文章编号:1000-6788(2006)02-0097-05

K-means 算法中的 k 值优化问题研究

杨善林,李永森,胡笑旋,潘若愚
(合肥工业大学计算机网络系统研究所,安徽 合肥 230009)

摘要: 在空间聚类中,最佳聚类数 k 求解的关键是构造合适的聚类有效性函数.典型 K-平均算法中的聚类数 k 必须是事先给定的确定值,然而,实际中 k 很难被精确地确定,使得该算法对一些实际问题无效.文章提出距离代价函数作为最佳聚类数的有效性检验函数,建立了相应的数学模型,并据此设计了一种新的 k 值优化算法.同时,给出了 k 值最优解 k_{opt} 及其上界 k_{max} 的条件,在理论上证明了经验规则 $k_{max} \leq \sqrt{n}$ 的合理性,实例结果进一步验证了新方法的有效性.

关键词: 空间聚类;K-平均算法;距离代价函数;k 值优化
中图分类号: TP391.4　　**文献标识码**: A

Optimization Study on k Value of K-means Algorithm

YANG Shan-lin, LI Yong-sen, HU Xiao-xuan, PAN Ruo-yu
(Institute of Computer Network System, Hefei University of Technology, Hefei 230009, China)

Abstract: In spatial clustering, the key factor to solve the problem of optimal class number is to construct a proper cluster validity function. The value of k must be confirmed in advance to exert K-means algorithm. However, it can not be clearly and easily confirmed in fact for its uncertainty. This paper recommends a distance cost function based on Euclidean distance to confirm the optimal class number, sets up a corresponding math model and designs a new optimization algorithm of k value. At the same time, the conditions of optimal solution k_{opt} and its up limit k_{max} are presented in this paper. The experiential rule which is usually expressed as $k_{max} \leq \sqrt{n}$ is theoretically proved to be reasonable. Results come from the example also show the validity of this new algorithm.

Key words: spatial clustering; K-means algorithm; distance cost function; optimization of k

1 引言

空间聚类与传统聚类方法的区别在于引入了空间距离维度,意在对具有空间分布属性的研究对象实现空间聚类.在空间聚类各算法中,最著名和经典的类别划分方法是 K-平均算法(K-means algorithm),K-中心算法(K-medoid algorithm)以及它们的变种[1,2].上述算法一般需要事先给定聚类数 k,但多数情况下,聚类数 k 事先无法确定,因此需要对最佳聚类数 k 进行优化处理.目前已提出了一些检验聚类有效性的函数[3~6],人们使用上述聚类有效性函数计算合适的聚类数 k,即最佳聚类数 k_{opt},但是,由于这些构造函数自身的缺陷,一般难以直接找到最佳聚类数 k_{opt},因此需要先确定一个搜索范围,就是要设定一个 k_{max},使得 $k_{opt} \leq k_{max}$.对于如何确定 k_{max},目前尚无明确的理论指导,多数学者使用的经验规则[4]为:$k_{max} \leq \sqrt{n}$.为了使研究更具针对性,本文仅对典型 K-平均算法的 k 值优化问题展开讨论.

典型的 K-平均算法以平方误差准则较好地实现了空间聚类,对于大数据集的处理效率较高[7].但是,该算法要求用户必须事先给出精确的 k(要聚类的数目)值,这在一定程度上影响和限制了其应用合理性.在实际中,k 值是难以准确界定的,用户无法知道采用什么样的 k 值聚类对自己更有利.因此,确定的 k 虽

收稿日期:2005-01-14
资助项目:国家自然科学基金(70471046);国家教育部博士学科点基金(20040359004)

作者简介:杨善林(1948-),男,安徽怀宁人,教授,博士生导师,主要研究方向为人工智能、计算机信息与控制系统智能决策支持系统等;李永森(1971-),男,安徽庐江人,博士生,主要研究方向为空间数据分析和处理、人工智能和空间决策支持系统等.

然对算法本身更方便、高效,但对一些实际问题并不有效. 据此本文提出了距离代价函数的概念,建立了相应的数学模型,并以距离代价最小准则实现了 k 值优化算法. 同时,给出了求 k 值最优解 k_{opt} 及其上界 k_{max} 的条件,并在理论上证明了经验规则 $k_{max} \leq \sqrt{n}$ 的合理性.

2 典型的空间聚类算法 ——K-平均算法

空间聚类是一种空间数据划分或分组的重要方法. 它是将研究对象的空间距离指标按照相似性准则划分到若干个子集中,使得相同子集中各元素间差别最小,而不同子集中各元素间差别最大. 通常的空间聚类算法是建立在各种距离基础上的,如欧几里得距离、曼哈顿距离和明考斯距离等. 其中,最常用的是欧几里得距离:

$$d(i,j) = \sqrt{[(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2]}, \tag{2-1}$$

式中, $i = (x_{i1}, x_{i2}, \cdots, x_{in})$ 和 $j = (x_{j1}, x_{j2}, \cdots, x_{jn})$ 是两个 n 维的数据对象.

根据空间聚类的一般原则,类别的划分应使得同一类(簇)的内部相似度最大、差异度最小,而不同类(簇)间的相似度最小、差异度最大. 空间聚类一般使用距离作为划分准则,即任一空间对象与该对象所属簇的几何中心之间的距离比该对象到任何其他簇的几何中心的距离都小.

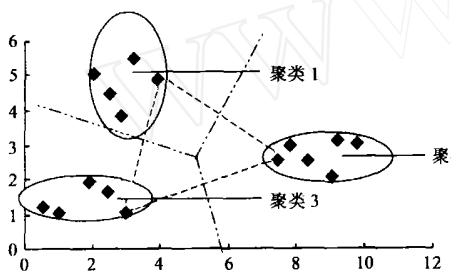

图 1 空间聚类过程

K-平均算法设计过程. 首先,由用户确定所要聚类的准确数目 k,并随机选择 k 个对象(样本),每个对象称为一个种子,代表一个簇(类)的均值或中心,对剩余的每个对象,根据其与各簇中心的距离将它赋给最近的簇. 然后重新计算每个簇内对象的平均值形成新的聚类中心,这个过程重复进行,直到下列(2-2)式准则函数收敛为止.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2. \tag{2-2}$$

这里, E 是所有研究对象的平方误差总和, p 为空间的点,即数据对象, m_i 是簇 C_i 的平均值. 按照这个准则生成的结果簇趋向于独立和紧凑(图 1).

3 距离代价函数及空间聚类 k 值优化算法

3.1 距离代价函数

对于如何求解最佳聚类数 k 和构造聚类有效性函数,不同学者给出了不同的答案,文献[8]归纳了以下几种常用的聚类有效性函数:

1) 分离系数:

$$F(U,k) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n} (u_{ij})^2, \tag{2-3}$$

假设 Ω 为所有聚类结果,那么 k 的最优选择由下式给出:

$$\max_k \{ \max_{\Omega} F(U,k) \}, \quad k = 2, 3, \cdots, n-1.$$

2) 分离熵:

$$H(U,k) = -\frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij} \log(u_{ij}), \tag{2-4}$$

其中, k 的最优选择由下式给出:

$$\max_k \{ \max_{\Omega} H(U,k) \}, \quad k = 2, 3, \cdots, n-1.$$

3) 紧致与分离性效果函数:

$$S(U,k) = \frac{\frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}^2 |x_j - c_i|^2}{\min_{i,j} |c_i - c_j|^2}, \tag{2-5}$$

k 的最优选择由下式给出:

$$\min_k\{\max_{\Omega} S(U,k)\},\quad k=2,3,\cdots,n-1.$$

上述几个聚类有效性函数,由于其自身的缺陷,如函数的非线性特征、单调性问题以及计算复杂度问题等,它们对于聚类有效性检验的效果并不理想,难以求解最佳聚类数 k. 比较而言,$S(U,k)$ 的性能较好,它反映了输入样本与它们相应聚类中心间距的平均值与聚类中心最小间距的比值,一个好的聚类应该使聚类中心的间距尽可能地大,而样本与其中心间距尽可能地小.

在研究空间聚类问题时,模型设计的合理性非常重要. 典型的 K-平均算法是在 k 被事先准确给定条件下实现的,由于算法本身较为完善、高效,K-平均算法一直占据着空间聚类算法的核心地位,但其对 k 值的严格要求大大限制了该算法的应用. 例如,在实际中,某物流公司准备在一个新的地区拓展其物流配送业务,公司根据该区域居民点密度和市场需求状况等拟投资建设一定数量的配送中心. 对于该公司来说,这是一个战略决策问题,实际上可以看作一个空间聚类问题来处理. 该用户对于建立多少个配送中心并不能提供一个准确的数据,他仅能根据以往的经验或直觉提供一个大致的数量范围. 这种情况下,直接运用 K- 平均算法难以奏效,必须对 k 值进行优化处理. 考虑到运输成本对于物流企业的核心作用,应该保证在完成相同业务量的同时使得运输成本达到最小. 据此,本文构造了距离代价函数,并以距离代价最小准则求解最佳聚类数 k,该方法对于实际问题具有一定的合理性.

定义 1　令 $K=\{X,R\}$ 为空间聚类的聚类空间,其中,$X=\{x_1,x_2,\cdots,x_n\}$,假设 n 个空间对象被聚类为 k 个簇,定义类际距离为所有聚类中心(簇内样本的均值)到全域中心(全体样本的均值)的距离之和:

$$L=\sum_{i=1}^{k}|m_i-m|,\tag{3-1}$$

式中,L 为类际距离;m 为全部样本的均值;m_i 为簇 C_i 所含样本的均值;k 为所要聚类的个数.

定义 2　令 $K=\{X,R\}$ 为空间聚类的聚类空间,其中,$X=\{x_1,x_2,\cdots,x_n\}$,假设 n 个空间对象被聚类为 k 个簇,定义类内距离为所有聚类簇内部距离的总和(其中,每个簇的内部距离为该簇内所有样本到其中心的距离之和):

$$D=\sum_{i=1}^{k}\sum_{p\in C_i}|p-m_i|,\tag{3-2}$$

式中,D 为类内距离;p 为任一空间对象,即样本;m,m_i,C_i,k 含义与式(3-1)相同.

定义 3　令 $K=\{X,R\}$ 为空间聚类的聚类空间,其中,$X=\{x_1,x_2,\cdots,x_n\}$,假设 n 个空间对象被聚类为 k 个簇,定义距离代价函数为类际距离与类内距离之和:

$$F(S,k)=L+D=\sum_{i=1}^{k}|m_i-m|+\sum_{i=1}^{k}\sum_{p\in C_i}|p-m_i|,\tag{3-3}$$

式中,$F(S,k)$ 为距离代价函数;其他变量的含义与式(3-1)、式(3-2)中相应变量的含义相同.

在运用距离代价函数作为空间聚类有效性检验函数时,本文确定了距离代价最小准则,即当距离代价函数达到最小值时,空间聚类结果为最优,k 的最优选择由下式给出:

$$\min_k\{F(S,k)\},\quad k=1,2,3,\cdots,n.$$

定理 1　令 $K=\{X,R\}$ 为空间聚类的聚类空间,其中,$X=\{x_1,x_2,\cdots,x_n\}$,假设 n 个空间对象被聚类为 k 个簇,L 为类际距离,D 为类内距离,当 $L=D$ 时,空间聚类数 k 达到优化,即符合经验规则:$k\leq\sqrt{n}$.

定理 1 的证明如下:令 $\bar{d}$ 为样本与其聚类中心的平均距离,$\bar{d}=D/n$;$\bar{l}$ 为聚类中心的平均距离,$\bar{l}=L/k$,当空间聚类具有分形几何特征时,即每个聚类内部的空间结构与整个聚类空间结构在形态上是相似的,此时应有:

$$\frac{\bar{l}}{L}=\frac{\bar{d}}{D/k}.\tag{3-4}$$

但是,实际空间聚类不一定具备分形几何特征,考虑问题的一般性,空间聚类应遵循紧致和分离性[8]要求,即一个好的空间聚类应该使各聚类中心的间距尽可能地大,而样本与其中心间距尽可能地小. 此时应有:

$$\frac{\bar{d}}{D/k} < \frac{\bar{l}}{L}. \tag{3-5}$$

当 $L = D$,即 $L = k\bar{l} = D = n\bar{d}$ 时,联立上述(3-4)和(3-5)两个方程,容易得到:$k^2 \leq n$,即 $k \leq \sqrt{n}$,这正是被很多学者所接收但又难以证明的经验规则[5].

3.2 基于距离代价函数的空间聚类 k 值优化算法

定理 1 为最佳空间聚类数的求解指出了一个途径,即可以先求出最优解的上界,这样便大大缩小了最优解的范围. 上文构造的距离代价函数由于其结构简单,计算复杂度较小,因而具有较好的检验效果. 本文据此设计了一个空间聚类 k 值优化算法. 该算法过程描述如下:

算法:在 K-平均算法基础上,通过距离代价函数优化 k 值.

输入:包含 n 个对象的数据库.

输出:距离代价函数最小条件下的 k^* 个簇.

方法步骤:

1) 根据经验规则计算和确定最优解的上界 $k \leq \sqrt{n}$;

2) 用 K-平均算法实现 $k \leq \sqrt{n}$ 所有数目下的空间聚类;

3) 根据距离代价函数分别计算不同聚类数目 k 下的 $F(S, k)$ 值;

4) 搜寻距离代价函数最小的 $F(S, k)^*$,并记下相应的 k^*;

5) 结束.

4 空间聚类 K 值优化求解实例

图 2 为某研究区内空间对象分布状况,图中共有 9 个研究对象(样本),其空间坐标如表 1 所示. 现在采用上文提出的空间聚类 k 值优化算法求解,考虑到经验规则 $k \leq \sqrt{n}$,应该有:$k \leq \sqrt{9} = 3$,因此 k 的取值范围可以缩小为 $k_1 = 1$, $k_2 = 2$, $k_3 = 3$,求解步骤如下:

表 1 研究对象的空间坐标

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
x	1	2	3	3	4	2	8	9	10
y	1	2	1	4	5	5	3	2	3

首先,按照经典的 K-平均算法,分别对 9 个研究对象进行 $k_1 = 1$, $k_2 = 2$, $k_3 = 3$ 时的空间聚类,形成如图 3,图 4,图 5 的空间聚类结果.

然后,根据上文构造的距离代价函数分别计算 $k_1 = 1$, $k_2 = 2$, $k_3 = 3$ 时的距离代价,结果为:$F(S,1) = 30.0$, $F(S,2) = 21.04$, $F(S,3) = 18.77$.

根据距离代价函数最小原则,$k_3 = 3$ 为最优解.

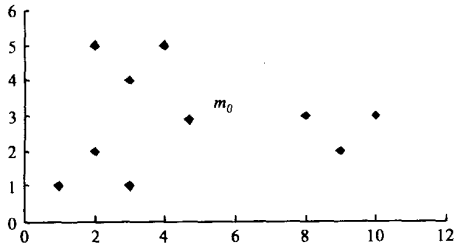

图 2 样本的空间分布状况(m_0 为样本均值)

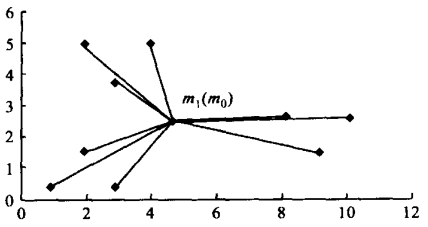

图 3 一个聚类及其距离代价(30.0)

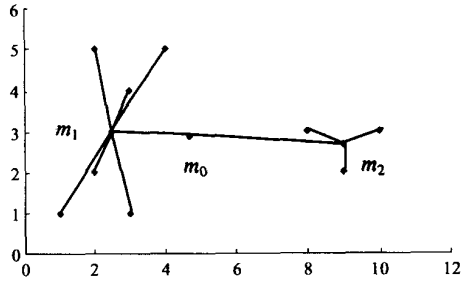

图 4　两个聚类及其距离代价(21.04)

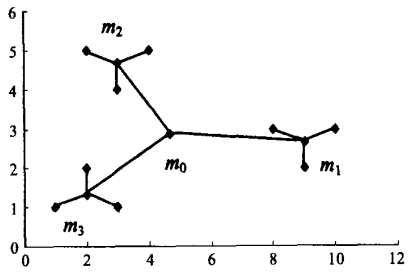

图 5　三个聚类及其距离代价(18.77)

在对最佳聚类数进行优化求解时,利用图像分析法可以快速找到最佳的 k 值. 由式(3-3)可知,距离代价函数 $F(S,k)$ 实际上是由两部分组成的,其中,L 为类际距离,是关于 k 的增函数,而 D 为类内距离,是关于 k 的减函数. $F(S,k)$ 的变化取决于两者的合成(图 6). 上文已经证明,当 $L = D$ 时,空间聚类能够达到优化,即 $k_{max} \leq \sqrt{n}$. 当然,这个解不一定就是最优解,但肯定是较优的解,它与最优解通常仅一步之遥. 实例中 L 和 D 的交点恰好为距离代价函数的最小值点,分析表明,$k_3 = 3$ 正是最优解.

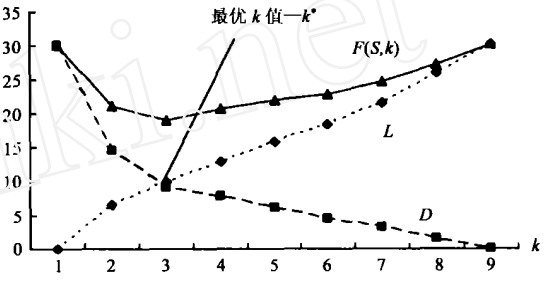

图 6　距离代价函数随 k 值变化趋势

5　结束语

典型的空间聚类 K-平均算法是在 k 被事先给定条件下实现的,但是,在实际应用中,k 难以准确确定,因此,在应用经典的 K-平均算法实现空间聚类时需要作进一步的优化和改进. 文章在分析和比较其他聚类有效性函数的基础上,构造了距离代价函数,并据此设计了空间聚类 k 值优化算法. 在对最佳聚类数优化求解过程中,确定了距离代价最小原则作为最优解的条件,进一步给出了最优解的上界 k_{max},从而大大缩小了最优解的范围,提高了新算法的效率,并从理论上论证了其合理性,算例结果和图像分析结果表明新方法是有效的.

参考文献:

[1] Treshansky A, McGraw R. An overview of clustering algorithms[A]. Proceedings of SPIE, The International Society for Optical Engineering [C]. 2001(4367):41 - 51.

[2] Clausi D A. K-means Iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation[J]. Pattern Recognition, 2002, 35:1959 - 1972.

[3] Bezdek J C, Pal N R. Some new indexes of cluster validity[J]. IEEE Transactions on Systems, Man, and Cybernetics _ Part B: Cybernetics, 1998,28(3) :301 - 315.

[4] Ramze R M, Lelieveldt B P F, Reiber J H C. A new cluster validity indexes for the fuzzy c-mean[J]. Pattern Recognition Letters, 1998,19:237 - 246.

[5] 于剑,程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学(E辑),2002,32(2):274 - 280.
Yu Jian, Chen Qiansheng. The range of optimal class number of fuzzy cluster[J]. Science of China (series E), 2002, 32(2):274 - 280.

[6] 范九伦,裴继红,谢维信. 聚类有效性函数:熵公式[J]. 模糊系统与数学,1998,12(3):68 - 74.
Fan Jiulun, Pei Jihong, Xie Weixin. Cluster validity function: Entropy formula[J]. Fuzzy Systems and Mathematics, 1998,12(3): 68 - 74.

[7] Jiawei Han,Micheline Kamber. 数据挖掘概念与技术[M]. 范明,孟小峰,等. 北京:机械工业出版社,2001.
Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques[M]. Fan Ming, Meng Xiaofeng, et al. Beijing: China Machine Press, 2001.

[8] 史忠植. 知识发现[M]. 北京:清华大学出版社,2002.
Shi Zhongzhi. Knowledge Discovery[M]. Beijing: Tsinghua University Press, 2002.