文章编号:1004-3918(2019)04-0507-07

基于孤立点自适应的K-means算法

杨莉云， 颜远海

（广东财经大学 华商学院，广州 511300）

摘 要：孤立点的存在使聚类中心的计算产生较大误差,影响K-means算法的聚类效果．针对该问题,引入谢林模型,使孤立点能够自动移动到其邻居所在位置,消除孤立点,同时,对K-means算法过程中的距离计算、初始聚类中心选取环节进行改进,提出基于孤立点自适应的K-means算法．该算法首先对原始数据进行归一化处理,以提高距离计算的准确性;然后,根据谢林模型的基本思想,将孤立点移动到其最近的多邻邻居;接着,由类簇的数目确定邻居样本的搜索范围,确定初始聚类中心;最后,根据移动后的数据集和初始聚类中心,进行K-means聚类．在UCI机器学习数据库中经典聚类数据集上的实验结果表明,该算法可显著提升聚类的精度,同时,簇的内聚性也比较好．

关键词：K-means算法；孤立点；谢林模型；初始聚类中心；误差平方和

中图分类号：TP 181 **文献标识码**：A

K-means Algorithm Based on Outliers Adaptive

YANG Liyun， YAN Yuanhai

（Huashang College，Guangdong University of Finance & Economics，Guangzhou 511300，China）

Abstract：The existence of outliers brings much deviation in the calculation of clustering centers and affects the clustering results of K-means algorithm. To solve this problem，this paper introduces Schelling model，makes the outlier automatically move to the location of its neighbors，and eliminates the outliers. At the same time，the distance calculation and the initial cluster center selection in K-means algorithm are improved，and the K-means algorithm based on outliers adaptive is proposed. In the algorithm，the original data are normalized to improve the accuracy of distance computation. Then，according to the basic idea of Schelling model，the outliers are moved to its nearest neighbor. The search range of neighbor samples is determined by the number of clusters，and the initial clustering centers are determined. Finally，the K-means clustering is carried out according to the moved data set and the initial clustering center. Experimental results on classical clustering data sets in UCI machine of learning database show that the algorithm can significantly improve the accuracy of clustering，and clusters also have a better cohesiveness.

Key words：K-means algorithm； outliers； Schelling model； initial clustering center； sum of the squared errors

聚类分析是数据挖掘领域的常用数据分析方法．聚类分析就是将数据对象分组成为多个类或簇，在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大[1]．目前,聚类分析的方法有很多,其中基于划分的K-means算法以其简单、快速并有效处理大规模数据等诸多优点,成为应用最广泛的聚类方法之一[2]．但是,K-means算法也存在需要用户预先设定k值、聚类质量依赖于初始解的选择、聚类质量容易受孤立点的影响等问题．

当前,有许多学者对K-means算法进行了改进．文献[1]以数据对象每一维的均值为中心,在均值两侧标准差范围内均匀分段,每一段的端点即为聚类中心在该维的坐标．文献[2]将数据样本点向每一维的坐标

收稿日期：2019-01-19

基金项目：广东省普通高校青年创新人才项目(2015WQNCX203,2017KQNCX266)

作者简介：杨莉云(1984-),女,讲师,主要研究方向为商务智能,机器学习

轴投影，根据数据样本的分布特性递归地将样本空间划分为若干个超立方体，超立方体稠密区域的几何重心点即为初始聚类中心．文献[3-5]的基本思想一致：在所有数据点中找出邻居数最多的点作为第一个聚类中心，将该点和其所有邻居删除，在剩余点中找邻居最多的点，直到找到k个点为止．文献[6]在文献[3]的基础上将孤立点分离出来，使孤立点不参与初始聚类中心的计算过程．与文献[3]过程类似，文献[7]选择方差最小的点作为聚类中心．文献[8]依据类间相似度逐步选择聚类中心．文献[9]将布谷鸟搜索算法引入K-means聚类，以搜寻最优的初始聚类中心．

以上研究人员的工作都对传统的K-means算法进行了不同程度的改进，其中文献[3]所提出的以数据点的密度选择初始聚类中心的算法符合聚类分析的基本思想且适用于多维数据聚类，比传统的K-means算法的精确度有所改进但仍不够理想．本文在文献[3]算法的基础上加入对孤立点的处理，同时对数据对象之间的距离度量方法、初始聚类中心选取进行改进，以获得更好的聚类效果．

1 K-means算法中的距离度量

数据点间亲密度或距离如何定义直接影响着聚类结果．距离度量的方法有多种，如欧氏距离、曼哈顿距离、切比雪夫距离、马氏距离、汉明距离等．对于很多数据集，用欧氏距离作为定义数据点间亲密度的基础，即可获得较好的聚类结果，欧氏距离是聚类分析中最为常见的数据点间距离定义方法[10]．两个数据点之间的欧氏距离如下：

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2},\tag{1}$$

其中：$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})$，$\boldsymbol{x}_j=(x_{j1},x_{j2},\cdots,x_{jp})$是两个$p$维向量．

由于数据对象各个属性的取值可能会存在比较大的差异，例如Wine数据集，共有13个属性，大部分属性的取值范围变化很小，标准差小于1，但第13个属性值范围变化很大，标准差为314.9．直接以欧氏距离计算，会导致数据对象之间的距离严重依赖于第13个属性而忽略其他属性，由此得到的聚类结果也将存在较大误差．同时，人们将数据对象到所属类中心的距离平方和作为聚类效果好坏的主要评判标准之一，也会得出错误的评价．

标准化欧氏距离是针对欧氏距离的缺点而作的一种改进方案．为了消除数据各维分量量纲的差异，标准欧氏距离将各个分量都“标准化”到均值、方差相等．假设原分量$\boldsymbol{x}$的均值为μ，标准差为σ，则标准化后的分量为$\boldsymbol{x}^*=\dfrac{\boldsymbol{x}-\mu}{\sigma}$，$\boldsymbol{x}^*$的均值为0，标准差为1．两个$p$维向量$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})$与$\boldsymbol{x}_j=(x_{j1},x_{j2},\cdots,x_{jp})$的标准化欧氏距离为：

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sqrt{\left(\frac{x_{i1}-x_{j1}}{\sigma_1}\right)^2+\left(\frac{x_{i2}-x_{j2}}{\sigma_2}\right)^2+\cdots+\left(\frac{x_{ip}-x_{jp}}{\sigma_p}\right)^2},\tag{2}$$

其中：σ_j是第j个分量的标准差，$j=1,2,\cdots,p$．与公式(1)相比，公式(2)可以看成是一种加权欧氏距离，权重为每一维数据方差的倒数．方差反映的是数据偏离中心的程度，数据分布越集中，方差越小，该维数据在计算距离时所占的权重越大；反之，数据越分散，该维数据在计算距离时所占的权重越小．标准化欧氏距离消除了各维数据方差分布不均带来的影响，但同时也使得“辨识度”高的分量在距离计算时权重降低．另外，若数据集中，某一维分量数值完全相同，则会出现数据点之间的距离无法用标准化欧氏距离度量的问题．

本文采用将数据“归一化”的方法来处理数据各维数据量纲差别过大的问题．对数据集中的第r维数据$\boldsymbol{x}_r=(x_{1r},x_{2r},\cdots,x_{nr})'$，其最大值$\max_r=\max(x_{1r},x_{2r},\cdots,x_{nr})$，归一化后的数据为：

$$\boldsymbol{x}_r^*=\left(\frac{x_{1r}}{\max_r},\frac{x_{2r}}{\max_r},\cdots,\frac{x_{nr}}{\max_r}\right)',\tag{3}$$

其中：$r=1,2,\cdots,p$；n为数据点的个数．

将归一化后的数据代入公式(1)，此时，两个p维向量$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})$与$\boldsymbol{x}_j=(x_{j1},x_{j2},\cdots,x_{jp})$间的距离公式变为：

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sqrt{\left(\frac{x_{i1}-x_{j1}}{\max_1}\right)^2+\left(\frac{x_{i2}-x_{j2}}{\max_2}\right)^2+\cdots+\left(\frac{x_{ip}-x_{jp}}{\max_p}\right)^2}\ . \tag{4}$$

2 孤立点处理

对孤立点的定义还没有统一的标准，有研究者用距离大小判断孤立点，认为孤立点是到其他所有点距离最大的那些点；也有研究者用密度值来判断孤立点，认为孤立点就是在给定半径内具有最大邻居数的点[11]．本文采用第二种定义．K-means算法中取一个类中所有对象的算术平均值作为聚类的中心，如果有孤立点，会严重影响聚类中心点，也就是说，K-means算法对孤立点敏感．对于K-means聚类算法中的孤立点，研究者或者不单独处理，或者采用“规避”的方法，即将孤立点分离出来，使其不参与聚类中心的选取[6,12]，或者使它向高密度区域移动[13]．本文根据美国著名经济学家托马斯•谢林于1971年提出的谢林模型的基本思想，将对孤立点进行移动处理．

谢林模型描述的是同质性对于空间隔离的影响和作用：人们需要和自己的同类(同种族的人、同收入群体等)居住在一起，当自己的邻居中同类邻居数量过少，他就会选择找一个新的、满意的地方居住，最终会导致隔离现象的产生，即使没有人刻意要求隔离的结果，但隔离也会出现[14-15]．

根据谢林模型，本文给予每一个数据对象以“主观能动性”，即数据对象是有感知能力的，当一个数据对象周围的邻居数量过少的时候，它就会移动到离它最近的、具有较多邻居的数据对象的位置．文献[13]利用Mean-shift算法，孤立点受到其他所有数据点的影响而发生漂移，移动到密度更高的区域，本文假设孤立点只受最近邻居的影响．

3 基于孤立点自适应的改进K-means算法

算法分为三个阶段．第一阶段，孤立点处理．假设孤立点在所有数据点中的占比小于等于$s\%$，邻居较多的点在数据点中的占比小于等于$w\%$．孤立点将会移动到距离它最近的、具有较多邻居的点上．第二阶段，选取初始聚类中心．在数据集中选取k个近邻密度较大的点作为初始聚类中心．第三阶段，聚类．基于孤立点自适应的K-means算法描述如下：

输入：含有n个样本的数据集合U，聚类个数k．

输出：k个簇的集合．

1)第一阶段：孤立点处理

①按照公式(3)对原始数据进行归一化处理．

②按照公式(1)计算归一化处理后的数据点两两之间的距离$d(\boldsymbol{x}_i,\boldsymbol{x}_j),\boldsymbol{x}_i,\boldsymbol{x}_j\in U$，对每一个数据点，按距离从小到大对其他数据点排序．

③计算数据点两两之间的平均距离meandist：

$$\text{meandist}=\frac{1}{C_n^2}\sum_{\boldsymbol{x}_i,\boldsymbol{x}_j\in U}d(\boldsymbol{x}_i,\boldsymbol{x}_j)\ , \tag{5}$$

式中：n为样本点总数；C_n^2是n个点中任取两个点的集合数．

④计算每个点的邻居数nabor(i)：

$$\text{nabor}(i)=\sum_{j=1}^{n}u\left(\frac{\text{meandist}}{2}-d(\boldsymbol{x}_i,\boldsymbol{x}_j)\right),\quad i=1,2,\cdots,n\ , \tag{6}$$

其中：$u(z)=\begin{cases}1 & z\geqslant 0\\ 0 & z<0\end{cases}$，若$d(\boldsymbol{x}_i,\boldsymbol{x}_j)\leqslant\frac{\text{meandist}}{2}$，$j$是$i$的邻居，否则不是．

⑤按邻居数从少到多对数据点排序.

⑥对于排在前面的$s\%$的每一个数据点o(孤立点):按距离从小到大依次检查o的每一个邻居,如果邻居i为多邻数据点(邻居数量排序>1−$w\%$),邻居i的坐标值赋值给o,停止检查.

否则,转到下一个邻居.

2)第二阶段:选取初始聚类中心

①复制由第一阶段得到的数据集合U到I.

②按照公式(1)重新计算数据点两两之间的距离$d(\boldsymbol{x}_i,\boldsymbol{x}_j),\boldsymbol{x}_i,\boldsymbol{x}_j\in I$,对每一个数据点,按距离从小到大对其他数据点排序.

③按照公式(5)重新计算数据点两两之间的平均距离meandist.

④计算每个点的近邻数量nabor(i):

$$\mathrm{nabor}(i)=\sum_{j=1}^{n}u\left(\frac{\mathrm{meandist}}{k+1}-d(\boldsymbol{x}_i,\boldsymbol{x}_j)\right),\quad i=1,2,\cdots,n,\tag{7}$$

其中:$u(z)=\begin{cases}1 & z\geqslant 0\\ 0 & z<0\end{cases}$,若$d(\boldsymbol{x}_i,\boldsymbol{x}_j)\leqslant\frac{\mathrm{meandist}}{k+1}$,$j$是$i$的近邻,否则不是.近邻搜索半径与类的数量$k$有关,$k$越大,搜索半径越小.

⑤按近邻数从大到小对数据点进行排序.

⑥将排序后的第1个点作为第1个聚类中心,将此点和与其距离小于$\frac{\mathrm{meandist}}{k+1}$的点从集合$I$中删除.

⑦重复此阶段步骤②~⑥,直到找到k个聚类中心为止.

3)第三阶段:聚类

①对第一阶段得到的数据集合U中每一个数据点,计算该点到每一个聚类中心的距离,找出最小距离,将该点添加到聚类中心对应的类中.

②重新计算每一类的中心.

③重复此阶段中步骤①、②,直到每个数据点所属的类不再发生变化.

在本算法中,孤立点的占比$s\%$、邻居较多点的占比$w\%$对聚类的结果有重要影响.通过不断地验证,取s=20,w=75.

4 实验结果与分析

4.1 实验数据与实验环境

为了检验上述算法的有效性,本文采用MATLAB 7.0编程环境,选择UCI数据库中的Iris、Wine、Zoo、Soybean 4个数据集作为测试数据集,这4个数据集为常用的知名数据集,已知其聚类结果可靠、并取得一致意见,适合做聚类分析的基准数据集[16].

Iris也称鸢尾花卉数据集,共有150条记录,通过花萼长度、花萼宽度、花瓣长度、花瓣宽度4个属性预测鸢尾花卉属于Setosa,Versicolour,Virginica 3个种类中的哪一类.Wine数据集包含来自3种不同起源的葡萄酒,共178条记录,具有13个属性,记录葡萄酒的13种化学成分,通过化学分析可以推断葡萄酒的起源.Zoo数据集也称动物园数据集,共有101个记录,由16个属性来描述样本,其中15个为布尔属性值{0,1},1个分类属性(腿的数量){0,2,4,6,8},分为7类.Soybean-small数据集也称大豆疾病数据,共有47个样本,具有35个属性,分为4类.

4.2 评价指标

聚类有效性指标主要可以分为三类:内部有效性指标、外部有效性指标和相对有效性指标.内部有效性指标指只依据数据集本身和聚类结果的统计特征对聚类结果进行评价.外部有效性指标指利用已知的外部信息与聚类结果进行比较,评价聚类效果.相对指标则是在聚类之前需要确定一个决策目标,然后使用不同的参数集运行聚类算法,基于之前建立的聚类准则评价聚类划分结果,并确定最优聚类划分和最佳聚类数[17].

本文采用外部有效性指标准Accuracy(*AC*)、Rand Index (*RI*)，*F*-measure(*F*)和内部有效性指标误差平方和(Sum of Squared Errors，*E*)对聚类结果进行评价．各指标计算公式如下：

$$AC=\frac{m}{N}\times 100\%\ , \tag{8}$$

其中：m为能够被正确分配到指定类的数据对象的个数；N为全体数据对象总数．

$$RI=\frac{\mathrm{tp}+\mathrm{tn}}{\mathrm{tp}+\mathrm{fp}+\mathrm{fn}+\mathrm{tn}}\times 100\%\ , \tag{9}$$

$$F=\frac{2\mathrm{tp}}{2\mathrm{tp}+\mathrm{fp}+\mathrm{fn}}\times 100\%\ , \tag{10}$$

其中：tp代表同类数据被分到同一簇的数据对个数；tn代表不同类数据被分到不同簇的数据对个数；fp代表不同类数据被分到同一簇的数据对个数；fn代表同类数据被分到不同簇的数据对个数．

$$E=\sum_{j=1}^{k}\sum_{x_l\in c_j}\left|\boldsymbol{x}_l-c_j\right|^2, \tag{11}$$

其中：$\boldsymbol{x}_l$是数据集中属于第j类的数据样本；c_j是第j类中所有样本的平均值．

4.3 实验结果

1)传统的误差平方和指标有效性

为了验证传统的误差平方和指标和聚类精度指标在评价聚类效果上是否一致，本文将完全正确的聚类结果与文献[7]得出的聚类结果对应的精度值和误差平方和(*E*)进行比较，结果如表1所示．

表1 误差平方和与聚类准确度的对比

Tab.1 Comparison between the accuracies and sums of squared errors

数据集	聚类结果	精确度r/%	误差平方和(*E*)
Iris	完全正确的聚类结果	100	89.30
	文献[7]得出的聚类结果	88.67	78.86
Wine	完全正确的聚类结果	100	5 232 680
	文献[7]得出的聚类结果	70.22	2 370 740
Zoo	完全正确的聚类结果	100	277.97
	文献[7]得出的聚类结果	73.27	224.07
Soybean-small	完全正确的聚类结果	100	218.06
	文献[7]得出的聚类结果	74.47	234.02

从表1可以看到，在Iris，Wine，Zoo 3个数据集上，组内误差平方和(*E*)和聚类精度对聚类结果的评价是相矛盾的；聚类精度为100%时，误差平方和反而越大．由于数据各维度量纲的不同，直接对原始数据聚类处理往往使结果存在比较大的偏差；对由原始数据得出的聚类结果直接以误差平方和(*E*)进行评价，往往也会得出错误的结论．

2)算法外部有效性

将本文算法与文献[3]、文献[7]所提算法的Accuracy(*AC*)、Rand Index (*RI*)，*F*-measure(*F*)在4个经典数据集上进行比较，结果如表2、图1、图2所示．

表2 三种算法的Accuracy值比较

Tab.2 Comparison of accuracy values with the three algorithms %

数据集	文献[3]算法	文献[7]算法	本文算法
Iris	88.67	88.67	96.00
Wine	70.22	70.22	89.33
Zoo	58.42	73.27	88.12
Soybean-small	74.47	74.47	97.87

从表2和图1、图2可以看到，与文献[3]、文献[7]中算法相比，本文算法在*AC*值、*RI*值，*F*值上均有明显的提高．

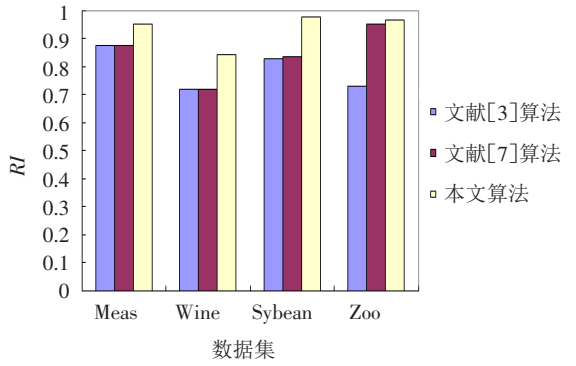

图1　三种算法的Rand Index值比较

Fig.1　Comparison of Rand Index values with the three algorithms

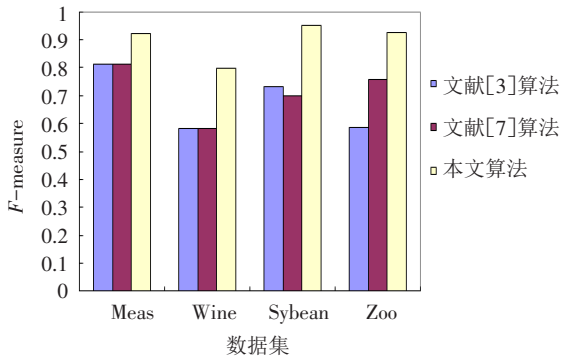

图2　三种算法的*F*−measure值比较

Fig.2　Comparison of *F*−measure values with the three algorithms

文献[3]没有考虑孤立点的影响，且在聚类中心选取过程中，以点与点之间的平均距离为邻居查找半径而忽视类簇的个数，很容易导致聚类中心的半径范围过大，聚类中心选取不够准确．文献[7]以数据点的方差作为选择数据中心的依据，很容易选出数据点的重心而非聚类中心．本文算法充分考虑了以上问题，取得了较好的聚类效果．

3）算法内部有效性

将3种算法的组内误差平方和（*E*）进行比较，结果如表3所示．

表3　三种算法的组内误差平方和（*E*）

Tab.3　Comparison of sums of squared errors with the three algorithms

数据集	文献[3]算法	文献[7]算法	本文算法
Iris	3.28	3.28	2.87
Wine	40.80	40.80	28.06
Zoo	202.06	156.82	116.52
Soybean−small	81.29	81.33	70.57

本文算法对原始数据进行了归一化处理，为了使*E*值具有可比性，在计算文献[3]和文献[7]算法的*E*值时，对这两种算法聚类后的数据进行了归一化处理．本文算法的*E*值是根据原始数据（非移动后的数据）与最终聚类中心计算得出．从表3可以看到，本文算法的*E*值明显低于另外两种算法．组内误差平方和与聚类精度对聚类效果的评价是一致的；聚类精度越高，组内误差平方和越小．因此，本文算法无论是从聚类的精确度还是从聚类结果的统计特征来看，都取得了良好的效果．

4）算法运行时间的比较

在相同的运行环境下，3种算法的迭代次数和运行时间如表4所示．

表4　三种算法的迭代次数和运行时间

Tab.4　Number of iterations and running times with the three algorithms

数据集	文献[3]算法		文献[7]算法		本文算法	
	迭代次数	运行时间/s	迭代次数	运行时间/s	迭代次数	运行时间/s
Iris	8	14.75	3	34.28	3	8.67
Wine	5	18.69	6	54.41	2	17.03
Soybean−small	4	0.83	5	2.09	4	1.72
Zoo	6	17.70	5	19.80	3	7.42

从表4可以看出，由于本文算法对离群点数据做了移动处理，初始聚类中心的选取也更合理，在进行K−means聚类时算法的迭代次数和运行时间与其他两种算法相比均有比较明显的降低．

5　结语

聚类分析是数据挖掘领域最重要的方法之一，K−means算法是一种最基础的聚类分析算法．离群点的

存在和初始聚类中心的选取是影响聚类结果的重要因素．本文首先对离群点问题进行研究，根据谢林模型的基本思想，使离群点移动到合适的位置，然后对初始聚类中心选取方法进行改进．实验结果表明，改进后的算法取得了较为理想的聚类效果．但是本文也存在一些不足之处：本文假设聚类个数k是已知的，但给定数据集，k往往是未知的，如果根据数据集的特征确定聚类个数k是作者今后研究的重点．

参考文献：

[1] 张文君，顾行发，陈良富，等. 基于均值－标准差的K均值初始聚类中心选取算法[J]. 遥感学报，2006，10(5)：715-721.
[2] 张健沛，杨悦，杨静，等. 基于最优划分的K-Means初始聚类中心选取算法[J]. 系统仿真学报，2009，21(9)：2586-2590.
[3] 韩凌波，王强，蒋正峰，等. 一种改进的K-means初始聚类中心选取算法[J]. 计算机工程与应用，2010，46(17)：150-152.
[4] 周炜奔，石跃祥. 基于密度的K-means聚类中心选取的优化算法[J]. 计算机应用研究，2012，29(5)：1726-1728.
[5] 黄敏，何中市，邢欣来，等. 一种新的K-means聚类中心选取算法[J]. 计算机工程与应用，2011，47(35)：132-134.
[6] 刑长征，谷浩. 基于平均密度优化初始聚类中心的K-means算法[J]. 计算机工程与应用，2014，50(20)：135-138.
[7] 谢娟英，王艳娥. 最小方差优化初始聚类中心的K-means算法[J]. 计算机工程，2014，40(8)：205-211.
[8] 程艳云，周鹏. 动态分配聚类中心的改进K均值聚类算法[J]. 计算机技术与发展，2017，27(2)：33-36.
[9] 王波，余相君. 自适应布谷鸟搜索的并行K-means 聚类算法[J]. 计算机应用研究，2018，35(3)：675-679.
[10] 金建国. 聚类方法综述[J]. 计算机科学，2014，41(11A)：288-293.
[11] 杨金花，刘显为. K-means聚类算法初始中心选择研究[J]. 河南科学，2016，34(3)：348-351.
[12] 叶施仁，杨英，杨长春，等. 孤立点预处理和Single-Pass聚类结合的微博话题检测方法[J]. 计算机应用研究，2016，33 (8)：2294-2297.
[13] CHENG Y. Mean shift，mode seeking，and clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1995，17(8)：790-799.
[14] THOMAS S. Dynamic models of segregation[J]. Journal of Mathematical Sociology，1972，1：143-186.
[15] THOMAS S. Micromotives and macrobehavior[M]. New York：W.W. Norton & Company，1978.
[16] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61.
[17] 开乐，杨善林，丁帅，等. 聚类有效性研究综述[J]. 系统工程理论与实践，2014，34(9)：2417-2431.

（编辑　张继学）