第41卷 第6A期
2014年6月

计 算 机 科 学
Computer Science

Vol. 41 No. 6A
June 2014

一种具有优良抗噪性能的初始聚类质心选择算法

马仕玉 李益才 蓝章礼

(重庆交通大学信息科学与工程学院 重庆 400074)

摘 要 K-means算法由于其固有的初始聚类质心敏感性,存在聚类结果不稳定、容易收敛到局部最优等问题。现有改进方案在处理无噪数据集时能够在降低迭代次数的同时得到近似全局最优解,但在处理有噪数据集时容易陷入局部最优,甚至聚类效果低于传统的K-means算法。在最远空间距离确定初始质心算法的基础上,提出一种基于空间距离差的初始质心选择算法。该算法的核心思想是通过计算非聚类质心点到已选质心的距离和,并排序,选取相邻距离差最大的两点中靠近已知质心的点作为下一个簇的初始质心而实现的。实验结果表明,所提算法在聚类迭代次数相当的情况下,对不含噪声数据集的聚类准确度增加约1%,对于含有噪声的数据集,聚类准确度达到90%以上。

关键词 K-means算法,初始质心,空间距离差,噪声数据

中图法分类号 TP391 **文献识别码** A

Novel Anti-noise K-means Algorithm Based on Spatial Distance Difference

MA Shi-yu LI Yi-cai LAN Zhang-li

(College of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China)

Abstract Due to the inherent initial clustering center sensitivity of K-means algorithm, it exists problems including result instability and being easy to fall into local optimum. The current improvement schemes can reduce the number of iteration and obtain an approximate global optimal solution when deal with noise-free data sets. But for noisy data sets, it would be easy to fall into local optimum, and the clustering result is lower than traditional K-means algorithm. Based on the algorithm that can find initial clustering centers according to the farthest spatial distance, the paper proposed a novel algorithm to select initial centers based on spatial distance difference. The main idea of the algorithm is calculating the sum distances between non-clustering center and all selected centers, then sort them. Choose the point which is the closer to the given centers as the new selected cluster center. Experimental results show that under the quite condition of iteration, when deal with noise-free data sets, the clustering accuracy of the proposed algorithm is improved about 1%. For noisy data sets, the classified accuracy is above 90%.

Keywords K-means algorithm, Initial centroid, Spatial distance difference, Noisy data

1 引言

聚类作为数据挖掘和机器学习领域中的重要数据分析工具[1],被广泛地应用在网络、经济、交通、生物等领域中。聚类分析具有多种不同的类型[2],包括基于层次的聚类、基于划分的聚类、基于密度的聚类、基于网格的聚类、基于模型的聚类。K-means算法[3]是最经典的基于划分的聚类算法,尽管K-means算法的目标函数是最小化等尺寸和等密度的球形簇,或者具有明显分离的簇,如果用户接受将一个自然簇分割成若干纯子簇,这种缺陷可以克服。其所具有的简单和适用于各种数据类型的优点,使它得到广泛的应用。但其对初始质心的过度依赖,导致K-means算法易陷入局部最优解,随机选择的初始质心会导致迭代次数过多[4]。针对K-means算法的不足,Fayyad等人采用多次取样数据集两次聚类以获取最优初始质心的思想,有效地解决了K-means算法对初始质心依赖性较大的问题[5];Adil等人通过寻找全局最佳初始质心和改进迭代算法来提高收敛速度[6];曹志宇等人通过计算最远空间距离寻找聚类初始质心以改善算法稳定性和聚类效果[7];张真等人提出基于移动网格技术的K-means质心点选取算法,在减少迭代次数的同时,近似得到误差平方和(Sum of the Squared Error, SSE)的全局最优[8]。

在基于射频识别(radio frequency identification, RFID)的车辆运行特征提取与分析的研究过程中发现,采集的车辆运行特征数据存在奇异点(噪声点数据),使得在利用上述改进方法进行数据处理时聚类效果不理想。针对前面研究者提出的算法对噪声点敏感、抗噪声性能不强的问题,结合车辆运行特征提取和分析的实际需要,本文提出了一种基于空间距离差的初始质心选择算法,用于K-means聚类分析中的初始质心选择。

本文受重庆市交通委员会科学计划项目:基于RFID的车辆非法营运监控与特征提取资助。

马仕玉(1989—),女,硕士生,主要研究方向为数据挖掘与决策支持;李益才(1970—),男,副教授,主要研究方向为人工智能与模式识别,E-mail:13452855299@163.com(通信作者);蓝章礼(1973—),男,教授,主要研究方向为交通信息化、交通图像处理与识别、太阳能发电技术、结构健康监测等。

2 基于最大距离差的 K-means 聚类算法

2.1 初始聚类质心对噪声数据集聚类效果的影响

K-means 聚类算法存在的初始质心敏感性，导致算法不稳定，常出现局部最优解等问题，而噪声点的存在将会加剧局部最优解的出现。在随机选取初始聚类质心过程中，噪声点被指定为初始值的概率是随机的，一旦其被指定为初始值，不仅会使簇质心严重偏离数据集样本空间的密集区域，还会出现噪声点被聚为独立簇的情况，有损于聚类结果的准确性。

假设 K-means 聚类数据对象为 p 维数据，所有数据都是 p 维空间中的一个点。考虑邻近性度量为欧几里德距离的数据，常使用 SSE 作为度量 K-means 聚类的目标函数，SSE 定义如式(1)所示：

$$SSE=\sum_{i=1}^{k}\sum_{x\in C_i}|x-c_i|^2 \tag{1}$$

其中，C_i 为第 i 个簇中元素的集合，x 是 p 维空间中的一个点，c_i 是第 i 个簇的聚类质心，使族 SSE 最小的聚类质心是均值，即 c_i 为：

$$c_i=\frac{1}{m_i}\sum_{x\in C_i}x \tag{2}$$

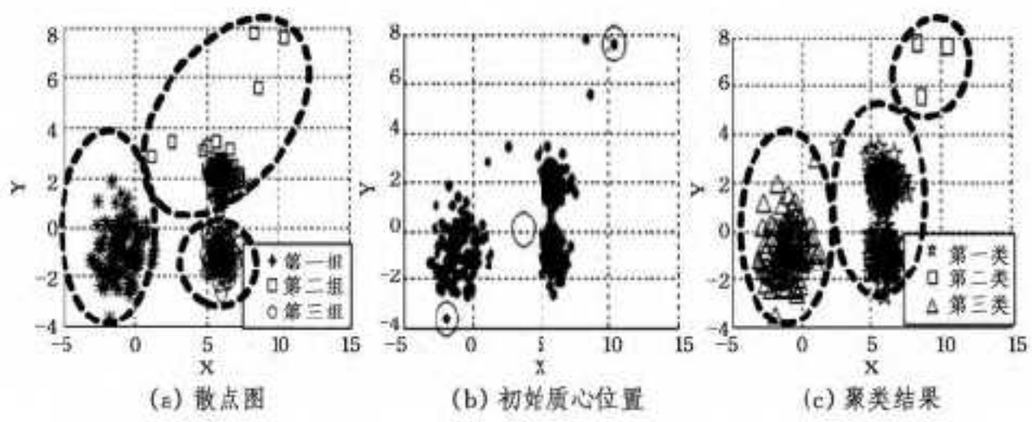
根据 K-means 算法的目标函数和聚类质心的计算规则可知：①当每一个簇对(两个相隔较近的簇)有两个聚类质心点时，不管这两个质心点在一个簇中或分别在两个簇中，随着聚类算法的执行，最终都能得到最优解。初始聚类质心分布不均匀时会使得算法中聚类的迭代次数增加。②当初始分配的聚类质心点由于某种原因而使得某个簇对只有一个质心点而另一个相隔较远的簇有多于两个初始聚类质心点时，则 K-means 算法最终执行的结果是两个自然簇的合并和一个自然簇的分离。③当初始分配的聚类质心使得某个簇对只有一个质心点而相隔较远的离群点被分配为聚类初始质心时，最终的结果是相隔较近的离群点作为一个簇，而真正的相隔相对较近的两个自然簇合并，从而造成聚类准确度的大幅降低。

在对 K-means 算法的研究中，文献[7]结合空间距离度量选择数据集中空间距离最远的点作为初始聚类质心，该算法能有效地解决上述分析中所提出的第二类问题，在处理无噪数据集时能求得全局最优值，并且由于初始聚类质心相距较远和均匀，算法的迭代次数也得到有效的降低。但由于是以距离最远为初始聚类质心的选择标准，因此在处理有噪数据集时，当噪声点作为远离数据密集区域的奇异点存在时，在空间最远距离法选取初始质心中，部分噪声点会被选为初始质心，将导致簇质心偏离数据密集区域以及相邻噪声点被聚为一类的结果，影响聚类结果的准确度，导致算法的分类效果不理想。

图 1 结合文献[9]所给出的方法，人工生成 3 组含有噪声的二维数据集。数据集中的样本分布符合正态分布，表 1 为各组数据的均值及标准差等参数，其中在第三组数据中加入了适量的噪声数据点。

表 1 随机二维数据集各参数设置

	第一组	第二组	第三组	噪声组
均值	$\mu_{x1}=-1$	$\mu_{x2}=6$	$\mu_{x3}=6$	$\mu_x=6$
	$\mu_{y1}=-1$	$\mu_{y2}=-1$	$\mu_{y3}=2$	$\mu_y=2$
标准差	$\sigma_1=1$	$\sigma_2=0.5$	$\sigma_3=0.5$	$\sigma=3$
数量	120	120	115	5



(a) 散点图 (b) 初始质心位置 (c) 聚类结果

图 1 空间最远距离法对人工噪声数据集的分析

图 1(a)为 3 组随机数据集的散点图，图中有明显的噪声数据点存在。在基于最远空间距离的初始聚类质心选择算法中，噪声点会被选为初始聚类质心，如图 1(b)所示。在下一步的聚类分析中，作为初始聚类质心的噪声点与临近的几个噪声点被分到同一个簇，并作为单独的簇存在，而其余部分非同组数据被错误地划分到了同一个簇中，如图 1(c)所示。

2.2 本文所提初始聚类质心选择算法

根据以上分析，选择初始聚类质心必须遵循两个原则：①不能选择到离群点中距离数据稠密区最远的点，否则将会导致初始聚类质心不能自动调整到数据稠密区，聚类目标函数陷入局部最优；②所选择的初始聚类质心应尽可能地分散，以有效降低迭代次数。

根据上述两种原则，文献[7]所提的以空间距离最远作为选择初始聚类质心的算法虽然能够保证初始聚类质心有效分散，但常常会选择到离群点中距离数据稠密区最远的点，从而导致聚类准确度的下降

因此，初始质心的选取不再以空间最远距离为标准，而是在空间距离的基础上，以空间距离差为标准，选取距离差最大的两个点中离已知质心较近的点为初始质心。这样既保证了所选择的初始聚类质心比较分散，也保证了所选择初始聚类质心靠近数据稠密区，有效地解决了偏远的离群点作为初始聚类质心所带来的 K-means 算法陷入局部最优的问题。

本文所提初始聚类质心选择算法如下所述：

1. 初始化 K 值和初始化质心集合 $C=\varnothing$。

2. 计算数据集 $D=\{X_1,X_2,\cdots,X_n\}$ 的均值：

$$c_1=\frac{1}{n}\sum_{X\in D}X \tag{3}$$

式中，$X=(x_1,x_2,\cdots,x_p)$ 是 p 维空间中的一个点。

3. 将 c_1 作为选择出的第一个初始聚类质心，放入聚类质心集合 C 中。

$$C=C\cup\{c_1\} \tag{4}$$

4. 设已选择初始聚类质心 C 中元素个数为 i，计算数据集 $D=\{X_1,X_2,\cdots,X_n\}$ 中所有非聚类质心元素到所有已选择初始聚类质心的距离之和，并进行排序。

$$d_j=\sum_{c_l\in C}|X_j-c_l|,j=1,2,\cdots,n-i+1 \tag{5}$$

式中，$X_j\in D$，且 $X_j\notin C$。

5. 按已排序的距离 $(d_1,d_2,\cdots,d_{n-i+1})$ 计算相邻两距离之间的距离差，并选择距离差的最大值。

$$d_{diff}=\max(|d_{j+1}-d_j|),j=1,2,\cdots,n-i \tag{6}$$

6. 从与 d_{diff} 相对应的两个元素中找出离 C 中元素距离和小的那个元素 X，作为一个初始聚类质心放入 C 中。

$$C=C\cup\{X\} \tag{7}$$

7. 如果 C 中元素数目达到 K 值，则初始质心寻找完毕，

返回聚类质心集合 C，否则返回第 4 步继续寻找下一个质心，直到得到 K 个质心为止。

当数据集中数据元素个数 n 很大时，采取抽样的方式来得到用于确定初始聚类质心的抽样数据集 D。此时抽样数据集中的元素个数 m 远小于数据集元素个数 n。

根据 K-means 算法思想，初始质心选择完毕后，将非质心点指派到距离最临近质心，形成 K 个簇，更新簇质心。重复执行指派和更新簇质心的步骤，直到簇质心不发生变化或 SSE 的变化小于给定阈值为止。这时就可以得到对整个数据集的聚类结果。

2.3 算法分析

传统 K-means 算法由于是随机指定聚类初始质心，其时间复杂为 $O(nkT)$[10]，其中 n 为数据集中样本个数，k 为分类个数，T 为算法迭代次数。空间最远距离法与本文算法增加了初始聚类质心的选择，如果采用数据抽样的方式进行，假设抽样数据个数为 m，则时间复杂度分别为 $O(km+mkt)$，$O(km^2+2km+mkt)$，其中 t 为聚类过程中的迭代次数，$O(km)$ 为空间最远距离法选择初始质心所花费的时间，$O(km^2+2km)$ 是本文算法消耗在初始质心选择上的时间。由于本文所选初始质心靠近数据分布密集区域，因此 $t<T$，又由于 $m \ll n$，因此本文算法相较传统 K-means 算法并未增加时间复杂度。

3 实验结果及分析

为检验本文所提改进算法的有效性，实验分别采用两类数据集对本文所提算法进行测试，一类是无噪声数据集，另一类则是上文给出的含噪声数据集。

3.1 聚类结果评价指标的说明

本文除采用常用聚类指标迭代次数和分类准确率外，还采用 Jaccard 系数、Rand 指数[11,12] 和 Adjusted Rand Index (RI) 参数[13] 以全面直观地观察不同算法的收敛速度和聚类质量。其中，后 3 个聚类有效性指标的定义如下：设 C 是数据集的已知划分，S 是数据集的一个聚类划分，对数据集中任意一对数据点，计算如下几项：

a) 属于 C 中同一类 S 中同一簇的数据对数目；

b) 属于 C 中同一类 S 中不同簇的数据对数目；

c) 属于 S 中同一簇 C 中不同类的数据对数目；

d) 属于 C 中不同类 S 中不同簇的数据对数目。

Rand 指数、Jaccard 系数和 RI 参数分别表示如下：

Rand 指数：$Rand=(a+d)/M$

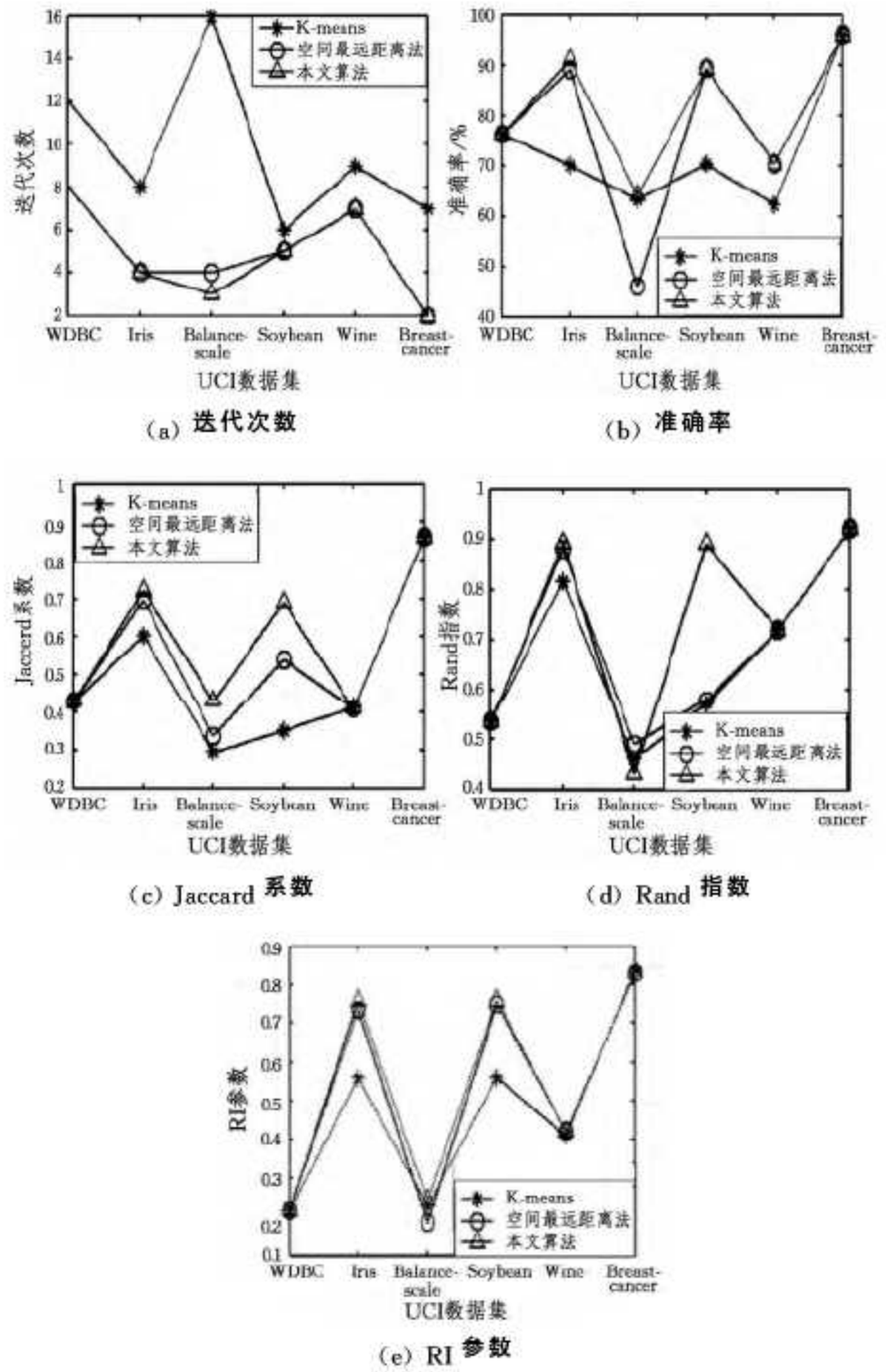
Jaccard 系数：$Jaccard=a/(a+b+c)$

RI 参数：$RI=\dfrac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$

从定义可知，以上 3 个聚类有效性评价指标都是有监督度量，其取值越大说明聚类效果越好。

3.2 无噪声数据集算法测试

无噪声数据集采用 UCI 数据库中的 WDBC、Iris、Balance-scale、Soybean、Wine 和 Breast-cancer 数据集，分别用传统 K-means 算法、空间最远距离法及本文算法对 6 个数据集进行聚类分析，利用迭代次数、聚类准确率、Rand 指数、Jaccard 系数和 RI 参数来评价 3 种算法的聚类效果。实验中对传统的 K-means 算法进行 10 次测试，取其均值进行 3 种算法的比较。UCI 数据集的实验结果如图 2 所示。



(a) 迭代次数　(b) 准确率

(c) Jaccard 系数　(d) Rand 指数

(e) RI 参数

图 2　3 种算法在 UCI 数据集上的分类结果比较

由图 1 中(a)—(e)图可以看出，改进后的两种算法与 K-means 算法相比，各指标参数都得到较为显著的优化，收敛速度平均提高 1 倍。在对 6 个数据集的处理中，本文改进算法又比空间最远距离法得到进一步的优化，准确率提高了 1% 以上，其它参数也有相应程度的提高。

3.3 有噪声数据集算法测试

为消除人工产生的随机数据的偶然性，实验中分别产生了 20 次随机数据集进行分析，并统计空间最远距离算法与本文算法下的分类结果。图 3 为两种算法的准确率和迭代次数曲线图。

经过 20 次不同的数据处理结果对比表明，本文所提算法在处理含有噪声点的数据集时分类效果明显优于空间最远距离法，聚类准确率保持在 90% 以上，迭代次数与空间最远距离法相当，聚类效果稳定。

实验表明，在数据集无噪声的情况下，本文所提算法在迭代次数与最远空间距离法相当的情况下，聚类准确率、Rand 指数、Jaccard 系数和 RI 参数都有不同程度的提高。在数据集有噪声的情况下，经过反复测试，结果表明本文所提算法在迭代次数与最远空间距离法相当的情况下，聚类准确率、Rand 指数、Jaccard 系数和 RI 参数等评价指标比最远空间距离法有显著的提高。其中，本文所提算法对有噪数据集的聚

(下转第 420 页)

[3] Coelho T,Calado P,Souza L,et al. Image retrieval using multiple evidence ranking[J]. IEEE Transactions on Knowledge and Data Engineering,2009,16(4):408-417

[4] Theobald M,Siddharth J,Paepcke A. SpotSigs:Robust and efficient near duplicate detection in large Web collections[C]//Proceedings of the 31st Annual International ACMSIGIR Conference on Research and Development in Information Retrieval. Singapore,2008:563-570

[5] Erkan G,Radev D. Lexrank:Graphbased lexical centrality as salienceinin text summarization[J]. Journal of Artificial Intelligence Research,2009,22(7):457-479

[6] Ko Y,Park J,Seo J. Improving text categorizationusingthe importance of sentences[J]. Information Processing and Management,2010,40(1):6579

[7] 中医药学语言系统 Wi-ki[EB/OL]. http://www.cintcm.com/yuyan/index.htm,2013-05-01

[8] 医药在线交易服务平台 Wi-ki[EB/OL]. http://www.yaol.cn/,2013-07-01

[9] VSMWi-ki[EB/OL]. http://en.wikipedia.org/wiki/VSM,2012-03-19

[10] Oleshchuk V. Ontology based semantic similarity comparison of documents [C]//14th International Workshop on Database and Expert Systems Applications,2003. 2003,1

[11] Ding C H Q. Research on Optimize Technology in Latent Semantic Indexing Based on Semantic Block[C]//Chinese Conference on Pattern Recognition,2009(CCPR 2009). 2009

[12] 刘群,李素建. 基于《知网》的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会,2002

[13] 晋耀红. 基于语境框架的文本相似度计算[J]. 计算机工程与应用,2004(16)

[14] 颜端武,成晓,甘利人. 基于领域本体和概念向量的中文文本相似性测度研究[J]. 中国图书馆学报,2007,33(6)

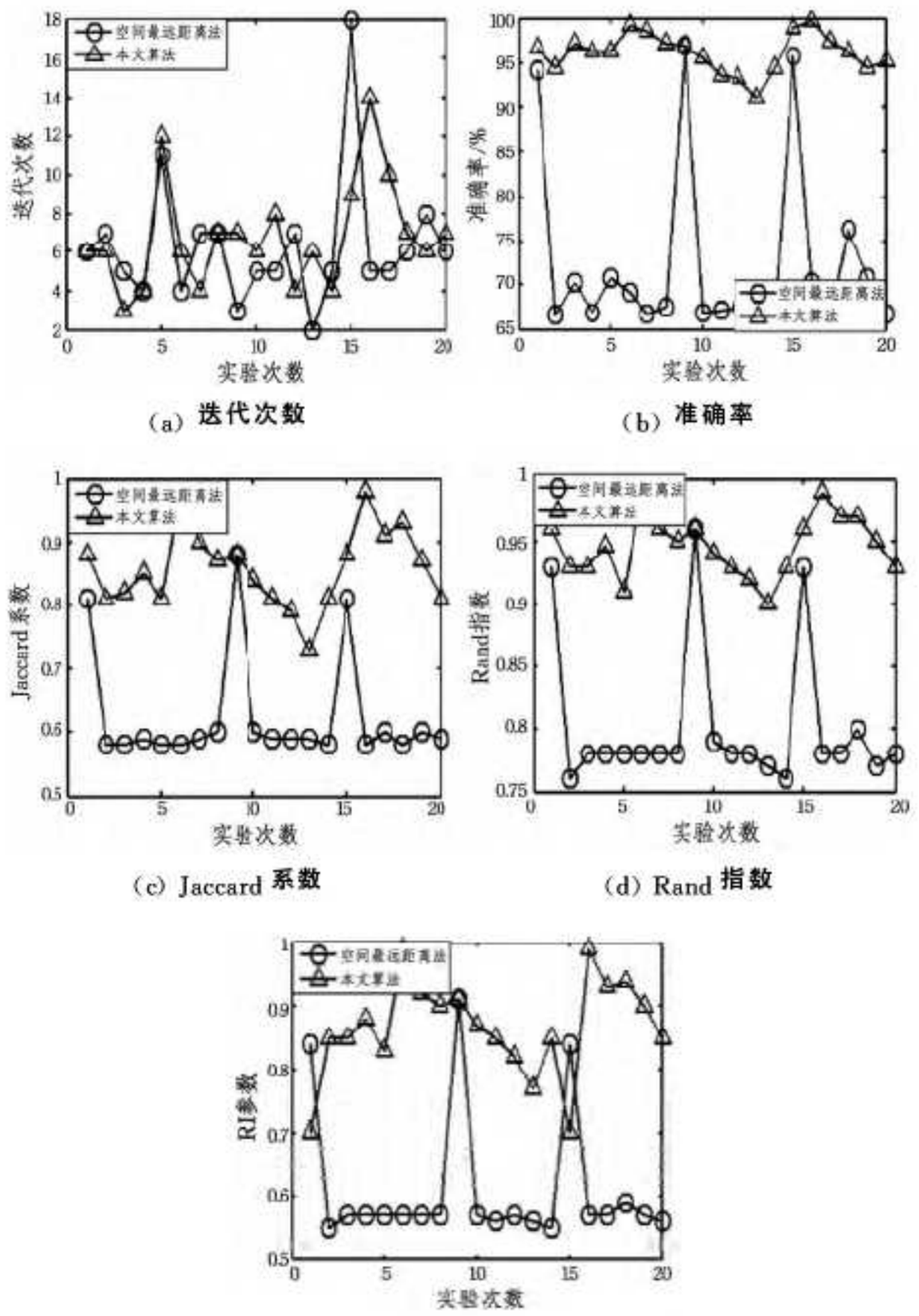
[15] 穗志方,俞士汶. 基于骨架依存树的语句相似度计算模型[C]//中文信息处理国际会议. 1998

[16] 黄慧,印鉴,侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报,2011(5):856-864

[17] TF-IDFWi-ki[EB/OL]. http://zh.wikipedia.org/wiki/TF-IDF,2013-05-01

(上接第 408 页)

类准确率达到 90% 以上,具有优良的抗噪性能。



(a) 迭代次数 (b) 准确率

(c) Jaccard 系数 (d) Rand 指数

(e) RI 参数

图 3 空间距离法与本文算法在噪声数据集上的分类效果比较

结束语 本文针对 K-means 聚类算法中初始质心的选择,改进基于空间最远距离方法存在的不足,采用空间距离差法寻找初始聚类质心,从而降低噪声点对初始聚类质心选择的影响,以提高 K-means 聚类分析的准确性。通过使用 UCI 数据库中的数据集和带有噪声点的人工随机数据测试,本文所提算法效果稳定,处理噪声数据集的能力得到提升,能够用于实际采集的含有噪声大数据集的分析和处理。

参 考 文 献

[1] 欧陈委. K 均值聚类算法的研究与改进[D]. 长沙:长沙理工大学,2011

[2] 张俊生. 数据挖掘中的聚类方法及其应用研究[D]. 天津:天津理工大学,2012

[3] Anil K J. Data clustering:50 year beyond K-Means[J]. Pattern Recognition Letters,2010,31(08):651-666

[4] 吴晓蓉. K-均值聚类算法初始质心选取相关问题的研究[D]. 长沙:湖南大学,2008

[5] Fayyad U,Reina C,Bradley P S. Initialization of Iterative Refinement Clustering Algorithms[C]//Proc of the Fourth International Conference on Knowledge Discovery and Data Mining. 1998:194-198

[6] Adil M,Bafirov,Julien U. Fast modified global k-means algorithm for incremental cluster construction[J]. Pattern Recognition,2011,44(4):866-876

[7] 曹志宇,张忠林,李元韬. 快速查找初始聚类质心的 K-means 算法 [J]. 兰州交通大学学报,2009,28(6):15-18

[8] 张真,任贺宇. 一种基于动态网格技术的 K-means 初始质心选取算法[J]. 微电子学与计算机,2013,30(6):101-104

[9] 谢娟英,蒋帅,王春霞. 一种改进的全局 K-均值聚类算法[J]. 陕西师范大学学报,2010,38(2):18-22

[10] 谢娟英,郭文娟,谢维信. 基于样本空间分布密度的初始聚类中心优化 K-均值算法[J]. 计算机应用研究,2012,29(3):888-892

[11] 杨燕,靳蕃,Kamel M. 聚类有效性评价综述[J]. 计算机应用研究,2008,25(6):1631-1632

[12] 张惟皎,刘春煌,李芳玉. 聚类质量的评价方法[J]. 计算机工程,2005,31(20):10-12

[13] Hubert L,Arabie P. Comparing partitions [J]. Journal of Classification,1985,2(1):193-218

[14] 王纵虎,刘志镜子,陈东辉. 基于粒子群优化的模糊 C-均值聚类算法研究[J]. 计算机科学,2012,39(9):166-169