Journal of System Simulation

Volume 30 | Issue 6 Article 13

6-14-2018

Clustering Method Based on Graph Data Model and Reliability Detection

Yanyun Cheng
1. Nanjing University Of Posts And Telecommunications, Nanjing 210023, China;;

Huisong Bian
1. Nanjing University Of Posts And Telecommunications, Nanjing 210023, China;;

Changsheng Bian
2. Nanjing College of Information Technology, Nanjing 210023, China;

Clustering Method Based on Graph Data Model and Reliability Detection

Abstract

Abstract: For the data in feature space, traditional clustering algorithm can take clustering analysis directly. High-dimensional spatial data cannot achieve intuitive and effective graphical visualization of clustering results in 2D plane. Graph data can clearly reflect the similarity relationship between objects. According to the distance of the data objects, *the feature space data are modeled as graph data by iteration.* Cluster analysis based on modularity is carried out on the modeling graph data. *The two-dimensional visualization of non-spherical-shape distribution data cluster and result is achieved. The concept of credibility of the clustering result is proposed, and a method is proposed, which the Page Rank algorithm is used to calculate the reliability of clustering results.*

Keywords

data mining, clustering, graph data modeling, modularity, Page Rank algorithm

Recommended Citation

Cheng Yanyun, Bian Huisong, Bian Changsheng. Clustering Method Based on Graph Data Model and Reliability Detection[J]. Journal of System Simulation, 2018, 30(6): 2102-2109.

第 30 卷第 6 期 系统仿真学报© Vol. 30 No. 6
2018 年 6 月 Journal of System Simulation Jun., 2018

基于图数据模型的聚类方法及可信度检测

程艳云[1]，边荟淞[1]，边长生[2]
(1. 南京邮电大学自动化学院，江苏 南京 210023；2. 南京信息职业技术学院，江苏 南京 210023)

摘要：对于特征空间中的数据，传统聚类算法通常直接在特征空间中进行聚类分析，因此高维空间数据无法在二维平面实现直观有效的聚类结果图形可视化，图数据可以明确反映对象之间的相似性关系，根据数据对象之间的距离，*通过迭代将特征空间的数据建模成图数据。*并对建模得到的图数据模型进行基于模块性的聚类分析，*实现对非凸球分布数据集的聚类及对聚类结果实现二维空间的图形可视化。提出了聚类结果关于类间邻近边界的可信度概念，并提出了一种利用 PageRank 算法实现对聚类结果可信度计算的方法。*

关键词：数据挖掘；聚类；图数据建模；模块性；PageRank 算法
中图分类号：TP391.9 文献标识码：A 文章编号：1004-731X (2018) 06-2102-08
DOI: 10.16182/j.issn1004731x.joss.201806013

Clustering Method Based on Graph Data Model and Reliability Detection

Cheng Yanyun[1], *Bian Huisong*[1], *Bian Changsheng*[2]
(1. Nanjing University Of Posts And Telecommunications, Nanjing 210023, China;
2. Nanjing College of Information Technology, Nanjing 210023, China)

Abstract: For the data in feature space, traditional clustering algorithm can take clustering analysis directly. High-dimensional spatial data cannot achieve intuitive and effective graphical visualization of clustering results in 2D plane. Graph data can clearly reflect the similarity relationship between objects. According to the distance of the data objects, *the feature space data are modeled as graph data by iteration.* Cluster analysis based on modularity is carried out on the modeling graph data. *The two-dimensional visualization of non-spherical-shape distribution data cluster and result is achieved. The concept of credibility of the clustering result is proposed, and a method is proposed, which the Page Rank algorithm is used to calculate the reliability of clustering results.*

Keywords: data mining; clustering; graph data modeling; modularity; Page Rank algorithm

引言

聚类分析是一个将数据对象划分成子集的过程。每个子集是一个簇，使得簇中的对象彼此相似，但与其他簇中的对象不相似。对于特征空间中的数据，经过几十年的发展，很多学者提出很多聚类分析方法，从算法的角度看，聚类算法主要可以分为：划分聚类算法[1-2]，层次聚类算法[3-4]，基于密度的方法[5]，基于网格的方法[6]和基于模型的方法[7]。这些方法通常在多维特征空间中进行分析计算，对于高维数据无法实现直观的图形可视化。同时，大多数算法只将各数据点定性归为一类或几类，无法计算各数据点属于该类的程度(硬聚类)。而在实际

收稿日期：2016-07-20 修回日期：2016-09-17;
基金项目：江苏省省级现代服务业(软件产业)发展专项引导资金(SJ214038);
作者简介：程艳云(1974-)，女，江苏姜堰，硕士，副教授，研究方向为大数据在移动通信网络中的应用及网络优化；边荟淞(1989-)，男，山东德州，硕士生，研究方向为聚类分析与张量异常块检测。

应用中，给出聚类结果中不同类之间邻近边界处的数据点属于该类的可信度是很有意义的。例如，将产品的用户进行聚类并进行精确营销时，若能得到各用户在当前聚类结果中关于类间邻近边界的可信度，就可以将有限的人力物力成本用于可信度较高的对象，以改善营销结果。模糊聚类算法[8-9]虽然可以根据隶属度给出软聚类结果，但不在与其他类的邻近边界处的数据点也可能具有较小的隶属度。同时基于距离的聚类方法只能处理凸球形分布的簇，对于非凸球分布的数据集，将产生不合适的聚类个数或错误的聚类结果[10-11]。

针对上述问题，本文将提出一种将特征数据映射到图数据的建模方法和利用 PageRank 算法对聚类结果中各节点进行可信度检测的方法，并对聚类结果实现二维平面的图形可视化。根据数据对象之间的相似度，将特征空间数据映射成图数据，使图中的节点对应数据对象，节点之间的连接关系反映成数据之间的相似性关系。再将数据点之间的相似关系映射为图数据中的加权边，可以得到各数据点对于当前聚类结果的可信度。由于将数据点映射为图数据中的节点，所以对于高维数据也可实现在二维平面的图形可视化，以直观反映数据点之间及不同类之间的关系。

1 基于距离的图数据建模

本文提出一种基于距离的建模方法，建模过程采用迭代方法。通过迭代，将数据从特征空间映射到图空间。对于有 n 个数据对象的数据集，迭代过程的初始化图数据为含有 n 个孤立节点，连接边数为 0 的非连通图。这里，我们使用图论中的术语“连通分支”(简称“分支”)，则对于初始图数据，分支个数等于节点个数 n。在建模过程中，分支个数随迭代次数的增加而减少，直到整个图数据的分支数为 1，即整个图连通，迭代终止。

定义 1 若将第 i 次迭代记为 $iter^{(i)}$，第 i 次迭代后的分支数记为 $\mathrm{comp}^{(i)}$，称 $\mathrm{comp}^{(i)} = F(iter^{(i)})(i = 0,1,...,N)$ 为迭代分支函数(iteration-component function，以下简称 IC 函数)。

不难得出，分支数随着迭代次数的增加而减少，因此 IC 函数是单调非增函数。

定理 1 设初始图有 n 个孤立节点，第一次迭代连接所有节点的最近邻节点(互为最近邻节点只连接一次)，则第一次迭代之后分支数 $comp$ 满足：

$$comp \leqslant \frac{n}{2}$$

当且仅当 n 为偶数，节点两两互为最近邻时等号成立。

定理 2 从初始图开始，如果每次迭代都连接各节点当前未连接且空间距离最近节点，则数据点在特征空间中的分布越均匀，即自然簇越不明显，所需的迭代步数越少。反之，自然簇特性越明显，所需的迭代步数越多。

证明：当数据点在特征空间中的分布相对均匀时，设两自然簇的最短距离为 $d_{\min}$，而对应这一最小距离的两个数据点分别为 c_1 和 c_2，对于 c_1，在其自然簇内，与其距离由近到远的数据点分别为 $c_{i1}, c_{i2}, ..., c_{iM}$ (i 表示自然簇标号，M 表示除 c_1 外该簇的数据点个数)。对应的距离分别为 $d_{i1} < d_{i2} < ... < d_{iM}$，假设 $d_{ik} < d_{\min}$，由于数据分布相对均匀，k 一定取较小的值，因此，需要较少的迭代步数就可以使各自然簇连通；反之，若自然簇特性明显，即簇间最小距离远大于簇内数据点距离，则 k 必然取较大的值，此时需要迭代更多步才能实现全部连通。证毕。

通过定理 2 可以得出，数据点在特征空间中的分布情况直接决定建模所需要的迭代步数。如图 1 所示的数据集(Hartigan，1975)和图 2 所示的数据集(Ruspini)都是用于实验分析的经典数据集。根据定理 1，我们选择 $c = n / 2^i$ 作为参考函数，图 1~4 分别显示了相对均匀和自然簇特性明显的空间分布情况和对应的 IC 函数。

DOI: 10.16182/j.issn1004731x.joss.201806013

第 30 卷第 6 期 系统仿真学报 Vol. 30 No. 6
2018 年 6 月 Journal of System Simulation Jun., 2018

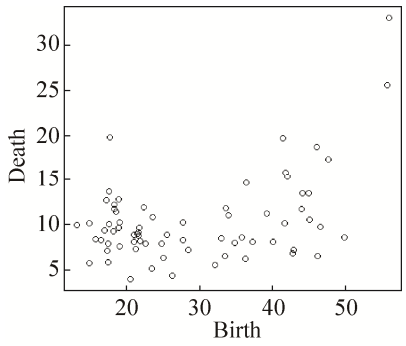

图 1 Hartigan 数据集的空间分布
Fig. 1 Spatial distribution of Hartigan dataset

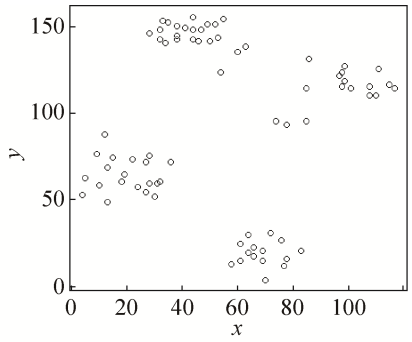

图 2 Ruspini 数据集的空间分布
Fig. 2 Spatial distribution of Ruspini dataset

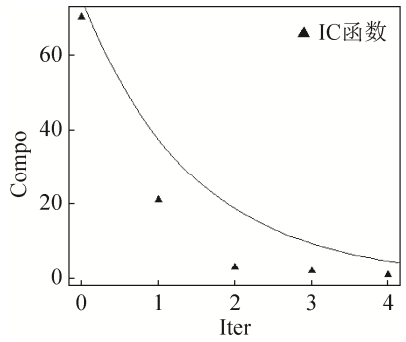

图 3 Hartigan 数据集迭代过程的 IC 函数
Fig. 3 IC function of iterative process of Hartigan dataset

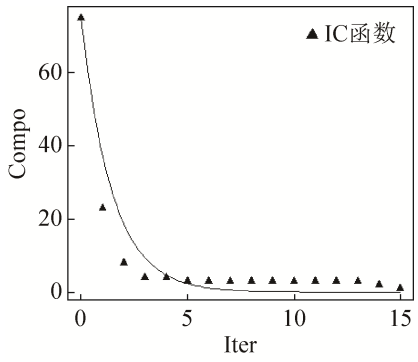

图 4 Ruspini 数据集迭代过程的 IC 函数
Fig. 4 IC function of iterative process of Ruspini dataset

2 聚类结果可信度检测

通过上述方法，特征空间中的数据根据相似度建模成加权图数据，多维数据点由图数据的节点表示，数据之间的相似性关系由加权边表示。通过对得到的图数据进行聚类分析，不仅可以得到原特征空间数据的聚类结果，而且可以通过任一节点与类内和类外节点的连接关系，给出该节点对于当前聚类结果的可信度，并能实现在二维平面的图形可视化。本文采用基于模块性的聚类算法[12-13]，实现对前面建模得到的加权图的聚类分析。进而实现对原特征空间中的数据对象的聚类分析。

对于聚类结果的可信度检测，本文提出了一种利用 PageRank 算法[14]对聚类结果中的各节点进行可信度检测的方法。本文通过对邻接矩阵的重新定义，将 PageRank 算法应用于可信度检测，可信度越高(PR 值越小)，表示该节点距离类间邻近边界越远，被正确聚类的可能性越大；反之，若可信度低(PR 值越大)，表示节点距离邻近边界越近，可能被错误聚类的可能性越大。对于已有的聚类结果，设 d^i 为节点 i 的度，d^i_{out} 表示节点 i 与其他类节点的连接边数，d^i_{in} 表示节点 i 与类内节点的连接边数。显然有 $d^i = d^i_{out} + d^i_{in}$。对于邻接矩阵 $\boldsymbol{A}$，a_{ij} 为 A 的第 i 行第 j 列元素，则有：

$$a_{ij} = \begin{cases} \dfrac{1}{d^j_{out}} & \text{，节点 } i \text{ 和节点 } j \text{ 连接且属于不同类} \\ 0 & \text{，其他} \end{cases}$$

利用上述邻接矩阵，通过 PageRank 算法迭代计算，可以得出各节点的 PR 值，进而得到各节点在当前聚类结果中，相对于邻近边界的可信度。

3 实验仿真

3.1 聚类准确性及可视化实验分析

这里我们使用 iris 数据集，它记录了 3 类共 150 朵鸢尾花的花萼长度，花萼宽度，花瓣长度和花瓣宽度，数据集中同时给出了这些鸢尾花的真实分类，以方便对聚类结果进行检验，数据集如表 1

所示。其中，1~50 为 setosa(set)类，51~100 为 versicolor(ver)类，101~150 为 virginica(vir)类。iris 数据集的数据点个数为 150，因此，初始化为包含 150 个孤立节点的图模型。我们设置连接阈值 $Q=30$，以避免过度冗余的迭代，并取距离-权重函数的参数 D 等于相似性矩阵中的最大距离 $dist_m=23.61$(特征空间中数据对象的最远距离)，进行迭代建模，然后对建模得到的图数据，利用基于模块性的聚类算法进行聚类分析，最终得到的聚类结果。

表 1 iris 数据集
Tab. 1 iris dataset

编号	花萼长度	花萼宽度	花瓣长度	花瓣宽度	真实分类
1	5.1	3.5	1.4	0.2	set
2	4.9	3.0	1.4	0.2	set
		……			
51	7.0	3.2	4.7	1.4	ver
52	6.4	3.2	4.5	1.5	ver
		……			
101	6.3	3.3	6.0	2.5	vir
102	5.8	2.7	5.1	1.9	vir
		……			

通过与真实分类进行比较：setosa 类 50 个数据对象全部被正确聚类，聚类正确率为 100%，versicolor 类有 5 个数据对象被错误聚类到 virginica 类，聚类正确率为 90%，virginica 类有 1 个数据对象被错误聚类到 versicolor 类，聚类正确率为 98%，因此，平均正确率为 96%。对数据集用传统 K-means 算法进行聚类对比，由于传统 K-means 算法具有随机性，本文进行 20 次聚类，并取平均值，最终聚类结果平均正确率为 89.3%。对数据集利用最大最小距离算法进行聚类分析，正确率为 90.7%。最后利用改进 K-means 算法[15]对数据集进行聚类分析，正确率为 96%。聚类结果正确率对比见表 2。

可以看出，本问提出的方法在聚类结果准确性上明显优于传统 K-means 算法和最大最小距离算法，与改进的 K-means 算法准确率持平。

表 2 聚类结果正确率对比
Tab. 2 Comparison of correct rate of clustering results

算法	准确率/%
传统 K-means 算法	89.3
最大最小距离算法	90.7
改进 K-means 算法	96
基于图模型算法	96

相比于在特征空间直接进行聚类分析的传统算法，本文的方法还实现了聚类结果直观有效的图形可视化，通过可视化结果，不仅可以得出最终的聚类结果，而且可以识别各类之间的关系，以及各类之间的重叠情况，如图 5 所示，每一个节点代表特征空间中的一个多维数据，节点之间的远近反映了数据点之间的相似度大小。可以看出，对于 iris 数据集，第 2 类和第 3 类的空间分布较近，且有一定的交叉，而第 1 类类则与第 2 类和第 3 类相距较远，且没有交叉。

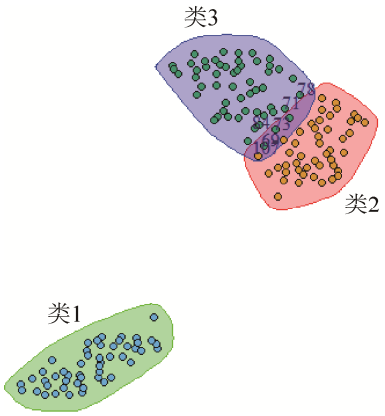

图 5 iris 数据集聚类结果的可视化
Fig. 5 Visualization of clustering result of iris dataset

3.2 聚类结果可信度检测实验分析

首先我们对上面 iris 数据集的聚类结果进行可信度检测。对于 iris 数据集，先列出被错误聚类的节点标号，如表 3 所示。

表 3 聚类错误的节点
Tab. 3 Incorrect points of clustering

类标号	错误节点标号
1	无
2	107
3	69，71，73，78，84

https://dc-china-simulation.researchcommons.org/journal/vol30/iss6/13
DOI: 10.16182/j.issn1004731x.joss.201806013

然后我们采用本文提出的算法进行可信度计算，计算结果如图 6 和图 7 所示。图 6 展示了各 PR 值在聚类结果中的分布情况，图 7 给出了图 6 中 PR 值与颜色的对照图及对应各 PR 值的节点个数(此处纵坐标采用对数坐标显示)。在图 6 中，圆形节点代表被正确聚类的数据对象，方形节点代表被错误聚类的数据对象。

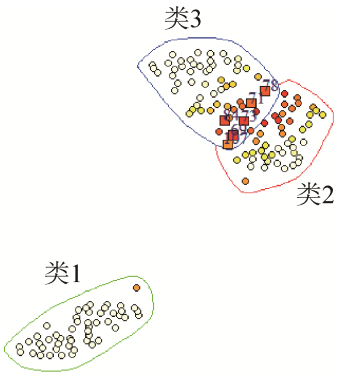

图 6 iris 数据集聚类结果及 PR 值分布
Fig. 6 Clustering result of iris dataset and PR distribution

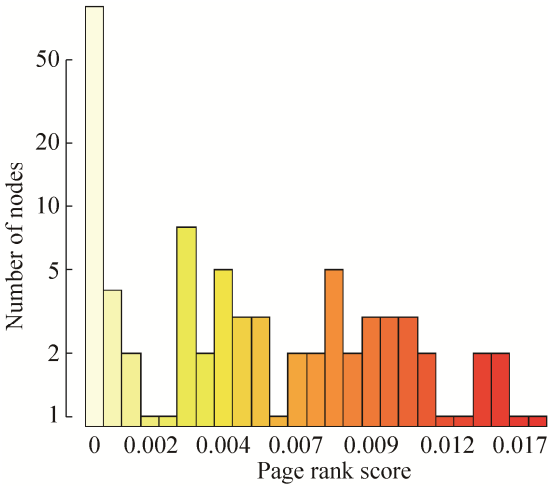

图 7 iris 数据集 PR 值对照图
Fig. 7 iris dataset and PR score comparison diagram

表 4 列出了被错误聚类节点对应的 PR 值。可以看出，越接近两个类的交叉边界，对应节点的 PR 值越高，即节点的可信度越低，而最终被错误聚类的节点都是具有较小可信度(较大 PR 值)的节点。

为了对比本文提出的可信度计算方法与模糊聚类给出隶属度的方法之间的区别，证明本文提出的方法更有效且更具有实际意义，我们采用 Nclus 开源数据集进行对比仿真。如图 8 所示，是利用模糊 C-均值进行聚类分析的聚类结果及隶属度分布。图 9 为图 8 中隶属度与颜色的对照图以及对应隶属度区间的数据点个数。

表 4 错误节点 PR 值
Tab. 4 PR score of incorrect points

错误节点标号	PR 值
69	0.013
71	0.011
73	0.013
78	0.019
84	0.012
107	0.010

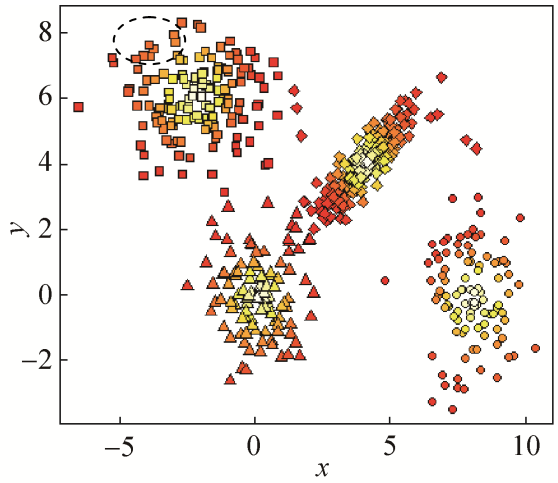

图 8 模糊 k 均值聚类结果的隶属度分布
Fig. 8 Membership distribution of fuzzy K mean clustering results

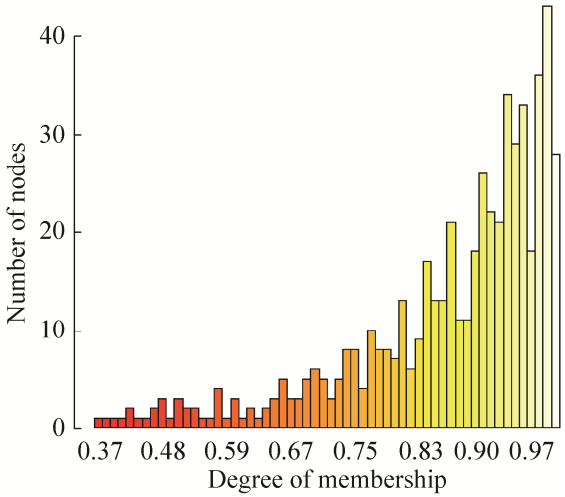

图 9 隶属度对照图
Fig. 9 Comparison diagram of membership

从图 8 中可以看出，直接利用隶属度作为可信度存在两个问题：

第一，具有较低隶属度的数据对象占总数据对象数量的比例过大，不具有实际参考价值；第二，很多不在类间邻近边界处的数据对象也具有较小的隶属度，如图 8 中的虚线圆圈内的数据对象。但从图中显然可以看出，圆圈内的数据对象不可能属于另外 3 类，即应具有较高的可信度。如图 10 所示为利用本研究提出的方法进行聚类分析的聚类结果及可信度分布。图 11 给出了图 10 中 PR 值与颜色的对照图以及各 PR 值区间的数据对象的个数。图 10 中只有类间邻近边界处的数据对象具有较小的可信度，且可信度较小的数据对象的数量较少，具有实际参考价值。

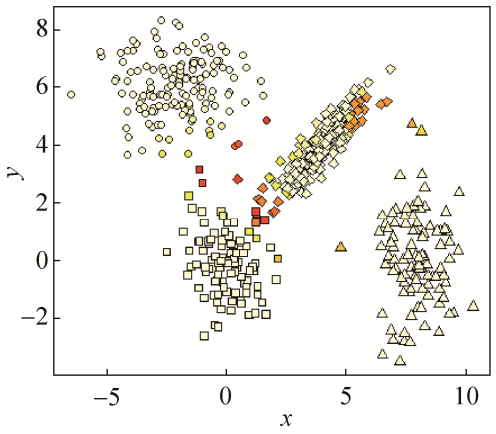

图 10 Nclus 数据集聚类结果 PR 值分布

Fig. 10 Nclus dataset clustering result and PR score distribution

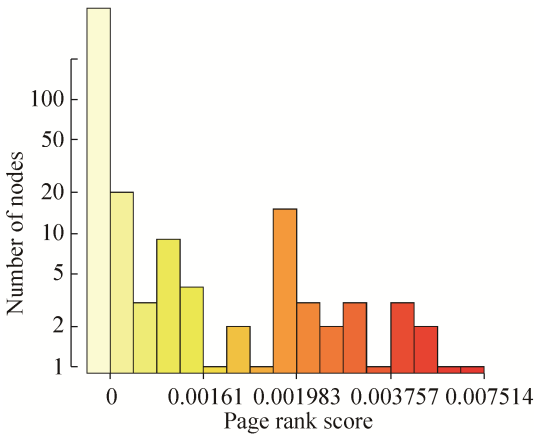

图 11 Nclus 数据集 PR 值分布图

Fig. 11 Nclus dataset PR score distribution diagram

3.3 非凸球分布数据的聚类结果分析

为直观表示本文提出的算法在非球形分布数据集上的聚类结果比传统基于距离的算法的优越性，本文采用 Cassini 数据集，其在二维空间的分布情况如图 12 所示，Cassini 数据集分为 3 个簇(上，中，下)，其中两个簇(上，下)为非凸球分布。利用本研究提出的方法对该数据集进行建模，距离-权重函数的参数取 D=7.83。图 12 为本文提出的聚类算法的聚类结果，可以看出，本文算法可以处理 Cassini 数据集，对于非凸球数据分布也能得到很好的聚类结果。图 13 为 K-means 算法的聚类结果，并用 Calinski Harabasz 指数确定最优聚类个数，可以得出，对于 Cassini 数据集所确定的最优聚类结果为 5。可以看出，利用 K-means 算法结合 Calinski Harabasz 指数将非凸球数据分布的簇(上，下)分为两个凸球分布，因此无法将数据集正确聚类。若用传统 K-means 算法，并指定聚类个数为 3，见图 14。

由于只能识别处理凸球数据分布，K-means 算法将数据集强行分割为 3 个凸球分布的簇，聚类结果与真实簇分布明显不符。因此，本文提出的方法基于距离为相似度度量，不仅可以准确聚类，且可以处理在空间中非凸球形分布的数据集。

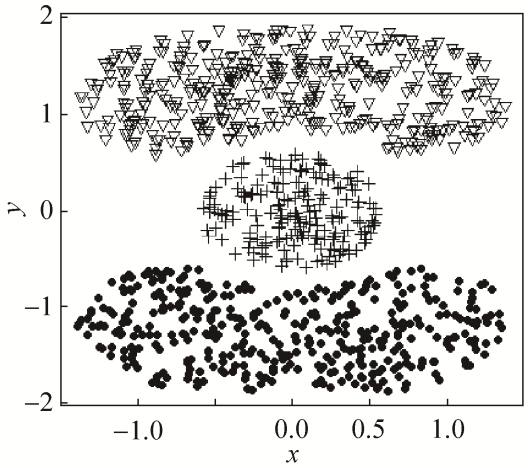

图 12 基于图数据的聚类结果

Fig. 12 Clustering result of graph-based data method

https://dc-china-simulation.researchcommons.org/journal/vol30/iss6/13
DOI: 10.16182/j.issn1004731x.joss.201806013

第 30 卷第 6 期 系统仿真学报 Vol. 30 No. 6
2018 年 6 月 Journal of System Simulation Jun., 2018

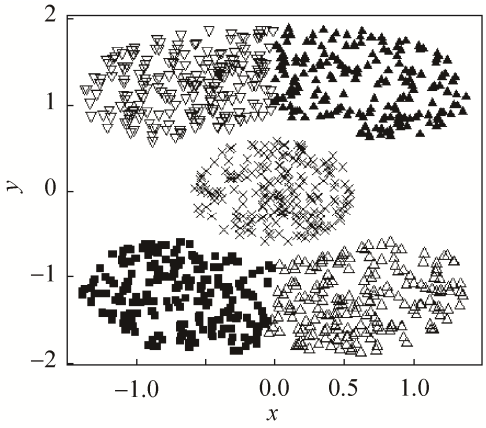

图 13 Calinski Harabasz 索引确定最优聚类个数
Fig. 13 Optimal cluster number based on Calinski Harabasz index

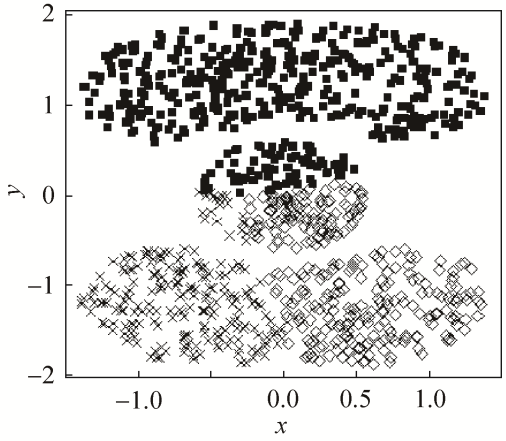

图 14 设置聚类个数为 3 的 K-means 聚类结果
Fig. 14 Clustering result of K-means with a cluster number of 3

4 结论

本研究提出了一种基于图数据模型的聚类方法，通过将特征空间数据建模成图数据，对特征空间数据进行聚类分析。通过实验仿真，可以证明，本研究提出的方法在聚类准确性上明显优于传统 K-means 算法和最大最小距离算法，与改进的 K-means 算法的正确率持平。同时，由于将特征空间数据映射到图数据空间，对于高维数据，实现了聚类结果在二维平面的图形可视化，通过可视化，不仅可以得到聚类结果，而且可以直观的展示各簇之间的远近关系及各类之间的交叉关系。最后，本研究利用图数据的连接关系，提出了各节点对于聚类结果的“可信度”度量，并提出了一种利用 PageRank 算法进行可信度检测的方法。可以看出，被错误聚类的数据对象都具有较低的可信度，且在邻近边界处数据对象的可信度明显小于远离邻近边界的数据对象。可信度的提出和运用对于成产生活有着重要的意义。

参考文献：

[1] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications (S0957-4174), 2009, 36(2): 3336-3341.

[2] Ng R T, Han J. CLARANS: A Method for Clustering Objects for Spatial Data Mining[J]. IEEE Transactions on Knowledge & Data Engineering (S1041-4347), 2002, 14(5): 1003-1016.

[3] Guha S, Rastogi R, Shim K. ROCK: A Robust Clustering Algorithm for Categorical Attributes[C]// International Conference on Data Engineering, 1999. Proceedings. 2002: 345-366.

[4] Karypis G, Han E H, Kumar V. Chameleon: hierarchical clustering using dynamic modeling[J]. Computer (S0018-9162), 1999, 32(8): 68-75.

[5] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: ordering points to identify the clustering structure[C]// Proceedings ACM SIGMOD International Conference on Management of Data, 1999, 28(2): 49-60.

[6] Sheikholeslami G, Chatterjee S, Zhang A. WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases[C]// Proceedings of the 24rd International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1998: 428-439.

[7] Fisher D H. Knowledge acquisition via incremental conceptual clustering[J]. Machine Learning (S0885-6125), 1987, 2(2): 139-172.

[8] Leski J. Towards a robust fuzzy clustering[J]. Fuzzy Sets & Systems (S0165-0114), 2003, 137(2): 215-233.

[9] Carvalho F D A T D. Fuzzy c -means clustering methods for symbolic interval data[J]. Pattern Recognition Letters (S0167-8655), 2007, 28(4): 423-437.

[10] 赵京胜, 孙梦丹, 张丽. 一种有效的 K-means 初始中心优化算法[J]. 信息技术与信息化, 2016(5): 77-79.
Zhao J S, Sun M D, Zhang L. An effective k-means algorithm with initial center optimization[J], Information Technology and Information. 2016(5): 77-79.

(下转第 2116 页)

