Advances in Applied Mathematics 应用数学进展, 2018, 7(10), 1349-1356
Published Online October 2018 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2018.710157

Hans汉斯

Three-Way Clustering Algorithms Based on Disturbances and K-Means

Dan Shen, Xiaolei Wang, Pingxin Wang*

School of Science, Jiangsu University of Science and Technology, Zhenjiang Jiangsu
Email: *pingxin_wang@hotmail.com

Received: Oct. 9th, 2018; accepted: Oct. 24th, 2018; published: Oct. 31st, 2018

Abstract

K-means algorithm is a traditional algorithm used for partition clustering, and its essence is a hard clustering, that is, the object studied only has two possible results, either belonging to this class or not belonging to this class, and its segmentation results are highly accurate. However, this algorithm has obvious disadvantages, and it is unable to deal with objects with features that are not obvious. The three-way clustering is a kind of fuzzy clustering division, which can deal with the non-obvious objects through the definition of core domain and boundary domain. This paper combines the ideas of three-way decision theory and K-means algorithm to form a new clustering algorithm. It cannot only maintain the original accuracy when clustering, but also make a more reasonable classification of relatively uncertain points. Then the core domain and boundary domain of the cluster are separated by perturbation processing.

Keywords

K-Means, Perturbation, Discrete Degree

基于K-means的三支聚类算法

沈　丹，王晓磊，王平心*

江苏科技大学理学院，江苏　镇江
Email: *pingxin_wang@hotmail.com

收稿日期：2018年10月9日；录用日期：2018年10月24日；发布日期：2018年10月31日

摘　要

K-means算法是一种传统的基于划分聚类的算法，其本质是一种硬聚类划分，即要求每个研究对象要么

*通讯作者。

文章引用: 沈丹, 王晓磊, 王平心. 基于 K-means 的三支聚类算法[J]. 应用数学进展, 2018, 7(10): 1349-1356.
DOI: 10.12677/aam.2018.710157

属于这个类，要么不属于这个类，其聚类结果具有严格的边界。然而将某些不确定的对象强制分配到某个类中往往容易带来较高的决策风险。三支聚类将确定的元素放入核心域中，将不确定的元素放入边界域中延迟决策，可以有效地降低决策风险。本文将三支决策理论和K-means算法相结合得到一个新的三支聚类算法，该算法利用K-means聚类的结果，对不确定的点做更加合理的分类，同时对聚类完成的结果做扰动处理，分离出聚类内部的核心域和边界域。

关键词

K-means，扰动，离散度

1. 引言

聚类[1]是用于研究多性质对象划分的方法，一直是机器学习等领域极具挑战的课题，长期以来受到广大研究者的关注。聚类算法是根据所研究对象的各个属性，通过属性的特征，按照数学的计算方法分析对象的异同，进而将对象进行划分。当数据无任何类别标记时，算法需要从初始数据中提取符合某种条件的规律，其自动划分的特性大大简化了人工成本，对于安全保障[2]，图像分析[3]，搜索引擎[4]以及相关的发展起到了巨大的促进和推动作用。

K-means 算法[5]是一个传统的用于聚类划分的算法，其划分迭代的准确性在当今各个领域得到广泛认可和应用。然而其本质是一种硬聚类的划分算法，即任意对象只能属于某一特定类簇。在点相对离散的情况下，这种基于距离的划分并不是特别规范。在实际问题中若强制区分性质不明显的对象，往往会带来较高的风险。面对这样的风险问题，本文提出了一种融合三支决策思想和 K-means 算法的三支聚类算法，即在用 K-means 算法做划分时考虑三支划分，通过核心域和边界域的界定软化硬聚类划分的弊端，在对于某个对象的研究当中，当这个对象在划分的时候由于性质不明显，可以同时划分进入两个类簇，这样的区域叫做边界域，而特征明显的点划分进入核心域。同样，在每个聚类内部的点也需要一个能够缓冲的区域，此区域能够软化聚类内部特征明显与不明显的点。此算法的可行性在于点的离散性和点集间隔性，在检测指标的选取上充分利用了这两种特性，用变异系数 CV [6]衡量离散型，采用类间 Separation (间隔性) [7]衡量点集的间隔程度。

2. 相关工作

2.1. K-means 算法

K-means 算法是经典的聚类划分算法，对于待研究的对象集合 $U=\{x_1,x_2,\cdots,x_n\}$，事先确定需要划分的聚类个数 k，任选 U 中的 k 个点作为初始聚类中心，计算 U 中所有对象到每个聚类中心的欧氏距离，将距离最小的点划分进此中心的归属，在此基础上得到一个初始聚类划分并重新计算聚类的中心，之后进行逐次迭代，以上阶段的聚类中心结果作为下阶段的初始中心作进一步划分，当聚类好的点趋于稳定时结束该过程，此过程的精度随着迭代次数的增加不断趋于理想状态。然而，K-means 的本质属于传统硬聚类范畴，即对于任意一个对象只能属于某一个聚类。

以二维图形图 1，此图通过 K-means 算法对特征明显的点直接做划分，但是在处理特征不明显的点

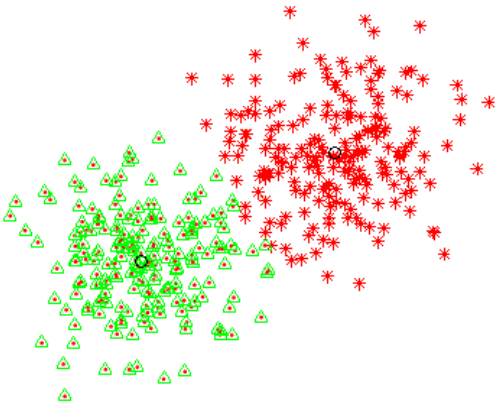

Figure 1. A schematic diagram by K-means
图 1. K-means 聚类结果示意图

的时候，却会出现比较明显的弊端，比如直观上观察到类的边界处某两个点的距离很近，但是却被分在不同的类中，显然这种划分并不完善，此划分过程虽然有严格的划分指标，但是却增加了划分的风险系数。

2.2. 三支聚类

设给定一个数据 $U=\{x_1,x_2,\cdots,x_n\}$，传统的二支聚类方法是用一个集合表示一个类，即寻找一组集合 $C_1,C_2,\cdots,C_k$ 满足 $U=\bigcup_{i=1}^{k}C_i$ 且 $C_i\cap C_j=\varnothing$，$(i,j=1,2,\cdots,k,i\neq j)$，其中 k 为聚类的个数。

基于三支决策的聚类思想是用三个互不相交的集合表示一个类，即 C_i^P,C_i^B 与 C_i^N，分别称为类的核心域，边界域和琐碎域，其中

$$C_i^P\cup C_i^B\cup C_i^N=U \tag{1}$$

核心域中的元素确定属于这个类，边界域中的元素可能属于也可能不属于这个类，而琐碎域中的元素则肯定不属于这个类。由关系式(1)，我们可以通过 C_i^P 和 C_i^B 来表示一个类。反过来，如果给定一组集合 C_i^P,C_i^B，$(i=1,2,\cdots,k)$满足

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

$$\bigcup_{i=1}^{k}\left(C_i^p\cup C_i^B\right)=U \tag{3}$$

我们称其为数据集 U 三支聚类。其中(2)要求正域非空，即每个类中至少有一个对象，而(3)保证每个对象至少被分到一个类中。和传统的硬聚类的结果 $C=\{C_1,C_2,\cdots,C_k\}$ 不同，三支聚类的结果应为如下形式，

$$TC=\left\{\left(C_1^P,C_1^B\right),\cdots,\left(C_k^P,C_k^B\right)\right\} \tag{4}$$

显然在三支聚类中如果有 $C_i^B=\varnothing,\left(i=1,2,\cdots,k\right)$，就变成了传统二支决策的聚类形式。因此，三支决策聚类形式是传统二支聚类方法的推广，这也是对一些不确定数据聚类问题提出的一种解决方案，针对那些目前知识体系下难以聚类的对象，我们无法确定其所属类别时将其归为某些类的边界域，等待新的信息以帮助进一步决策。

3. 问题分析

K-means 算法在划分点的过程中采用逐步迭代的方法，每一次迭代过程可以优化一次点的划分，当

迭代到最后阶段时，每个元素都有自己的归属，此过程的精确度逐步提高，因此在迭代过程中做优化必定会降低原有的精确性，所以做进一步优化的过程首选迭代结束之后，在避免精确性降低的情况下做优化处理。

上一阶段的聚类处理仅仅解决了聚类边界的划分，对于每个聚类内部，如果缺少过渡的集合，势必也会增加强制划分的风险系数，因此对于内部对象同样也需要进行划分，即特征明显的点为核心域，不明显的点为边界域，在做此划分之后，提取出的边界域和优化后的 K-means 算法找出的缓冲区域共同构成此聚类的边界域。

对于类中的每个点都有其属性，但是每个点的属性差异不一，对于差异很大的点可以轻易地做出划分，但是对于差异较小的点，可以通过放大属性的思想做划分。划分的方法多种多样，本文在划分核心域的时候，通过每个点对中心的干扰程度做出判断，干扰程度越大，属性差异越大，即可划分到边界域当中。

4. 基于 K-means 的扰动三支聚类算法

4.1. 步骤一：改进 K-means

算法：

Step 1：对于整个数据集 U，通过聚类的个数 k 任选 k 个点作为聚类的中心 $c_1,c_2,\cdots,c_k$；

Step 2：比较聚类中的每个点到 k 个点的距离，通过距离的大小做一次划分，即距离越近的点划分到中心所属聚类当中；

Step 3：重新计算每个聚类的中心；

Step 4：若每个新聚类中心和上一次的聚类中心的距离在一个很小的范围内，转 Step 5，否则转 Step 2；

Step 5：比较聚类中的每个点到 k 个点的距离，若最小的点和次小的点的距离相差很小，则把它划分到边界域当中。

Step 6：输出边界域 $U^{12},U^{21},U^{13},U^{31},\cdots,U^{(k-1)k},U^{k(k-1)}$ 与聚类 $U_1,U_2,\cdots,U_k$。

由于此方法是划分类间的算法，在对原先 K-means 的缺陷作了优化，仍然是采用 K-means 的思想，通过距离的大小来判断点的归属，在最后一步弱化了 K-means 的要求，产生了软化硬聚类的效果，但是 K-means 的精确性却没有发生任何改变。此步骤作为三支聚类的先导步骤，考虑的是类间的划分效果，在此结果之上需做进一步划分。

4.2. 步骤二：单个点集做三支聚类

算法：

Step 1：将步骤一中的各个聚类(不包含边界域) $U_1,U_2,\cdots,U_k$ 带入步骤二；

Step 2：对单个聚类中的每个点 $p_{ji}\left(j=1,2,\cdots,k\right)$ 增加所占比重，增加 n 个相同的点构成一个新的数据集 U_*。

Step 3：重新计算 U_* 的中心 c_{j*}，比较 c_{j*} 与原先中心 c_j 的偏移程度，若偏移程度大于参数，则将它划分到边界域当中，若偏移程度较小，则划分到核心域当中。

Step 4：是否所有聚类都判断完毕，若判断结束，则转 Step 5，否则转 Step 2。

Step 5：输出核心域 $U_1^C,U_2^C,\cdots,U_k^C$ 和边界域 $U_1^M,U_2^M,\cdots,U_k^M$。

此步骤是通过放大点的性质来区别点的划分，考虑的是聚类内部的准确划分。由于步骤一和步骤二是相互独立的，因此需要将两个步骤的数据进行合理合并，合并的规则基于边界域的相似性，即对于性

质相似的边界域进行合并。

4.3. 步骤三：合并边界域

算法：

Step 1：将步骤一的边界域$U^{12},U^{21},U^{13},U^{31},\cdots,U^{(k-1)k},U^{k(k-1)}$、步骤二的边界域$U_1^M,U_2^M,\cdots,U_k^M$和步骤二的核心域$U_1^C,U_2^C,\cdots,U_k^C$带入到步骤三。

Step 2：对于$U_j^M\ (j=1,2,\cdots,k)$和$U^{il}\ (i=1,2,\cdots,k;l=1,2,\cdots,k)$且$i\neq l$，若$j=i$或$j=l$，则将$U_j^M$和$U^{il}$合并为$U_{j*}^M$。

Step 3：若所有边界域都判断完毕，转 Step 4，否则转 Step 2。

Step 4：输出核心域$U_1^C,U_2^C,\cdots,U_k^C$和边界域$U_{1*}^M,U_{2*}^M,\cdots,U_{k*}^M$。

5. 实验结果

5.1. 实验指标

由于对二维数据的处理可以通过图形的形式展示效果，但是无法通过图形展示三维甚至高维的数据，因此需要合理的评价指标对高维数据进行分析。本文采用变异系数 CV [5]比较类内的离散程度，用公式表示为：

$$\mathrm{CV}=\delta/\mu$$

其中$u=\dfrac{\sum_{i=1}^{n}p_i}{n}$，$\delta=\sqrt{\dfrac{\sum_{i=1}^{n}(p_i-u)^2}{n}}$。

CV 是标准差和平均数的比值，所以此指标无量纲，可以消除单位和(或)平均数不同对两个或多个聚类变异程度的影响，一般来说，指标的数值越小，点的分布越集中。本文需比较三个 CV 值，分别为 Kmeans 划分后得到的聚类 CV 值，优化后 K-means 算法划分得到的聚类 CV 值和核心域 CV 值，分别记为 CV1，CV2，CV3，根据划分的过程，从理论上有 CV1 > CV2 > CV3，由于聚类的个数多于一个，因此在求 CV 值时，取所有聚类 CV 值的平均值。

CV 适用于类内点的分析，却没有考虑类间的关系，仅仅一个衡量指标并不能完全阐述点的分布状态，因此选用 Separation (间隔性) sp [6]来衡量类间的划分，将原先的 K-means 算法划分出的聚类和用优化后 Kmeans 算法去除边界域后的聚类做比较。间隔性越大，聚类效果越明显。

$$\mathrm{sp}=\frac{2}{k^2-k}\sum_{i=1}^{k}\sum_{j=i+1}^{k}\left\|w_i-w_j\right\|_2$$

其中$\left\|w_i-w_j\right\|_2$表示两个聚类中心的距离，sp 越高表明类间聚类距离越远。本文需要比较两个 sp 值，分别为传统 K-means 算法得到的聚类 sp 值和优化后 K-means 算法得到的 sp 值，分别记为 sp1，sp2，由于优化后的结果优于传统算法，因此从理论上有 sp1 < sp2 成立。此指标没有考虑类内数据，因此和 CV 指标共同衡量划分的优劣性。

5.2. 人工数据

本小节选取一组人工数据集 U，选取规则为协方差 δ，均值为 μ 的正态分布中抽取 $n*m$ 的矩阵 Array，其中 n 为样本总数，m 为矩阵维度，为直观描述点的发布，此处选用矩阵维度为 2 的数据集。其中$\mu=\begin{pmatrix}\mu_1 & \mu_2\\ \mu_3 & \mu_4\end{pmatrix}$，Array 中每一行是以 μ 对应行作为正态分布选取的一个样本。其中

$$\mu_j = \left(rand\left(5,2\right)-1\right)*11+1,\ j=1,2,3,4$$

$\delta = \begin{pmatrix} \delta_1 & \delta_2 \\ \delta_3 & \delta_4 \end{pmatrix}$，其中 $\delta_j = \left(rand\left(5,2\right)*1\right)/5+2,\ j=1,2,3,4.$

本次实验中选取两组点，每组 200 (400)个，初始中心为 400 (800)个点任取两个，比较最后的图形结果以及评价指标结果。图 2 为 400 个点构成的集合，图 3 为 800 个点构成的集合，在优化后 K-means 基础上的划分和指标结果如表 1，表 2 所示。

图 2 实验结果符合 K-means 软化的预期效果，点的个数越多，聚类的效果越明显，其实验指标也符合预期设想，然而这只是对于二维的数据分析，对高维数据同样需要做具体分析。

5.3. UCI 数据集

本文选择五组 UCI 数据集 Seed-Nor (SN)、Breast Tissue (BT)、Congressional Voting (CV)、Ecoli-Nor (EN)、Contraceptive Method (CM)，最后的检测指标数据如表 3~表 5 所示。

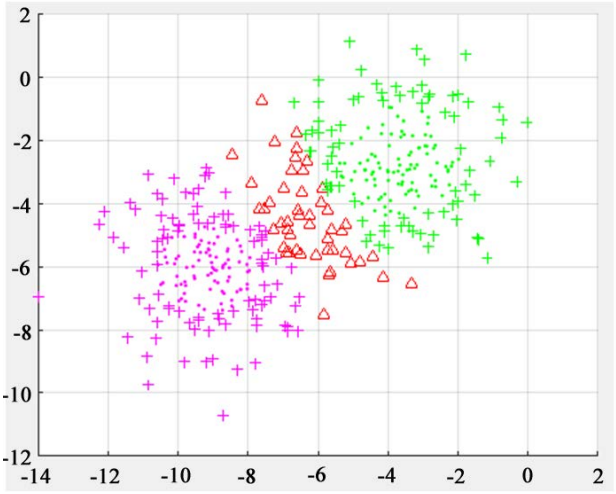

Figure 2. Clustering result of artificial data sets containing 400 elements
图 2. 400 个点的人工数据集聚类结果

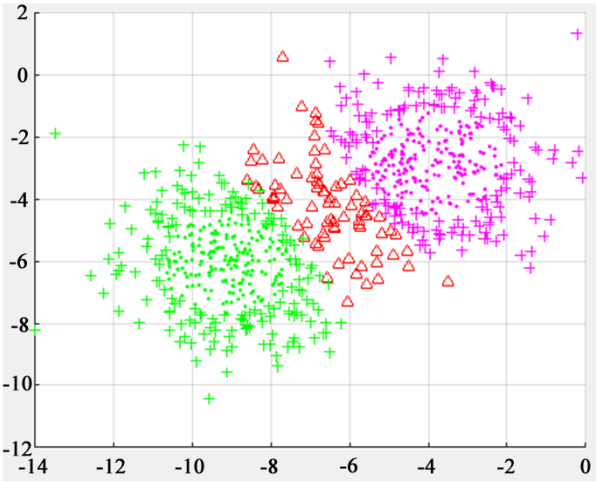

Figure 3. Clustering result of artificial data sets containing 800 elements
图 3. 800 个点的人工数据集聚类结果

Table 1. Experimental result of artificial data sets containing 400 elements
表 1. 400 个点人工数据集实验结果

CV 类别	CV 值	sp 类别	sp 值
CV1	7.4043	sp1	5.8240
CV2	7.3268	sp2	6.4444
CV3	4.5234		

Table 2. Experimental result of artificial data sets containing 800 elements
表 2. 800 个点人工数据集实验结果

CV 类别	CV 值	sp 类别	sp 值
CV1	10.0617	sp1	5.7539
CV2	9.8488	sp2	6.2464
CV3	6.1931		

Table 3. Data sets of UCI
表 3. UCI 数据集

数据集	样本数	样本维数	类别
SN	210	7	3
BT	106	9	6
CV	435	16	2
EN	336	7	8
CM	1473	9	3

Table 4. CV values of UCI data sets
表 4. UCI 数据集实验结果的 CV 值

数据集	CV1	CV2	CV3
SN	3.2822	2.8721	1.1776
BT	1.7118	1.4381	0.2075
CV	5.0783	4.0719	3.0711
EN	2.5529	1.9499	0.3143
CM	7.3864	6.5790	5.7671

Table 5. Sp values of UCI data sets
表 5. UCI 数据集实验结果的 sp 值

数据集	sp1	sp2
SN	0.9363	1.0149
BT	0.8970	0.9508
CV	1.2341	1.3499
EN	0.6976	0.7621
CM	0.9914	1.0894

6. 总结

本文的思想是基于传统算法的基础之上，通过合理放大点的性质对点进行划分。对于本文所选实验指标，都充分达到预期效果，但是明显可以看出实验指标相差并不大，特别是在对间隔性的研究中，sp

值的变化很小，接下来所要做的就是寻找放大实验指标的方法对本文实验指标进行优化处理。同时，不同数据集的数量级各不相同，本文是基于距离的算法，因此，在对参数的选取上需要根据数量级的大小做不同变化。

基金项目

本文受国家自然科学基金(61503160, 61572242)，江苏省高校自然科学基金(15KJB110004)，江苏科技大学本科生创新计划项目资助。

参考文献

[1] Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) Data Clustering: A Review. *ACM Computing Survey*, **31**, 264-323. https://doi.org/10.1145/331499.331504

[2] Kalyani, S. and Swarup, K.S. (2011) Particle Swarm Optimization Based K-Means Clustering Approach for Security Assessment in Power Systems. *Expert Systems with Applications*, **38**, 10839-10846. https://doi.org/10.1016/j.eswa.2011.02.086

[3] El Alami, M.E. (2011) Supporting Image Retrieval Framework with Rule Base System. *Knowledge-Based Systems*, **24**, 331-340. https://doi.org/10.1016/j.knosys.2010.10.005

[4] Martín-Guerrero, J.D., Palomares, A., Balaguer-Ballester, E., *et al*. (2006) Studying the Feasibility of a Recommender in a Citizen Webportal Based on User Modeling and Clustering Algorithms. *Expert Systems with Applications*, **30**, 299-312. https://doi.org/10.1016/j.eswa.2005.07.025

[5] Hartigan, J.A. and Wong, M.A. (1979) A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*: *Series C* (*Applied Statistics*), **28**, 100-108.

[6] 王文森. 变异系数——一个衡量离散程度简单而有用的统计指标[J]. 中国统计, 2007, 2007(6): 41-42.

[7] Adil, F., Najlaa, A., Zahir, T., Abdullah, A., *et al*. (2014) A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions on Emerging Topics in Computing*, **2**, 267-279. https://doi.org/10.1109/TETC.2014.2330519