第 25 卷第 2 期　　四川理工学院学报(自然科学版)　　Vol. 25　No. 2
2012 年 4 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Apr. 2012
文章编号:1673-1549(2012)02-0077-04

K -均值算法中聚类个数优化问题研究

韩凌波

(中共湛江市委党校理论信息室, 广东　湛江　524032)

摘　要:在传统的 K -均值聚类算法中,聚类数 K 必须事先给定,然而,实际中 K 值很难被精确的确定, K 值是否合理直接影响着 K -均值算法的好坏。针对这个缺点,提出一种优化聚类数算法,根据聚类算法中类内相似度最大差异度最小和类间差异度最大相似度最小的基本原则,构建了距离评价函数 $F(S,K)$ 作为最佳聚类数的检验函数,建立了相应的数学模型,并通过仿真实验进一步验证了新算法的有效性。

关键词:k -均值算法;聚类个数;距离价值函数
中图分类号:TP311.12　　**文献标识码**:A

1967 年,MacQueen 首次提出了 K -均值算法。K -均值算法是典型的基于距离的聚类算法,是聚类分析中使用最广泛的算法之一。K -均值算法采用距离作为相似性的评价指标,即认为两个对象的距离越近,其相似度就越大。通过计算每个数据对象与 K 个聚类中心的距离,将数据对象划分到距离它最近的一个类,然后调整聚类中心,如上反复迭代直到聚类中心不再发生变化。传统的 K -均值算法中,聚类数 K 需要事先给定,但在实际中,聚类数 K 难以准确界定, K 值的选取是否合理直接决定着聚类效果的好坏。

针对这个问题,一些文章提出了一些检验聚类有效性的函数指标,通过对聚类有效性指标计算合适的聚类数 K, 即最佳聚类数 K_{opt},但是由于有效性指标自身的缺陷,一般难以直接找到最佳聚类数 K_{opt},需要确定一个合理的范围,使的 $K_{min} \leqslant K_{opt} \leqslant K_{max}$。对于如何界定 K_{min} 和 K_{max},目前尚无明确的理论指导[3],多数学者使用经验规则[4]认为:$1 \leqslant K_{opt} \leqslant \sqrt{n}$。该文对 Systems Engineering - theory & Practice 杂志上杨善林等人提出的一种距离代价函数作为聚类算法的有效性指标进行进一步的改进,根据 K -均值算法的基本原则:类间差异度最大相似度最小、类内相似度最大差异度最小,提出了一种新的聚类有效性评价指标,建立了相应的数学模型,并通过仿真实验进一步验证了新算法的有效性。

1　K -均值聚类算法

K -均值聚类算法是一种空间数据划分或分组处理的重要手段和方法。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差别最大。K -均值算法设计过程中,首先需由人工确定所要聚类的准确书目 K, 并随机选择 K 个对象,每一个对象代表一个簇(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的平均值形成新的聚类中心,这个过程重复进行,直到聚类准则函数收敛。

算法步骤如下:

(1)针对数据集 $\{x_1, x_2, \cdots, x_n\}$,任选 K 个样本作为初始聚类中心 $(z_1, z_2, \cdots, z_K)$。

(2)对每个样本 x_i 找到离它最近的聚类中心 z_v,并将其分配到 z_v 所标明的类。

(3)采取平均的方法计算重新分类后的各类心。

(4)计算 $D = \sum_{i=1}^{n} [\min_{r=1,2,\cdots,K} d(x_i, z_r)^2]$。

(5)如果 D 值收敛,则 return $(z_1, z_2, \cdots, z_K, U)$ 并终止算法,否则转至步骤(2)。

收稿日期:2011-11-09
基金项目:广西科学基金项目(0640067);广西研究生教育创新计划项目(2007106020812M73)
作者简介:韩凌波(1982-),男,山西晋中人,硕士,主要从事人工智能、模式识别和数据挖掘方面的研究,(E-mail)hanlingbo@163.com

2 基于距离评价函数的 K 值优化算法

2.1 确定聚类数 K 的取值范围

确定聚类数 K 的最佳范围[K_{min}, K_{max}],就是要确定 K_{min} 和 K_{max}, $K_{min}=1$ 指样本均匀分布,无明显特征差异,通常聚类数 K 最小为 2,即 $K_{min}=2$。对于 K_{max} 如何确定,目前尚无明确理论指导[3],多数学者使用经验规则: $K_{max} \leqslant \sqrt{n}$。文献[6-7,10]中数据集样本在理论上的分类数和实际的分类数也不符合上述规则。文献[4]中指出 K_{max} 小于等于 n 的平方根在不确定性函数 $f(x)$ 单调递减的情况下是合理的,并对 5 组 UCI 数据集进行试验,从理论上对 $K_{max} \leqslant \sqrt{n}$ 做出了一个合理的解释。综上所述, $K_{max} \leqslant \sqrt{n}$ 作为一种经验规则,不具有普遍性,但对大多数自然分布的样本是适用的,本文采用上述规则来确定聚类数 K 的取值范围。

2.2 距离评价函数

评价聚类结果优劣的过程称为聚类有效性分析。一般来说,一个好的聚类划分应尽可能反映数据集的内在结构,使得类内样本尽可能相似,类间样本尽可能不相似。从距离测度考虑,就是使类内部距离代价最小而类间距离代价最大的聚类最优。鉴于这种规则,本文设计了距离评价函数作为一种新的聚类有效性指标,该指标可以对 K -均值算法的聚类效果和最佳聚类 K 进行评价。

K -均值算法是一种数据划分或分组的聚类方法。通常基于划分的聚类算法是建立在各种距离的基础上的,本文采用欧氏距离[1]:

$$d(x_i,x_j) = (|x_{i1}-x_{j1}|^2 + |x_{i2}-x_{j2}|^2 + \cdots + |x_{ip}-x_{jp}|^2)^{\frac{1}{2}} \quad (1)$$

其中, $x_i=(x_{i1},x_{i2},\cdots,x_{ip})$ 和 $x_j=(x_{j1},x_{j2},\cdots,x_{jp})$ 是两个 p 维的数据对象。

根据空间聚类算法的一般规则,类的划分应该使得同一类的内部相似度最大、差异度最小,类与类之间的相似度最小,差异度最大,即任一空间对象与该对象所属的类的几何中心之间的距离比该对象到任何其他类的几何中心的距离都小,此时聚类准则函数收敛。

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p-m_i|^2 \quad (2)$$

其中,E 是所有研究对象的误差平方总和, p 为数据对象, m_i 是 C_i 的平均值。

据此上述基本思想构造了距离评价函数,并以距离评价函数最小为准则求解最佳聚类数 K。

样本数据集合: $S=\{m_1,m_2,\cdots,m_n\}$, K 为聚类个数。

定义 1 令 $I=\{S,K\}$ 为聚类空间,类间距离为所有聚类中心(类内样本均值)到全域中心(全体样本的均值)的距离之和:

$$D_{out} = \sum_{i=1}^{k} |m_i - m| \quad (3)$$

其中, D_{out} 为类间距离, m 为样本均值, m_i 为类 C_i 中所有样本的均值。

定义 2 令 $I=\{S,K\}$ 为聚类空间,类内距离为所有类内部距离的总和,类内部距离是指每个类内所有对象到类中心的距离之和:

$$D_{in} = \sum_{i=1}^{k} \sum_{p \in c_i} |p-m_i| \quad (4)$$

其中, D_{in} 为类内距离; p 为任一空间对象; m_i 为类 C_i 中所有样本的均值。

定义 3 令 $I=\{S,K\}$ 为聚类空间,根据文献[3],当 $D_{in}=D_{out}$ 时 K 值越接近最优,定义距离评价函数为:

$$F(S,K) = \left|\frac{D_{in}}{D_{out}} - 1\right| = \left| \frac{\sum_{i=1}^{K} \sum_{p \in c_i} |p-m_i|}{\sum_{i=1}^{k} |m_i - m|} - 1 \right| \quad (5)$$

在运用距离评价函数作为聚类有效性检验函数时,确定了距离代价最小准则,即当距离代价函数达到最小时,空间的聚类结果为最优,K 的最优选择为:

$$\underset{k}{Min}\{F(S,K)\}, K = 1,2,3\cdots,n \quad (6)$$

定义 4 C 令 $I=\{S,K\}$ 为聚类空间,根据经验规则[4],聚类数 K 的范围为:

$$1 < K \leqslant \sqrt{n} \quad (7)$$

2.3 基于距离评价函数的 K 值优化算法描述

根据距离评价函数设计了 K 值优化算法,该算法运算过程描述如下:

算法:在 K -均值算法的基础上,通过距离评价函数优化 K 值。

输入:包含 n 个数据对象的数据集。

输出:距离评价函数最小时的 K 值。

(1)根据定义 4 和公式(7)确定最优 K 值最佳范围。

(2)用 K -均值算法和第一步中确定的 K 值范围进行聚类有效性分析。

(3)根据距离评价函数分别计算不同聚类数目 K 下的距离评价函数 $F(S,K)$ 值。

(4)搜索距离评价函数 $F(S,K)$ 的最小值,并记下相应的 K 值。

(5)输出最佳聚类数 K。

3 实例分析

实验采用 MATLAB 7.0 开发环境编程实现,在 Inter

Pentium(R) CPU 2.79GHz,2G 内存,Windows Xp 操作系统的计算机上运行。

为了证实距离评价函数的有效性,本文采用二维空间分布的 12 个数据作为研究对象,12 个数据对象的分布如图 1 所示,12 个数据对象的二维坐标见表 1。

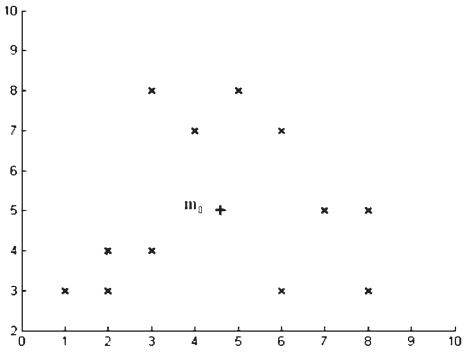

图 1　12 个样本二维空间分布图

表 1　12 个数据对象的二维坐标

	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}
x	1	4	8	6	7	2	8	3	5	6	3	2
y	3	7	5	7	5	3	3	4	8	3	8	4

首先,根据定义 4 和式(7),求 K 值范围:$1 < K \leqslant \sqrt{12}$,$K$ 取整后,K 的取值范围为 $K \in \{2,3,4\}$。

然后,根据 K－均值算法,对 12 个对象进行 $K=2$,$K=3$,$K=4$,时的聚类分析,形成如图 2,图 3,图 4 所示的二维空间聚类结果。

根据距离评价函数分别计算 $K=2$,$K=3$,$K=4$ 时的 $F(S,K)$ 值,$F(S,K)$ 的值如图 2、图 3、图 4 所示。

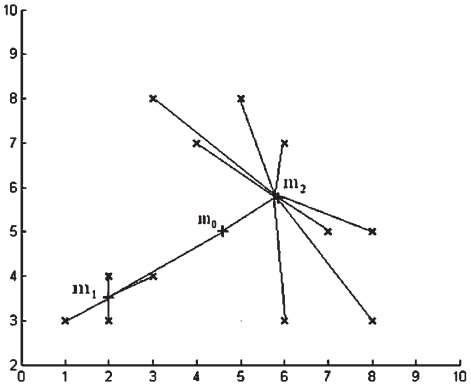

图 2　K=2 时 F(S,2) =0.55 值图

从图 5 可以看出,距离评价函数 $F(S,K)$ 主要取决于两个因素 D_{in} 和 D_{out}。D_{out} 是类间平均距离,是关于 K 的增函数,D_{out} 随着 K 的增加而增加;D_{in} 是类内平均距离,是关于 K 的减函数,D_{in} 随着 K 的增加而减小;当 D_{in} 和 D_{out} 越接近,距离评价函数 F(S,3)的值越小,空间的

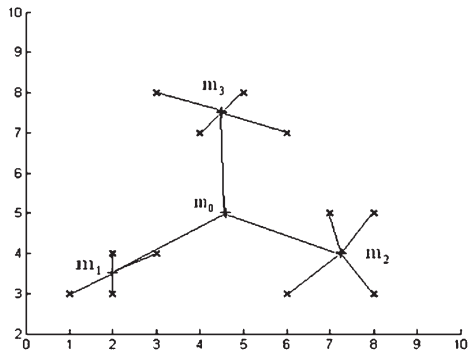

图 3　$K=3$ 时 $F(S,3)=0.04$ 值图

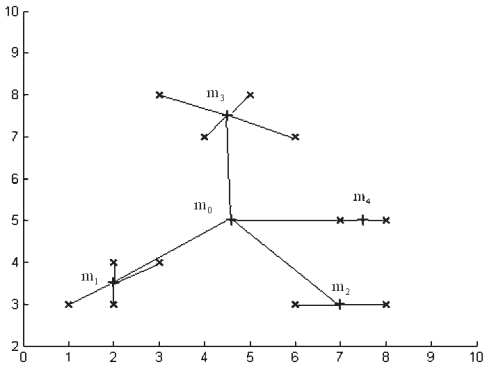

图 4　$K=4$ 时 $F(S,4)=0.31$ 值图

聚类效果达到优化,同时符合 $1 \leqslant k_{opt} \leqslant \sqrt{n}$。需要指出的是,距离评价函数最小时取得的解不是最优解,而是相对较优解,从图 5 中看出 D_{in} 和 D_{out} 的交点恰好为距离评价函数最小值,分析表明 $k=3$ 是最优解。

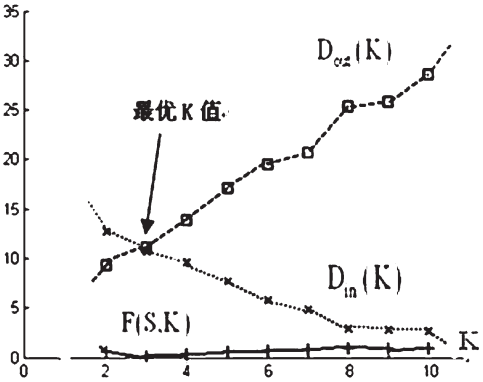

图 5　距离评价函数随 K 值变化趋势

4　结束语

本文针对传统的 K－均值算法中,难以确定聚类数 K 何时为最优的问题,提出一种基于距离评价函数对 K 值进行检验,阐述了算法的实现过程,并通过具体的实例分析了这种算法的可行性,实验证明基于距离评价函数的 K 值优化算法能够得到相对较好的 K 值。

参 考 文 献:

[1] Tan Pangning,Michael Steinbach,Vipin Kumar.In-troduction to Data Mining[M].Addison Wesley.2005.

[2] 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,1988,19:237-246.

[3] 杨善林,李永森,胡笑旋,等.K-means 算法中的 k 值优化问题研[J].系统工程理论与实践,2006(2):97-101.

[4] 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.

[5] 李永森,杨善林,马溪骏,等.空间聚类算法中的 K 值优化问题研究[J].系统仿真学报,2006,18(3):573-576.

[6] 孙吉贵,刘 杰,赵连宇.聚类算法研究[J].软件学报,2008,19(1):48-61.

[7] 孙即祥.现代模式识别[M].长沙:国防科技大学出版社,2002.

[8] Calinski R,Harabasz J.A dendrite method for cluster analysis[J].Communications in Statistics,1974,3(1):27.

[9] Kapp A V,Tibshirani R.Are clusters found in one dataset present in another dataset[J].Biostati-stics,2007,8(1):9-31.

[10] Frey B J,Dueck D.Response to comment on"clustering by passing messages between data points"[J].Science,2008,319.

[11] Frey B J,dueck D.Clustering by passing messages between data points[J].science,2007,315:972-976.

[12] 洪志令,姜青山,董槐林,等.模糊聚类中判别聚类有效性的新指标[J].计算机科学,2004,31(10):121-125.

[13] Dudoit S,Fridlyand J.A prediciton-based resampleing method for estimating the number of clusters in a dataset[J].Genome Biology,2002,3(7):1-21.

[14] 范九伦,裴继红,谢维信.模糊相关度与聚类有效性[J].西安电子科技大学学报,1998,25(1):13-16.

[15] 诸克军,苏顺华,黎金玲.模糊 C 均值中的最优聚类与最佳聚类数[J].系统工程理论与实践,2005,3:52-61.

Optimization Study on Class Number of K-means Algorithm

HAN Ling-Bo

(Department of Theory and Information, Zhanjiang Party Institute, Zhanjiang 524032, China)

Abstract: In traditional K-means algorithm, the class number must be confirmed in advance. However, it can not be clearly and easily confirmed in fact for its uncertainty. Whether the class number is optimized has a direct impact on the performance k-means algorithm. Considering this defection, a new improved algorithm is proposed. According to the basic principles of clustering algorithm that the Within-class similarity is Maximum and the within-class difference is least, the inter-class difference is maximum and the inter-class similarity is least, a distance cost of function $F(S,K)$ to confirm the optimal class number is recommended in this paper. A corresponding math model is set up, and example results further verify the effectiveness of the new algorithm.

Key words: K-means algorithm; clustering center; distance cost