distance of adjacent points respectively as the weight labeled with \((1, 2, \cdots, 1)\) of the two points. After the average weight \(W_{\text{avg}}\) is computed as Eq.(4), compute the neighbors of each data with a circular scope with \(W_{\text{avg}}\) as radius and itself as center. Then label the number of neighbors as the density \(d_i\) of \(i\)th data.

\[
W_i = \|D^i_j - D^j_i\|, 1 \leq i = j - 1 < n
\]

\[
W_j = \sum_{i=1}^{n} W_i
\]

\[
W_{\text{avg}} = \frac{1}{n} W_t
\]  

\[
(4)
\]

**Step 3** Rearrange the data in \(D\) on the principle of decreasing density, then the first data (which have the max density) in \(D\) is inserted into a seeds candidate set \(S\), on the principle that the distance between any two seeds must be more than or equal to \(2W_{\text{avg}}\), select the data from \(D\) in turn and make sure whether it can be a seed; if yes, then put this data at the end of \(S\). This procedure can be stopped when \([S]\) is equal to \(3k\). Therefore, the seeds candidate set \(S = \{S_1, S_2, \cdots, S_3\}\) is generated. But in case the number of seeds in \(S\) is less than \(3k\), the distance between any pair of two seeds should be decreased by \(0.1W_{\text{avg}}\) from \(2W_{\text{avg}}\) in turn until the number of candidate seeds is equal to \(3k\).

**Step 4** According to the scattered principle of initial seeds seeking strategy, we should choose \(k\) seeds from \(S\) as initial seeds. Let us use \(S^\ast = \{S_1^\ast, S_2^\ast, \cdots, S_n^\ast\}\), which stands for initial seeds set.

(1) With the similar method to step1, find the dimension with the largest gap, then choose the biggest and the smallest data of this dimension in \(S\) as initial seeds, here \(S^\ast = \{S_1^\ast, S_2^\ast\}\). Let \(S'' = (S - S^\ast)\) be the rest seeds in \(S\) set.

(2) Find which data in \(S''\) can meet the demand of \(\max_{p, s \in S''} \left( \sum_{p, s \in S''} \|p - p_i\| \right)\), and then put this data into \(S''\) and remove it from \(S''\).

(3) Repeat (2), until \([S''\) = \(k\).

**Step 5** End of the algorithm.

We can use SSPG to obtain “good” initial seeds set for a given data set, because: firstly, the average weight \(W_{\text{avg}}\) can explain the distribution of data set well and truly, if the data are scattered densely or more loosely, \(W_{\text{avg}}\) will decrease or increase. In comparison with other algorithms [6,7,12], SSPG algorithm avoids the artificial influence. Secondly, in order to remedy the problem that all seeds may be chosen in one area, in step3 we choose seeds with some restrictions such as the distance between any pair of two seeds. Finally, however, one may argue that the procedure of SSPG has an associated cost, but we can draw a conclusion that, with SSPG, \(k\)-means algorithm will converge even faster. It is because, with SSPG, the initial seeds used in \(k\)-means algorithm are even closer to the centroid of the final cluster, so the times of iteration will decrease dramatically; therefore, the efficiency will be even higher.

### 3.2 Experiments

Using SSPG in the \(k\)-means algorithm with \(k\) equal to three of data set DS1, we compute the average weight and with it as radius the density of each data is counted as shown in Fig.4(a), then with the distance restriction of any pair of two seeds, the initial seeds set that contains \(\{p_1, p_3, p_{13}\}\) is brought forward as shadowed points in Fig. 4(a) and (b), which explains the density and scattering principle primarily, and thereby, the optimal clustering as Fig.3(a) shows is obtained.

![Fig.4 Procedure of SSPG used in DS1](image-url)
us, then with the scattering principle we obtain an initial seeds set that contains three points with a circle around, which follows the $\theta_{\text{min}}$ principle and therefore can lead the clustering process to an optimal result.

In the area of Wireless Sensor Network (WSN)\textsuperscript{[2]}, in order to transmit information between nodes and the base station, the nodes should be grouped into some clusters in advance, and in each cluster, a sink node should be assigned to take charge of the communication affairs of all nodes in self cluster. The assigned sink node will consume more energy than other nodes in the cluster. As shown in Fig.5(b), we propose intuitively that there are four clusters; so with $k$ equal to four, we use SSPG strategy to seek initial seeds, and the final result also proves that the SSPG strategy is efficient and has excellent performance.

Furthermore, with SSPG strategy used in CURE\textsuperscript{[13]} algorithm to find the “scattered well” points as representatives, or with it in $k$-Medoids algorithms to seek initial seeds, similarity to the experiments for DS1 to DS4, the result may be relatively good. For large database which contains thousands of objects, we could recur to some sample algorithms such as that proposed in Reference\textsuperscript{[15]} to reduce the complexity in step1 of SSPG strategy. And some experiments in Reference\textsuperscript{[13]} showed that the sample drawing have little influence on the accuracy of SSPG.

The discussion above shows the methods of obtaining optimal clustering with restriction such as a given cluster number $k$. However, in practice, it is so difficult for users to determine an appropriate value of $k$, so we should attempt to get an optimal reference value of $k$. The next section will give a method based on $\xi$ curve to present a suitable value for cluster number.

### 4 Optimal $k$ estimation

According to the definition of clustering, clustering optimization is a combinational problem which minimizes $WC$ and simultaneously maximizes $BC$. And with further experiments, we find that an optimal reference $k$ could be made out by the data set itself.

With SSPG strategy, we could get the optimal initial seeds set and hereby an optimal clustering result can be obtained with a certain cluster number $k$. So, with integer $k$ tuned in the range of $[2,k-1]$, and with each value of this range, when the optimal clustering is obtained, there exist the determinate values of $WC$ and $BC$ according to the given $k$, i.e., we can work out the optimal reference $k$ by analyzing $WC$ and $BC$ with combinational theory. When $k$ increases, $WC_k$ decreases and $BC_k$ rises simultaneously. Let us take DS1 as an example, we can run $k$-means algorithm with each $k$ of range $[2,7)$, so $WC_k$ and $BC_k$ are worked out. With combinational theory, $k$ value which can minimize $WC$ and maximize $BC$ simultaneously is optimal. So we propose a curve of $\xi_k$ as Fig.8 shows which can be attained by Eq.(5).

$$\xi_k = |1 - \theta_{\text{min}}^k| \times BC_k$$

According to Eq.(5), we can find that with the variants of $WC$ and $BC$, $\exists k \in [2,n], \forall \min(\xi_k)$. And the value of $k$ which can minimize $\xi$ is what we seek. To the $\xi$ curve, it is easy to prove that with a certain $k$, the curve to the left or right of $k$ is monotonous.

We do experiments on DS1 with $k$ ranging from integer 2 to 7, using SSPG strategy and thereby the optimal results that correspond to $\theta_{\text{min}}$, and the $\xi_k$ is computed subsequently as shown in Fig.6(a). According to Fig.6, we find that when $k$ is equal to 3, $\xi_k$ is minimized. Therefore, we can draw a conclusion that this value ($k=3$) is the optimal reference for
the number of clusters. Similarly, from Fig.6(b) which demonstrated the experimental result for dataset DS 4, the value of \( k=4 \) correspond to the minimum \( \xi = 2.889 \), so \( k=4 \) is the optimal reference for DS4.

The experimental results prove that the reference number of optimal cluster is correct and efficient. Actually, in practice, we do not need to compute every value of \( \xi \) because the curve will change the monotony when the optimal reference is reached; so, we could stop the procedure of seeking optimal reference when this happens. Thereby this algorithm will run with even higher efficiency.

5 Conclusion

In this paper, we address the problems of spatial clustering optimization, and our contributions are as follows:

- Draw a conclusion that the research on spatial clustering optimization should be under some restrictions such as the number of clusters. And an evaluation indicator \( \theta \) is proposed to appraise the clustering result.
- Bring forward the \( \theta_{\min} \) principle to seek the “good” initial seeds.
- Propose an initial seeds seeking strategy—SSPG, and with the initial seeds SSPG generated, we can obtain the optimal clustering with high probability.
- Propose a method called \( \xi \) curve to obtain the optimal reference of the number of final optimal clusters.

So, for a given dataset, with a given \( k \), we can obtain the optimal clustering by virtue of \( \theta_{\min} \) principle. In case the user cannot offer \( k \), it may be returned to the \( \xi \) curve method to get an optimal reference. With this value as an input variant in the procedure of clustering algorithm, and with SSPG used to seeking “good” initial seeds, then the optimal clustering result can be obtained. However, this result may not meet the user’s demand; but actually, from the point of view of mathematics, it is the optimal result for sure.

References

[1] Wang Quanhong, Hassanein H, Takahara G (2004) Stochastic modeling of distributed, dynamic, randomized clustering protocols in wireless sensor networks [C]. ICPP 2004 Workshop on Ad Hoc and Sensor Networks, Montreal
[2] Bandyopadhyay S, Coyle E J (2003) An energy efficient hierarchical clustering algorithm for wireless sensor networks [C]. Proc. IEEE INFOCOM’03, San Francisco
[3] Han Jiawei, Kamber M (2000) Data mining: concepts and techniques [M]. San Francisco: Morgan Kaufmann
[4] MacQueen J B (1967) Some Methods for classification and Analysis of Multivariate observations [C]. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley
[5] Kaufman L, Rousseuw P J (1990) Finding Groups in Data: an introduction to cluster analysis [M]. New York: John Wiley & Sons
[6] Zhang Tian, Ramakrishnan R, Livny M (1996) BIRCH: an efficient data clustering method for very large databases [C]. Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, Canada
[7] Beckmann N, Kriegel H P, Schneider R (1990) The R-tree: an efficient and robust access method for points and rectangles [C]. Proc. of ACM SIGMOD, Atlantic City, NJ
[8] Li Yongsen, Yang Shanlin, Ma Xijun, et al. (2006) Optimization study on \( K \) value of spatial Clustering [J]. Journal of System Simulation, 18(3): 573-576
[9] Knorr E, Ng R (1998) Algorithms for mining distance-based outliers in large datasets [C]. Proc of the VLDB Conference, New York
[10] Barnett V, Lewis T (1994) Outliers in Statistical Data [M] New York: John Wiley & Sons
[11] Tibshirani R (2001) Estimating the number of clusters in a data set via the gap statistic [R]. Tech.Rep., Stanford University
[12] Li Suohua, Sun Zhihui, Zhou Xiaoyun (2006) Distributed clustering algorithm based on feature vector [J]. Computer Applications, 26(2): 379-382.
[13] Suditch G, Rajeev R, Kyuseok S (1998) CURE: an efficient clustering algorithm for Large Database [C]. Proceedings of the ACM SIGMOD Conference on Management of Data, Seattle, WA, USA.
[14] Guo Jiasheng, Zhao Yi, Shi Pengfei (2001) An efficient dynamic concept clustering algorithm for data mining [J]. Journal of Software, 12(4): 582-591
[15] Vitter J (1985) Random sampling with a reservoir [J]. ACM Transactions on Mathematical software, 11(1): 37-57