DOI：10.7652/xjtuxb201410018

采用万有引力定律自动确定类数的 K 均值算法

杜辉[1,2]，王宇平[1]，董晓盼[1]
（1. 西安电子科技大学计算机学院，710071，西安；2. 西北师范大学计算机科学与工程学院，730070，兰州）

摘要：针对传统 K 均值算法需要提前指定聚类数目且易陷入局部最优的问题，提出了一种采用万有引力定律自动确定类数的 K 均值算法（Gravity K 均值算法，GK 均值算法）。所提算法利用正交设计方法在数据空间均匀投放若干探测器，探测器根据万有引力定律移动，当两个探测器的距离小于给定阈值时合并为一个，当探测器处于稳定状态时，探测器的个数就是聚类的数目。将得到的探测器作为 K 均值算法的初始中心点，有效地避免了 K 均值算法陷入局部最优。实验结果表明：相比传统 K 均值算法，本文提出的方法可以自动确定聚类数目，并给出较好的初始中心，算法的迭代次数至少减少了 25%，聚类正确率平均提高了 14%，DB(Davies and Bouldin)聚类评价指标平均降低了 0.19。

关键词：万有引力；聚类；K 均值；探测器
中图分类号：TP301.6 **文献标志码**：A **文章编号**：0253-987X(2014)10-0115-05

An Improved *K*-Means Algorithm with Auto-Determined Clustering Number by Using Gravity

DU Hui[1,2], WANG Yuping[1], DONG Xiaopan[1]
(1. School of Computer Science and Technology, Xidian University, Xi'an 710071, China;
2. College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China)

Abstract: An improved K-means algorithm to auto-determine the clustering number by using gravity is proposed to solve the problems that one needs to specify the number of clusters in the traditional K-means algorithm and it is easy to fall into local optimum. The orthogonal design method is used to uniformly place several detectors into the data space, and then these detectors are moved according to the law of universal gravitation. When two detectors's distance less than the given threshold, they are merged. The number of detectors is just the number of the clusters when the algorithm converges. Then the resulting detectors are used as the initial center points to avoid the local optimum. Experimental results and comparisons with the traditional K-means algorithm show that the proposed method can automatically determine the number of clusters and gives the better initial centers, and that the iteration number of the algorithm is reduced by at least 25% and clustering accuracy is increased by an average of 14% and the performance evaluation DB(Davies and Bouldin) index is reduced by an average of 0.19.

Keywords: gravity; cluster; K-means; detector

聚类分析是用于探索给定数据集基本结构的一种工具，它被应用于各种各样的工程和科学领域，例

收稿日期：2014-03-30。 作者简介：杜辉(1976—)，女，博士生；王宇平(通信作者)，男，教授，博士生导师。 基金项目：国家自然科学基金资助项目(61272119)。
网络出版时间：2014-09-23 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20140923.1746.003.html

如医学、社会学、心理学、生物学、模式识别和图像处理等[1]。聚类问题可以描述为在数据集中发现类，同一个类中的数据相似度高，而类和类之间的数据相似度低。目前，解决聚类问题的方法有很多，K 均值算法[2]是其中最为经典的方法，学者们对该算法做了很多的研究与改进。Estivill 等增加了 K 均值算法对噪声和离群点的处理能力，使得算法更健壮[3]。苏木春等将算法中相似度的度量改用了“点对称距离”，使得算法处理非凸数据时也可以得到较好的聚类结果[4]。Likas 等采用增量的方法构造了一种全局的快速 K 均值算法[5]。D'Urso 等提出了一种健壮的模糊 K 均值模型用于处理区间值数据[6]。Hua 等在算法中考虑到每个数据点的三角关系以及其最近的聚类中心，使得算法的效率得到提高[7]。Timmerman 等在缩减的空间建立中心和类残差的模型，使得 K 均值算法可以处理更广泛的数据类型[8]，一些科研工作者在这些工作的基础上，研究了 K 均值算法自动确定类数及初始中心点选取的问题。Pelleg 等提出了 X 均值算法[9]，该算法试探多个聚类数，并为每种聚类数的聚类结果使用贝叶斯信息标准打分，然后选择得到最高评分的聚类数。Hamerly 等提出的 G 均值算法[10]假设在同一个类中的数据是服从高斯分布的。该算法首先给定一个较小的聚类数，然后利用统计检验的方法来判断被分配在同一类中的数据是否服从高斯分布，如果不服从将该类分裂，以此类推，最终确定出一个较理想的聚类数。Argyris 等提出了 Dip 均值算法[11]，该算法与文献[10]不同的是假设每一个类中的数据服从一种单峰分布。Bagirov 等引入一个辅助聚类功能用于生成一组初始点，这些点在数据集的不同部分[12]。Grigorios 等为了解决 K 均值算法初始化的问题，设计了 MinMax K 均值算法[13]，该算法根据每个类的协方差给中心点赋权值，然后对算法的目标进行优化。

本文针对 K 均值算法自动确定类数及初始中心点选取的问题，提出了一种采用万有引力定律自动确定类数的 K 均值算法——GK 均值(Gravity K 均值)算法，该算法可以自动确定聚类数，并为 K 均值算法提供了较理想的初始中心。

1 相关知识

1.1 万有引力定律

万有引力定律把地球上物体运动的规律和天体运动的规律统一了起来，自然界中任何两个物体都是相互吸引的，引力的大小与它们的质量的乘积成正比，而与它们距离的平方成反比

$$\boldsymbol{F}(t) = \boldsymbol{G}m_x m_y / \boldsymbol{d}^2(\boldsymbol{x}(t), \boldsymbol{y}(t)) \tag{1}$$

式中：t 为时间；$\boldsymbol{G}$ 为重力加速度；物体 y 在物体 x 引力的作用下会沿着向量 $\boldsymbol{d}(t)=\boldsymbol{x}(t)-\boldsymbol{y}(t)$ 的方向运动，则万有引力公式又可表示为

$$\boldsymbol{F}(t) = \boldsymbol{G}m_x m_y / |\boldsymbol{d}(t)|^2 \tag{2}$$

1.2 正交设计

文献[14]提出一种正交设计的方法，该算法用于构造正交矩阵 $\boldsymbol{L}_M(Q^N)$，M 代表采样的个数，N 是数据的维数，Q 称为层次，就是每一维数据可取值的个数，一般取奇数，满足 $M=Q^J$ 的条件，J 是满足式(3)的最小正整数。

$$N \leqslant Q^J - 1/Q - 1 \tag{3}$$

例如，$N=2$，$Q=3$，$J=2$，$M=9$，所构造的正交矩阵如图 1a 所示，图 1b 为正交矩阵中 9 个点均匀分布的情况，算法的具体流程参见文献[14]。利用该方法可以在任意数据空间构造均匀分布的点。

2 GK 均值算法

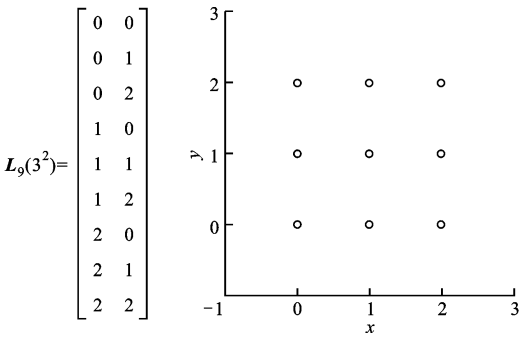
2.1 GK 均值算法思想

宇宙中的天体运行状况是小天体往往都围绕着质量较大的天体稳定运行，求聚类算法的中心点与此类似。具体地说，利用正交设计方法在数据空间均匀投放一些探测器，它们依照万有引力的规律运动，也就是质量小的探测器会被质量大的所吸引，并朝着质量大的探测器方向移动。当两个探测器的距离小于给定阈值时合并为一个；当探测器处于稳定状态时，探测器的数目就是聚类的个数，并且探测器的位置就是初始中心点的位置。

2.2 GK 均值算法步骤

GK 均值算法的步骤简单地说就是先投放探测器，然后移动、合并它们，算法收敛时得到聚类数目和初始中心，最后利用 K 均值算法对数据集聚类。

探测器的投放。首先利用文献[14]提出的正交设计的方法求出一个正交矩阵，将该矩阵中每行作为一个坐标，则得到如图 1b 所示的均匀分布的点；然后将这些点映射在真实数据空间，就得到均匀分布的探测器，如图 2 所示。映射的方法是：正交矩阵的每一列就对应数据集的一维，每列中的 0 对应所在维中的最小值 min，1 对应 min+(max−min)/(Q−1)(Q 就是每一维数据可取值的个数)，以此类推就可以完成映射。根据图 2 所示的例子，探测器的值如图 3 所示。



(a)正交矩阵　　(b)空间分布

图 1　正交矩阵及分布

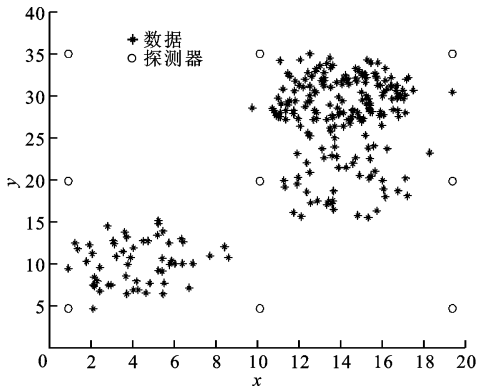


图 2　探测器的投放

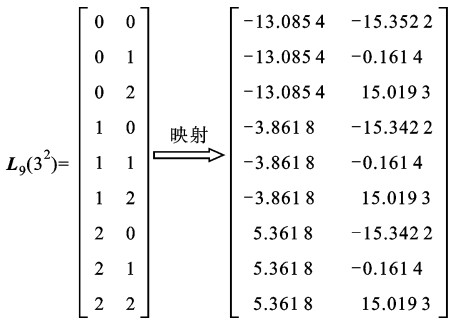


图 3　探测器的值

探测器的运动。利用万有引力的规律探测器开始运动，每个探测器的质量就是以自己为中心在半径为 E 的邻域内所含数据点的个数。探测器运动的步骤如下。

步骤 1　首先确定半径 E，利用式(4)进行计算

$$E = pS/(2Q-1) \tag{4}$$

式中：S 是数据集中离得最远的 2 个数据点之间的距离；Q 见式(3)；p 是调整参数，用来调控 E 的大小，依据实验得它的取值范围为 0.8～1.2。

步骤 2　求出每个数据点的质量，质量定义为以该点为中心在 E 邻域内所含数据点的个数(包括它自己)。

步骤 3　利用式(2)求出两两探测器之间的引力，为了计算简单将公式中的 $\boldsymbol{G}$ 简化为 1。

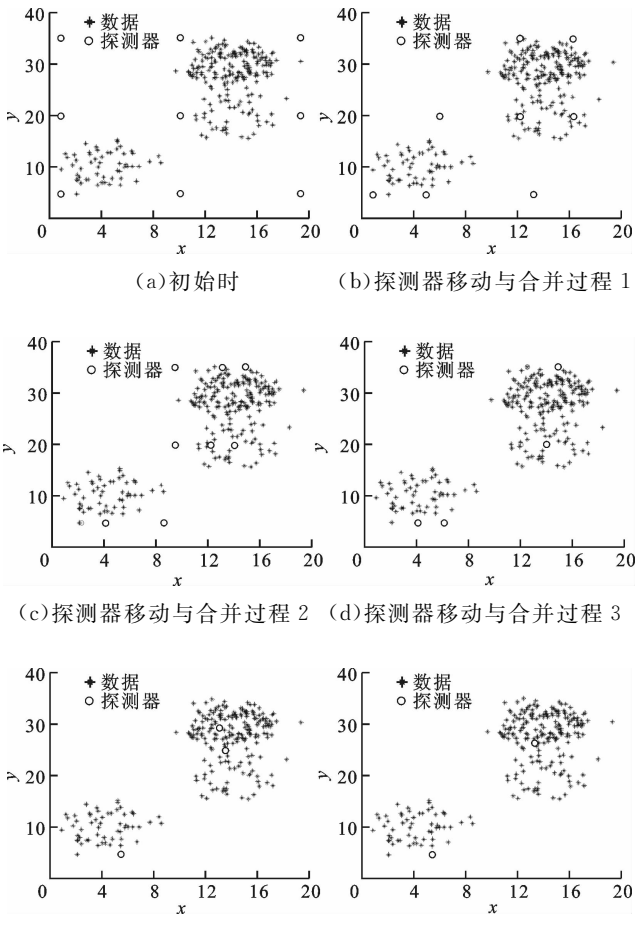
步骤 4　对每个探测器 y 分别求出在 E 邻域内对它引力最大的探测器 x，这里规定质量大的吸引质量小的，然后每个探测器 y 依次按照式(5)朝着 x 的方向移动。特别情况，当探测器 x 的 E 邻域内质量最大的就是自身时，探测器 x 不移动。

$$y(t+1) = (1-\lambda)y(t) + \lambda x(t) \tag{5}$$

式中：t 代表时间；λ 代表步长，即 y 朝着 x 移动的长度和它们距离的比值，$0<\lambda<1$，我们在实验中设定 $\lambda=1/3$。

步骤 5　如果 2 个探测器之间的距离小于 αE，就将它们合并成一个探测器，即保留质量较大的那个。这样探测器的数目就减少了。α 用来控制合并的速度，$0<\alpha<1$，实验中设定 $\alpha=0.25$。

步骤 6　重复步骤 2～5 直到探测器不再移动与合并，此时探测器的数目就是聚类的个数，探测器的位置就是初始的聚类中心。以图 2 中的数据为例，探测器的移动与合并过程如图 4 所示。特别情况，当 E 太小时，投放的探测器没有移动与合并，它们就全部作为初始中心，但初次分配数据时会出现有的探测器没有被任何一个点选中作为初始中心的情况，



(a)初始时　(b)探测器移动与合并过程 1

(c)探测器移动与合并过程 2　(d)探测器移动与合并过程 3

(e)探测器移动与合并过程 4　(f)终止时

图 4　探测器的移动与合并

这时就将实际被选中的探测器数目作为聚类的个数。

最后，将得到的类数与初始中心带入 K 均值算法进行聚类。

3 实验结果及分析

3.1 性能评价函数

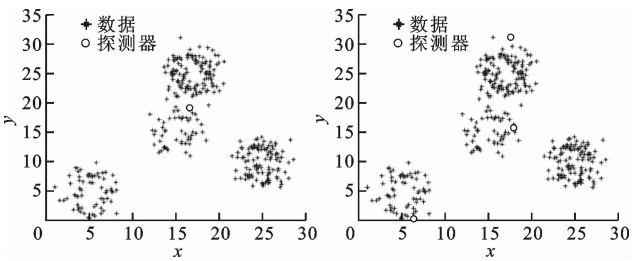
聚类算法的优劣通常是由聚类评价函数指标来判定的。本实验采用 DB(Davies and Bouldin)评价指标[15]，计算了类内元素间的分散度与类间距离的比率。具体定义为：第 i 个类 C_i 的内部分散度为 $S_i=(1/n_i)\sum_{x\in C_i}\|x-v_i\|^2$，其中 n_i 为类 C_i 的数据个数，v_i 为类 C_i 的中心；类间距离指的是各类中心点的距离，定义为 $d_{ij}=\|v_i-v_j\|^2$；任意两个类的内部分散度与类间距离之比为 $R_{ij}=(S_i+S_j)/d_{ij}$。DB 评价指标 I_{DB} 定义为

$$I_{\mathrm{DB}}=\frac{1}{c}\sum_{i=1}^{c}\max_{j\neq i}R_{ij} \tag{6}$$

式中：c 为聚类个数。显然 I_{DB} 值越小聚类效果越优，即聚类后类内分散度较小，而类和类之间距离较大。

3.2 实验结果

为了测试 GK 均值算法的性能，设计了 4 组实验：①类数随参数 p 变化；②对比迭代次数；③对比评价指标 I_{DB}；④对比聚类正确率。图 5 是第 1 组实验结果，采用人造数据，图中圆圈代表探测器最终的位置，实验表明类数随参数 p 变化。



(1)$p=1$，类数为 1　　(2)$p=0.9$，类数为 3

图 5 类数随 p 值的变化情况

第 2 组实验采用的数据同图 4 中的人造数据，实验结果见表 1。表中 p 是式(4)中的调整参数，C 代表类数。当参数 p 设定为 1.1 时，由 GK 均值算法得到聚类的数目为 2，评价指标 I_{DB} 值为最小。采用此数据集实验时，K 均值算法不会陷入局部最优，所得结果与采用 GK 均值算法所得结果一致，因此 I_{DB} 值也是相同的，故仅比较两种算法的迭代次数。由于 GK 均值算法会得到较好的初始中心，所以算法的迭代次数明显小于 K 均值算法。

表 1 两种算法的迭代次数对比

p	C	I_{DB}	迭代次数	
			GK 均值	K 均值
0.9	3	0.158 5	5	10
1.0	3	0.158 5	6	12
1.1	2	0.135 4	6	8

第 3 组实验采用的数据是 UCI 数据库中的 iris，实验结果见表 2。由于 K 均值算法初始点的选择是随机的，所以每次实验的结果可能不相同，表中列出了 10 次实验所得的 I_{DB} 的最大值、最小值和均值。对比两种算法所得的 I_{DB} 发现，GK 均值算法明显优于 K 均值算法。特别情况，当 $p=0.8$ 时，得到的 E 太小，9 个探测器没有移动与合并，初次分配时有 3 个探测器没有被选中作为中心点，所以此时的类数为 6。实验结果表明，当 $p=1.1$ 时 GK 均值算法可以得到理想的类数 3 和较好的初始中心，所以 I_{DB} 值最小。

表 2 iris 数据集 I_{DB} 值对比

p	C	GK 均值	K 均值		
		I_{DB}	$I_{\mathrm{DB}}^{\min}$	$I_{\mathrm{DB}}^{\max}$	$I_{\mathrm{DB}}^{\mathrm{ave}}$
0.8	6	0.497 1	0.605 2	0.955 1	0.681 4
1.0	4	0.446 9	0.443 5	0.450 3	0.447 1
1.1	3	0.290 3	0.294 4	0.870 1	0.409 5
1.2	1				

第 4 组实验采用了 UCI 数据库中的数据 iris、glass、wine 和 ecoli，用于比对两种算法的聚类正确率，实验结果见表 3。聚类的正确率用式(7)计算

$$\text{正确率}=\frac{\text{正确聚类的样本个数}}{\text{总样本个数}}\times 100\% \tag{7}$$

表 3 中列出了重复 10 次 K 均值算法所得的 $I_{\mathrm{DB}}^{\mathrm{ave}}$ 和正确率均值。实验结果表明，与 K 均值算法相比，GK 均值算法得到的 I_{DB} 值小而且聚类正确率高，所以它的性能优于 K 均值算法。

表 3 两种算法的聚类正确率对比 (%)

数据	p	C	GK 均值		K 均值	
			I_{DB}	正确率	$I_{\mathrm{DB}}^{\mathrm{ave}}$	正确率均值
iris	1.10	3	0.290 3	89.33	0.651 5	47.60
glass	0.82	6	0.475 8	44.86	0.782 6	36.45
wine	1.00	3	0.209 3	38.20	0.212 9	37.86
ecoli	0.90	8	0.847 5	18.75	1.155 8	11.96

4 结　论

本文利用正交设计算法在数据空间投放若干探测器，它们根据万有引力定律移动与合并，当探测器不再移动时，探测器的个数就是聚类的类数，探测器的位置就是 K 均值算法的初始中心。实验结果表明，当参数 p 在[0.8，1.2]内取值时，GK 均值算法可以得到理想的聚类数目；由于 GK 均值算法提供了较理想的初始点，所以得到了较好的聚类结果。本文提出的自动确定类数的算法框架也可以用在其他需要指定类数的聚类算法中。

参考文献：

[1] PENA J M, LOZANO J A, LARRANAGA P. An empirical comparison of four initialization methods for the K-means algorithm [J]. Pattern Recognition Letters, 1999, 20(10): 1027-1040.

[2] MACQUEEN J. Some methods for classification and analysis of multivariate observations [C]// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, USA: University of California Press, 1967: 281-297.

[3] ESTIVILL C V, YANG Jianhua. Fast and robust general purpose clustering algorithms [J]. Data Mining and Knowledge Discovery, 2004, 8(2): 127-150.

[4] SU Muchun, CHOU Chienhsing. A modified version of the K-means algorithm with a distance based on cluster symmetry [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 674-680.

[5] LIKAS A, VLASSIS M, VERBEEK J. The global K-means clustering algorithm [J]. Pattern Recognition, 2003, 36(2): 451-461.

[6] D'URSO P, GIORDANI P. A robust fuzzy K-means clustering model for interval valued data [J]. Computational Statistics, 2006, 21(2): 251-269.

[7] HUA Chunsheng, CHEN Qian, WU Haiyuan, et al. RK-means clustering: K-means with reliability [J]. IEICE Transactions on Information and Systems, 2008, E91-D(1): 96-104.

[8] TIMMERMAN M E, CEULEMANS E, DE R, et al. Subspace K-means clustering [J]. Behavior Research Methods, 2013, 45(4): 1011-1023.

[9] PELLEG D, MOORE A. X-means: extending K-means with efficient estimation of the number of clusters [C]// Proceedings of the 17th International Conference on Machine Learning. New York, USA: ACM, 2000: 727-734.

[10] HAMERLY G, ELKAN C. Learning the k in K-means [C]// Proceedings of the 17th Annual Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2003: 281-288.

[11] KALOGERATOS A, LIKAS A. Dip-means: an incremental clustering method for estimating the number of clusters [C]// Proceedings of the 6th Annual Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2012: 2402-2410.

[12] BAGIROV A M, UGON J, WEBB D. Fast modified global k-means algorithm for incremental cluster construction [J]. Pattern Recognition, 2011, 44(4): 866-876.

[13] TZORTZIS G, LIKAS A. The minmax k-means clustering algorithm [J]. Pattern Recognition, 2014, 47(7): 2505-2516.

[14] LEUNG Yiuwing, WANG Yuping. An orthogonal genetic algorithm with quantization for global numerical optimization [J]. IEEE Transactions on Evolutionary Computation, 2001, 5(1): 41-53.

[15] DAVIES D L, BOULDIN D W. A cluster separation measure [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 1(2): 224-227.

[本刊相关文献链接]

董哲，伊鹏. 采用链路聚类的动态网络社团发现算法. 2014，48(8)：73-79. [doi：10.7652/xjtuxb201408013]

张永斌，陆寅，张艳宁. 域名请求行为特征与构成特征相结合的域名变换检测. 2013，47(8)：54-60. [doi：10.7652/xjtuxb201308010]

许亚美，卢朝阳，李静，等. 手写维文字符分割中的多信息融合路径寻优方法. 2013，47(8)：68-73. [doi：10.7652/xjtuxb201308012]

王学恩，韩德强，韩崇昭. 采用不确定性度量的粗糙模糊 C 均值聚类参数获取方法. 2013，47(6)：55-60. [doi：10.7652/xjtuxb201306010]

夏虎，庄健，于德弘. 面向高维特征故障数据的进化软子空间聚类算法. 2013，47(5)：115-120. [doi：10.7652/xjtuxb201305021]

华莉琴，许维，王拓，等. 采用改进的尺度不变特征转换及多视角模型对车型识别. 2013，47(4)：92-99. [doi：10.7652/xjtuxb201304016]

李清华，康海燕，苑晓姣，等. 个性化搜索中用户兴趣模型匿名化研究. 2013，47(4)：131-136. [doi：10.7652/xjtuxb201304022]

杨攀，桂小林，田丰，等. 一种高效的用于话题检测的关键词元聚类方法. 2012，46(10)：24-28. [doi：10.7652/xjtuxb201210005]

（编辑　武红江）