改进的 k-means 算法在三支决策中的应用研究*

蔺艳艳 陆介平 王郁鑫 傅廷妍
（江苏科技大学计算机学院 镇江 212001）

摘 要 针对传统 k-means 算法不适用有不确定因素存在的环境和现有的三支 k-means 聚类分析中并未避免传统 k-means 算法随机选择初始簇中心而导致聚类结果不稳定的问题，论文提出一种改进的 k-means 算法，借助层次聚类算法和数学抽样方法，结合定义的聚类结果评估有效性指数，获得一组较优的初始中心，并将其作为 k-means 算法的初始簇中心，然后引进三支决策聚类理论方法进行聚类结果的优化，使其适应具有不确定因素的环境。实验表明，此方法在 UCI 数据集上的聚类效果、准确率和稳定性均有所提高。

关键词 聚类；有效性指数；k-means 算法；三支聚类
中图分类号 TP391.41 **DOI**：10. 3969/j. issn. 1672-9722. 2020. 06. 007

Application Research of Improved k-means Algorithm in Three Decisions

LIN Yanyan LU Jieping WANG Yuxin FU Tingyan
（School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang 212001）

Abstract The traditional k-means algorithm is not applicable to the environment with uncertain factors and the existing three k-means clustering analysis does not avoid the problem that the traditional k-means algorithm randomly selects the initial cluster center and leads to unstable clustering results. In this paper, an improved k-means algorithm is proposed. By using hierarchical clustering algorithm and mathematical sampling method, combined with the defined clustering results to evaluate the validity index, a set of better initial centers is obtained and used as k-means algorithm. The initial cluster center, then introduces three decision clustering theory methods to optimize the clustering results to adapt to the environment with uncertain factors. Experiments show that the clustering effect, accuracy and stability of this method on the UCI dataset are improved.

Key Words clustering, validity index, k-means algorithm, three branch clustering
Class Number TP391.41

1 引言

随着互联网在生产生活中应用的越来越广泛，随之产生的是大量的数据。这些数据往往潜藏着用户的行为动向或某个行业的发展规律。如何从这些海量数据中挖掘到有价值的信息是我们今天的研究热点。聚类是数据挖掘中的一个非常重要的分支，它是根据信息相似度原则在预先不知道预划分类的情况下进行信息聚类的一种方法。k-means 算法[1]是一种典型的基于划分的聚类算法，自 1967 年 MacQueen 提出后，由于其具有算法简单易懂且收敛速度快的优点，得到较普遍的应用[2]。比如，利用 k-means 聚类方法来对客户进行准确的分类，其结果可以作为企业优化营销资源的重要依据等等。

但是，传统的 k-means 算法是二支聚类，它不适用具有不确定因素的环境，并且 k-means 算法需要预先随机选取初始聚类中心和设定聚类数目[3]，这些因素将导致聚类结果不稳定，影响其精确性。有许多学者对 k-means 算法进行研究和改进，Feng

* 收稿日期：2019 年 12 月 10 日，修回日期：2020 年 1 月 20 日
作者简介：蔺艳艳，女，硕士研究生，研究方向：数据挖掘、粗糙集、语义分析。陆介平，男，博士，硕士生导师，研究方向：数据挖掘、粗糙集、自然语言处理。王郁鑫，男，硕士研究生，研究方向：语义分析、数据挖掘。傅廷妍，女，硕士研究生，研究方向：语义分析、数据挖掘。

等改进 k-means 算法[4] 是根据数据点的距离构造最小生成树，并加以剪枝来构造样本分布情况，从而动态选取初始聚类中心。Yu 等提出了一种结合关系矩阵和度中心性（Degree Centrality）的分析方法，从而确定 k-means 算法初始的 k 个中心点[5]。为了能自动选择 k-means 算法的初始 k 值，Debatty 等提出了 G-means 算法[6]，Yuan 等加入密度参数并通过计算平均密度来降噪[7]，Kettani 等提出 AK-means 算法[8]等。但是，这些方法对局部最优、聚类结果不稳定和总迭代次数多等问题进行优化，并且在处理不确定性信息时，考虑到当前获取到的信息不够充分的特点，如果强制将其中的元素划分到一个类中，往往容易带来较高的错误率或决策风险。基于此，Hoppner 等提出用模糊集表示聚类结果的模糊聚类理论方法[9]，Lingras 提出了粗糙聚类方法，将聚类结果由粗糙集的正域，边界域和负域来表示[10~13]，Yao 等用区间集来表示聚类结果中的一个类[14]等，这些方法计算复杂，对指标权重矢量的确定主观性较强。三支决策聚类理论的提出，有效地改善了传统聚类方法处理具有不确定性因素的问题，并且可与传统聚类算法结合，计算相对简单。

三支决策聚类是一种重叠聚类，早先由 Yao、Yu[15~17]等提出，它采用核心域、边界域和琐碎域来表示每个类别，其中边界域中的元素是介于核心域和琐碎域之间的元素，集中对边界域中的元素判断处理，可以较好地处理具有不确定性对象的聚类问题。在三支决策理论中，传统的 k-means 二支聚类算法是一种特殊的三支聚类，即边界域中的元素为空。有学者 Li[18] 利用传统 k-means 聚类算法产生的结果和每个类中元素的邻域所在的集合进行收缩与扩张，来研究将二支聚类转化为三支聚类的方法，达到提高聚类结果的数据结构的目的。但是，这些方法均是直接使用传统 k-means 算法，聚类结果的准确性和确定性受 k-means 算法缺点影响。

针对传统 k-means 算法不适用具有不确定因素的环境和现有的基于 k-means 的三支聚类分析中并未避免传统 k-means 聚类随机选择初始簇中心而导致聚类结果不稳定的问题，本文提出一种改进的 k-means 聚类方法，避免了传统 k-means 算法由于初始簇中心选择的随机性而导致聚类结果不稳定的现象；其次，为了避免传统 k-means 算法在处理不确定性信息时，强制将其中的元素划分到一个类中带来的错误率或决策风险，将改进的 k-means 算法与 Wang 提出的将二支聚类结果转换成三支聚类方法结合起来，研究本文所提出的方法在三支决策中的应用，以提高聚类结果的结构和精度，实验结果证明这种方法是有效的。

2　相关知识

2.1　k-means 算法

k-means 算法[19] 原理简单易懂，时间复杂度低，仅为 $O(kNT)$，并且具有计算简单、高效等特点，广泛应用在生产生活的各个领域。k-means 算法基于样本间相似度原则，采用两样本间的欧氏距离远近作为衡量标准进行数据集划分。k-means 的算法理论是：在数据集 D 中，先随机选取 k 个样本作为初始中心，再计算剩下的所有样本到这一组初始中心的欧氏距离，根据距离最近原则将各个样本归入到相应的聚类中心所在的类，然后计算每个类的新均值，重新修正聚类中心。不断迭代更新，直到误差平方和函数稳定在最小值。

设数据集集合 $D=\{x_1, x_2, \cdots, x_n\}$，$x_i=(x_{i1}, x_{i2}, \cdots, x_{ir})$，$x_j=(x_{j1}, x_{j2}, \cdots, x_{jr})$，则样本 x_i 和样本 x_j 之间的欧氏距离为

$$d(x_i, x_j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ir}-x_{jr})^2} \tag{1}$$

误差平方和准则函数如下：

$$J_c=\sum_{i=1}^{K}\sum_{j=1}^{r_i}\left\|x_j-n_i\right\|^2 \tag{2}$$

其中：k 为聚类类别数，r_i 为第 i 类中的样本的个数，n_i 为第 i 类中样本的平均值。

2.2　三支决策聚类的相关概念

三支决策理论早先是由 Yao[14] 等提出，其核心思想是在粗糙集决策聚类上将传统的二支决策聚类拓展为三支决策聚类，使它成为更符合现实情况，更适应用来作为决策依据，降低因聚类错误而产生的成本。Yu[15] 将三支决策的思想引入聚类中，提出用三个互不相交的集合表示一个类的三支聚类，这三个集合是核心区域、边界区域和琐碎域，分别用 C_i^P, C_i^B, C_i^N 来表示。核心区域中的元素确定属于这个类，边界区域中的元素可以属于也可能不属于这个类，琐碎域中的元素肯定不属于这个类。

设一个给定数据集 $U=\{x_1, x_2, \cdots, x_i, \cdots x_n\}$，则 C_i^P, C_i^B, C_i^N 分别满足：

$$C_i^P \cup C_i^B \cup C_i^N = U \tag{3}$$

$$C_i^P \neq \varnothing,\ i=1, 2, \cdots, k \tag{4}$$

$$\bigcup_{i=1}^{k}\left(C_i^P \cup C_i^B\right)=U \tag{5}$$

其中式(3)表明,可通过 C_i^P 和 C_i^B 来表示一个类,式(4)要求正域非空,即每个类中至少有一个对象,而式(5)保证每个对象至少被分到一个类中。与二支聚类结果不同,三支聚类的结果可由下式来表示:

$$TC=\left\{\left(C_1^P, C_1^B\right),\left(C_2^P, C_2^B\right),\cdots,\left(C_k^P, C_k^B\right)\right\}$$

2.3 基于传统k-means算法的三支聚类

二支决策的聚类结果是对象一定属于两个类中的一个,而三支决策的聚类结果是:对象确定属于某类、可能属于某类或确定不属于某类。可以说二支决策是三支决策类结果中的一种,即不存在边界区域。

基于传统的k-means算法的三支聚类其实是对二支聚类结果的进一步优化。假设数据集 $U=\{x_1,x_2,\cdots,x_i,\cdots x_n\}$,利用传统k-means算法对 U 进行二支决策聚类的结果是:$C=\{C_1,C_2,\cdots,C_k\}$。假如是图1所示的数据集。

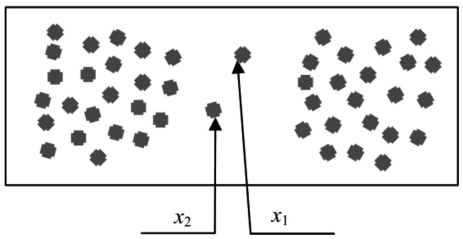

图1 数据集图

如果删除 x_1 和 x_2,则很容易地将图1聚成两个结构特征非常好的类,而无论将 x_1 或 x_2 放到哪一个类中均会降低这个类的紧致性。基于三支决策聚类的核心思想,将 x_1 和 x_2 放到两个类的边界域当中去。定义一个 θ 域(距离该点最近的 θ 个点组成的集合),使 θ 域内的点在二支聚类的结果下不完全包含于某个类中,例如 x_1 和 x_2 这类点。这样先采用k-means聚类的结果,再结合 θ 域(边界域)的进行决策聚类的方法,便是基于传统k-means算法的三支聚类。

3 改进的k-means算法在三支决策中的应用研究

3.1 改进的k-means算法

传统的k-means算法的缺点是随机地选取任意 k 个样本作为初始聚类中心,这种随机性会影响最终聚类结果的稳定性。本文提出的k-means算法的改进能够克服上述问题。首先,先对数据集进行凝聚层次聚类,并采用轮廓系数对不同层次划分进行评估,获得较为合理中心数 K。再对数据集进行 n 次样本抽取(n 次抽取的样本总数要大于等于原始样本数),并以层次聚类所得 K 值做为输入对其进行k-means聚类,从而得出一组中心。然后计算这n组中心的误差平方和准则函数值,选择值最小所对应的聚类中心作为初始聚类中心。最后将其作为k-means三支聚类算法的初始中心和 k 值的输入,避免了k-means算法随机选择初始中心和 k 值而导致最终三支聚类结果不稳定的问题。其中,多次抽取样本可以保持随机性不被破坏。层次聚类算法可以在其聚类结果上采用轮廓系数对数据集的不同层次划分进行评估,在初始中心确定前先初步优化簇数 K,最终的初始中心是由收敛函数来选取。为防止因适应误差平方和准则函数而陷于局部最优或导致簇过度划分,可以采取设定初始聚类中心数大于指定的 K 值,并且设定准则函数收敛标准,使达到收敛后的初始聚类中心数在合并距离近的簇之后数量减少到指定的 K 值。

层次聚类中需要用的公式具体如下:

设数据集 $D=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$,每个数据对象具有 p 个特征,即 $x_i=\{x_{i1},x_{i2},\cdots,x_{ip}\}$。在层次聚类算法中,计算数据对象间的两两距离的欧氏距离公式为

$$dist\left(x_i,x_j\right)=\sqrt{\sum_{m=1}^{p}\left(x_{im}-x_{jm}\right)^2} \tag{6}$$

在计算各类间的相似度时,本文采用了均链接(average-linkage),具体公式如下:

$$d_{avg}\left(C_i,C_j\right)=\frac{1}{\left|C_i\right|\left|C_j\right|}\sum_{x\in C_i}\sum_{z\in C_j}\left\|x-z\right\| \tag{7}$$

其中:$d_{avg}\left(C_i,C_j\right)$ 表示 C_i、C_j 间的相似度,$\left\|x-z\right\|$ 表示数据对象 x 和 z 的距离,$\left|C_i\right|$ 表示类 C_i 中数据的个数,$\left|C_j\right|$ 表示类 C_j 中数据的个数。

3.2 改进的k-means算法在三支决策中的应用研究

本文将层次聚类和科学抽样法引入到k-means算法中,是为了获取最优初始簇类的数目,避免k-means算法随机选择 k 值和初始聚类中心而导致分类结果误差较大、聚类结果不稳定等问题。然后将改进后的k-means算法聚类的结果做为三支聚类的初始输入,实现最优类别划分的问题。

前文提到的 θ 域这里给出详细定义:

定义（θ 域）：设数据集 $U=\{x_1,x_2,\cdots,x_n\}$ 采用 k-means 聚类算法的结果为 $C=\{C_1,C_2,\cdots,C_k\}$，$x_m\in C_i(i=1,2,\cdots,k)$，$\theta(x_m)$ 是在 C_i 距离 x_m 点最近的 θ 个点组成的集合，则 $\theta(x_m)$ 是 x_m 的 θ 域。

因为 θ 的选取也会对最终聚类结果产生影响，所以选择合适的 θ 很重要，这里选取的 θ 是数据集中样本数目的开平方，即 $\sqrt{N}$ 。

算法 1：改进的 k-means 算法在三支决策中的应用

输入：数据集 $U=\{x_1,x_2,\cdots,x_n\}$，参数 θ，

输出：

$$TC=\left\{\left(C_1^P,C_1^B\right),\left(C_2^P,C_2^B\right),\cdots\left(C_k^P,C_k^B\right)\right\}$$

Step1：通过式（6）、式（7）将 U 中所有样本都合并成一类，利用轮廓系数对聚类结果进行评估，得到较为合理的 K 值；

Step2：对 U 进行多次抽样，通过式（1）对每次抽取的样本进行 k-means 聚类，得到一组中心；

Step3：通过式（2）计算这 n 组中心的准则函数值，选择具有最小准则函数的对应聚类中心作为初始聚类中心 $C'=\{C'_1,C'_2,\cdots C'_k\}$。

Step4：利用改进的 k-means 算法得到初始中心，将其作为 k-means 算法的初始中心输入，进行聚类，得到聚类结果 $C'=\{C'_1,C'_2,\cdots C'_k\}$。

Step5：对于每一类 C_i，任取 $x_j\in C_i$，如果 $\theta(x_m)\subset C_i$，那么 $x_j\in C_i^P$，否则 $x_j\in C_i^B$；

Step6：对于每一类 C_i，任取 $x_j\notin C_i$，如果 $\theta(x_m)\cap C_i\neq\varnothing$，那么 $x_j\in C_i^B$；

Step7：通过 Step5 和 Step6 得到 C_i 的 C_i^P 和 $C_i^B(i=1,2,\cdots,k)$，返回

$$\left\{\left(C_1^P,C_1^B\right),\left(C_2^P,C_2^B\right),\cdots,\left(C_k^P,C_k^B\right)\right\}$$

4　几种聚类性能度量的指标

4.1　平均轮廓系数

轮廓系数（silhouette coefficient）[20]是聚类效果好坏的一种评价方式，它结合内聚度和分离度两种因素，在相同原始数据的基础上，评价不同算法或算法的不同运行方式对聚类结果所产生的影响。

计算某一个点的轮廓系数公式：

$$S(i)=\frac{b(i)-a(i)}{\max\left(a(i),b(i)\right)}$$

其中：$a(i)$ 表示 i 点向量到与它同簇的其他点的平均距离；$b(i)$ 表示 i 点向量到与它异簇的点的平均距离最小值。由上式可知，轮廓系数的值为 $[-1,1]$，如果轮廓系数 $S(i)$ 值越大，则表明 i 点所在的簇就越紧密。

对于整个数据集来说，其轮廓系数的计算公式如下：

$$SC=\frac{\sum_{i=1}^{n}S(i)}{n}$$

其中：n 表示数据集中的样本总数；SC 值越大，则聚类效果越好，反之越差。

4.2　Davies-Bouldin-Index 评价准则

Davies-Bouldin-Index（DBI）[21]，聚类结果的 DBI 越小，聚类效果越好。公式如下：

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i,j=1,2,\cdots,k}\left(\frac{avg\left(C_i\right)+avg\left(C_j\right)}{d_{cen}\left(\mu_i,\mu_j\right)}\right)$$

其中 k 是划分的类数，$d_{cen}\left(\mu_i,\mu_j\right)$ 为聚类中心 μ_i 和 μ_j 之间的距离，$avg\left(C_i\right)$ 为第 i 类中所有样本到聚类中心 μ_i 的距离的平均值。

4.3　准确率

准确率（accuracy）是常见的一种评价聚类性能的外部指标。准确率越高，聚类效果越好。计算公式为

$$ACC=\frac{1}{N}\sum_{i=1}^{k}\theta_i$$

其中 N 是所有已被确定类别的对象的总数，k 是聚类数，θ_i 是第 i 个类中正确划分的数据对象的个数。

5　实验结果与分析

本节选用 5 组标准 UCI[22]数据集对本文提出的算法进行测试实验来验证方法的性能。表 1 列出了实验中所使用的 5 组测试数据集的基本信息。对于每个数据集，重复进行了 200 次实验，用 200 次的平均值作为算法性能差异比较的依据。

表 1　实验所使用的数据集

数据集	样本个数	样本维数	类别数
Hill	1212	100	2
Iris	150	4	3
Sonar	208	60	2
Wdbc	569	30	2
Wine	178	13	3

本文设置两个实验来说明提出的改进的 k-means 算法以及基于改进 k-means 算法的三支决策的聚类效果和准确率。将改进的 k-means 算法记为 k-means PA。

第一个实验是二支聚类的测试,通过对比一些已有的算法,即首先用fuzzy c-means、k-medoid、传统k-means和k-means PA分别在表1中的数据集上进行聚类来验证本文提出的k-means PA算法的聚类结果DBI和准确率,以及聚类结果的稳定性。

第二个实验是在第一个实验的基础上,结合本文介绍的三支决策方法进一步优化聚类结果。最后通过实验数据分别比较二支聚类和三支聚类两组实验结果的DBI和ACC。

5.1 实验一

图2是fuzzy c-means、k-medoid、传统k-means和k-means PA四种算法表1中的数据集上进行二支聚类后结果的DBI对比图。由图可知,本文提出的算法k-means PA在UCI的5个数据集上,其DBI均低于fuzzy c-means、k-medoid和传统k-means的DBI,在Sonar数据集上尤其明显。在Wine数据集上,本文算法与传统k-means基本相同,低于fuzzy c-means和k-medoid。

图3是fuzzy c-means、k-medoid、传统k-means和k-means PA四种算法表1中的数据集上进行二支聚类后结果的准确率对比图。由图可知,本文提出的算法k-means PA在Iris数据集上的准确率没有fuzzy c-means、k-medoid的ACC高,但比传统k-means的准确率要高;在Hill、Sonar和Wine数据集上,k-means PA的准确率要高于其他三种算法;在Wdbc数据集上,k-means PA的准确率和其他三种算法的准确率不分伯仲。

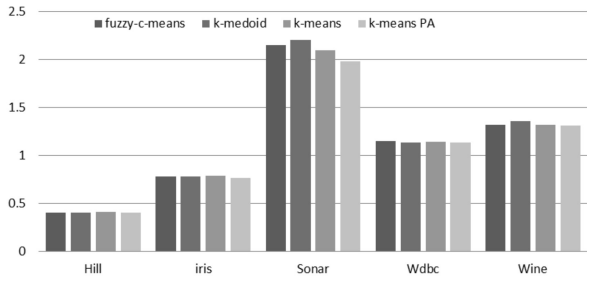

图2 算法DBI实验结果

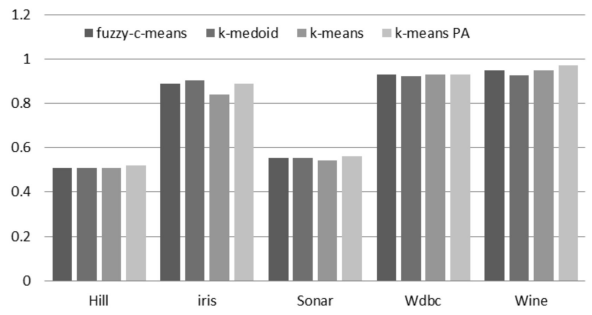

图3 算法准确率实验结果

综合图2和图3来,本文提出的改进算法在UCI的5个数据集上获得了较好的DBI和准确率,因此,本文的算法能适应于不同数据集的聚类挖掘。

5.2 实验二

实验二是在实验一的基础上进行的三支聚类,图4和图5是基于fuzzy c-means、k-medoid、传统k-means和改进k-means算法结合三支聚类理论的聚类结果的DBI和ACC对比图。

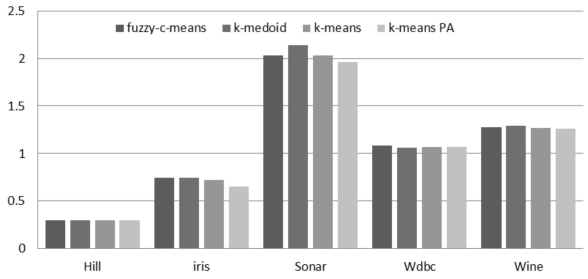

图4 三支聚类结果DBI对比图

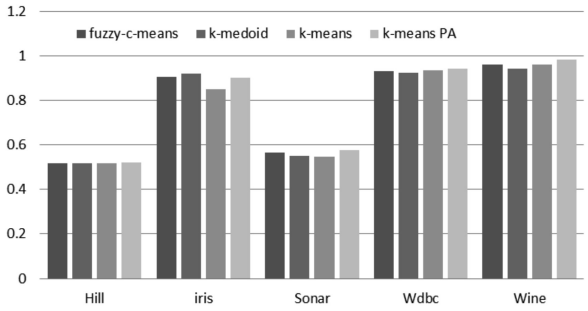

图5 三支聚类结果ACC对比图

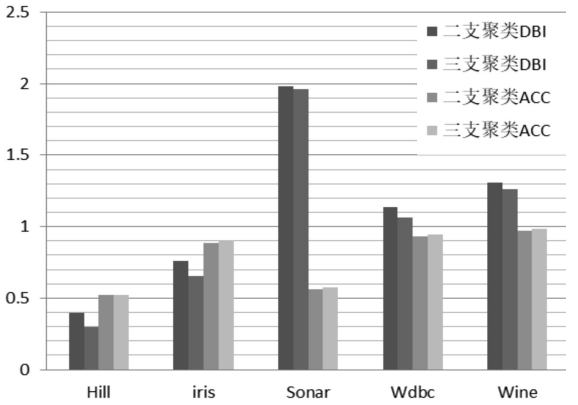

图6 基于k-means PA的二支聚类和三支聚类结果的DBI和ACC对比图

由图可知,三支聚类结果的DBI和ACC对比结论同实验一。但从图6中可以明显看出基于本文提出的k-means PA算法的三支聚类结果的DBI比k-means PA二支聚类结果的DBI低,基于本文提出的k-means PA算法的三支聚类结果的ACC比k-means PA二支聚类结果的ACC高,这表明结合三支聚类方法的聚类效果更好,准确率更高。综上所述,本文提出的改进的k-means算法在三支聚类

算法中应用是有效的。

5.3　稳定性

由于传统 k-means 算法在聚类时结果存在不稳定现象，因此，对提出改进行算法的聚类结果进行了稳定性实验。选取其中一个数据集并进行 30 次运行实验，其结果如图 7 所示。从图 7 可以看出，传统的 k-means 算法稳定性较差，结果会随着运行次数不同而呈现不同的聚类结果，而本文提出的聚类算法的聚类结果呈现出较好的稳定性。

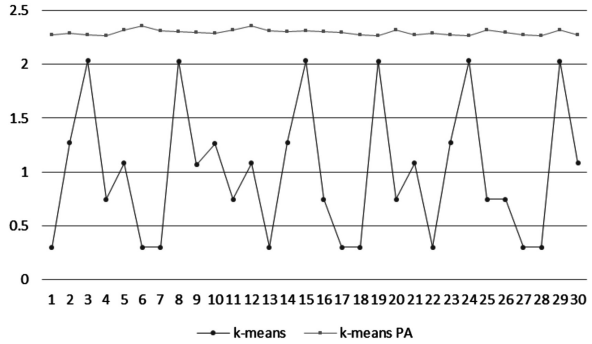

图 7　算法稳定性比较

6　结语

本文针对传统 k-means 聚类算法初始中心的选取做了改进，并结合 Yu 等学者研究的基于邻域的三支聚类理论，提出一种改进的 k-means 算法并结合三支聚类的算法，解决了初始中心和 k 值的选取问题和处理不确定性信息时最优类别划分的问题。实验结果证明，本文所提出的算法可以有效地避免传统 k-means 因随机选取初始簇而导致了聚类不稳定的现象，并且算法在准确率上有所提高，DBI 表明聚类的效果更好。但计算的时间有所增加，如何快速有效地获取一个最优参数 k 来使聚类效果和精度能达到一个最佳理想的结果，还有研究近邻参数 θ 的值将会是下一步的研究工作内容。

参 考 文 献

[1] RESNICK P, IACOVOU N, SUCHAK M, et al. GroupLens: anopen architecture for collaborative filtering of netnews [C]//CSCW'94: Proceedings of the 1994 ACM Conference on ComputerSupported Cooperative Work. New York: ACM, 1994: 175-186.

[2] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions [J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(6): 734-749.

[3] JAIN A K. Data clustering: 50 years beyond k-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[4] 冯波，郝文宁，陈刚，等. K-means 算法初始聚类中心选择的优化[J]. 计算机工程与应用，2013，49(14)：182-185，192.
FENG Bo, HAO Wenning, CHEN Gang, et al. Optimization of initial clustering center selection for K-means algorithm [J]. Computer Engineering and Applications, 2013, 49(14): 182-185, 192.

[5] 郁启麟. K-means 算法初始聚类中心选择的优化[J]. 计算机系统应用，2017，26(05)：170-174.
YU Qilin. Optimization of initial clustering center selection for K-means algorithm [J]. Computer Systems& Applications, 2017, 26(05): 170-174.

[6] Debatty T, Michiardi P, Mees W, et al. Determining the k in Kmeans with Mapreduce [C]//Proc of EDBT/ICDT Workshops, 2014: 19-28.

[7] Yuan Qilong, Shi Haibo, Zhou Xiaofeng. An optimized initialization center K-means clustering algorithm based on density[C] //Proc of IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems. Piscataway, NJ: IEEE Press, 2015: 790-794.

[8] Kettani O, Ramdani F, Tadili B. AK-means: an automatic clustering algorithm based on K-means[J]. Journal of Advanced Computer Science&Technology, 2015, 4 (2): 231-236.

[9] Hoppner F, Klawonn F, Kruse, R, et al. Fuzzy cluster analysis: methods for classification, data analysis and image recognition[M]. Chichester: Wiley Press, 1999: 1-48.

[10] Lingras P. Rough K-Medoids clustering using Gas [C]// Proceedings of the 8th IEEE International Conference on Cognitive Informatics. IEEE Press, 2009: 315-319.

[11] Lingras P, Hogo M, Snorek M. Interval set clustering of web users using modified Kohonen self-organizing maps based on the properties of rough sets [J]. Web Intelligence and Agent Systems: An International Journal, 2004, 2(3): 217-225.

[12] Lingras P, Hogo M, Snorek M, et al. Temporal analysis of clusters of supermarket customers: conventional versus interval set approach [J]. Information Sciences, 2005, 172(1): 215-240.

[13] Lingras P, West C. Interval set clustering of web users with rough k-means [J]. Journal of Intelligent Information Systems, 2004, 23(1): 5-16.

[14] Yao Yiyu, Lingras P, Wang Ruizhi, et al. Interval set cluster analysis: a re-formulation[C]//LNCS 5908: Proceedings of the 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Delhi, India, Dec 15-18, 2009. Berlin, Heidelberg:

(下转第 1353 页)

destrian detector to a specific traffic scene [C]//CVPR 2011. IEEE, 2011: 3401-3408.
[4] Ge W, Collins R T. Marked point processes for crowd counting[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 2913-2920.
[5] Idrees H, Soomro K, Shah M. Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37 (10): 1986-1998.
[6] Lin Z, Davis L S. Shape-based human detection and segmentation via hierarchical part-template matching [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(4): 604-618.
[7] Chan A B, Liang Z S J, Vasconcelos N. Privacy preserving crowd monitoring: Counting people without people models or tracking[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-7.
[8] Chan A B, Vasconcelos N. Bayesian poisson regression for crowd counting[C]//2009 IEEE 12th international conference on computer vision. IEEE, 2009: 545-551.
[9] Chen K, Loy C C, Gong S, et al. Feature mining for localised crowd counting[C]//BMVC,2012, 1(2): 3.
[10] Lempitsky V, Zisserman A. Learning to count objects in images [C]//Advances in neural information processing systems,2010: 1324-1332.
[11] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions [C]//Proceedings of the IEEE conference on computer vision and pattern recognition,2015: 1-9.
[12] Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.
[13] Wang C, Zhang H, Yang L, et al. Deep people counting in extremely dense crowds[C]//Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015: 1299-1302.
[14] Zhang Y, Zhou D, Chen S, et al. Single-image crowd counting via multi-column convolutional neural network [C]//Proceedings of the IEEE conference on computer vision and pattern recognition,2016: 589-597.
[15] Boominathan L, Kruthiventi S S S, Babu R V. Crowdnet: A deep convolutional network for dense crowd counting[C]//Proceedings of the 24th ACM international conference on Multimedia. ACM, 2016: 640-644.
[16] Pham V Q, Kozakaya T, Yamaguchi O, et al. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation [C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 3253-3261.
[17] Li Y, Chen Y, Wang N, et al. Scale-aware trident networks for object detection [J]. arXiv preprint arXiv: 1901.01892, 2019.
[18] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines [C]//Proceedings of the 27th international conference on machine learning (ICML-10),2010: 807-814.
[19] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection [C]//Proceedings of the IEEE international conference on computer vision,2017: 2980-2988.
[20] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch [C] //Proceedings of the 31st conference on neural information processing systems (NIPS 2017),2017:1-4.

(上接第 1299 页)

Springer, 2009: 398-405.
[15] Yu Hong, Chu Shuangshuang, Yang Dachun. Autonomous knowledge-oriented clustering using decision-theoretic rough set theory [J]. FundamentaInformaticae, 2012, 115(2/3): 141-156.
[16] Yu Hong, Liu Zhanguo, Wang Guoyin. An automatic method to determine the number of clusters using decision-theoretic rough set[J]. International Journal of Approximate Reasoning,2014, 55(1): 101-115.
[17] Yu Hong, Zhang Cong, Wang Guoyin. A tree- based incremental overlapping clustering method using the three-way decision theory[J]. Knowledge- Based Systems, 2016, 91(C): 189-203.
[18] Li Jinhai, Huang Chenchen, Qi Jianjun, et al. Three-way cognitive concept learning via multi-granularity[J]. Information Sciences, 2017, 378: 244-263.
[19] 黄韬,刘胜辉,谭艳娜. 基于k-means聚类算法的研究[J]. 计算机技术与发展,2011,21(07):54-57,62.
HUANG Wei, LIU Shenghui, TAN Yanna. Research Based on k-means Clustering Algorithm[J]. Computer Technology and Development,2011,21(07):54-57,62.
[20] Davies D L, Bouldin D W. A cluster separation measure [J]. IEEE transactions on pattern analysis and machine intelligence, 1979 (2): 224-227.
[21] Rousseeuw P J. Silhouettes: a graphical aid to the inter-pretation and validation of cluster analysis[J]. Journal of computational and applied mathematics, 1987, 20: 53-65.
[22] UCI machine Learning Repository [EB/OL]. 2005, Available: http://www.ics.uci.edu/mlearn /MLRepository.html.