文章编号: 1003-2843(2009)01-0198-03

基于 K-means 聚类算法的研究

步媛媛[1], 关忠仁[2]
(1. 成都信息工程学院计算机系, 四川成都 610225; 2. 成都信息工程学院网络中心, 四川成都 610225)

摘要: 原始的 k-means 算法[4]是从样本点的集合中随机选取 K 个中心, 这种选取具有盲目性和随意性, 它在很大程度上决定了算法的有效性. 为消除选取初始中心的盲目性, 应充分利用已有数据样本点的信息. 采取对数据进行预处理的方式来选取初始中心. 实验证明新的初始点的选取不仅提高了算法的计算效率, 也提高了算法最终确定的聚类的精度.

关键词: 数据挖掘; 聚类; k-means 算法; 聚类中心
中图分类号: TP392　　文献标识码: A

1 引言

聚类分析是数据挖掘中的一个重要功能, 目前已应用于许多方面:数据挖掘和知识发现、模式识别和模式分类、数据压缩和向量量化. 关于聚类分析有很多种方法, 这些方法包括分割与合并方法、随机化方法和神经网络方法. 其中在欧氏空间中的k-means聚类算法是最流行和最受关注的一种聚类分析算法.

k-means 是一种基于划分的聚类算法, 它的思想是当一个类确定后, 将类中数据点的几何平均值取为类的中心. 其中初始聚类中心的选择对聚类结果的影响是很大的. 如图所示, 图1是三个类的实际分布, 图2是选取了较好的初始聚类中心(+字标记的数据对象是聚类中心)得到的结果, 图3是选取不大好的初始聚类中心得到的结果. 从中可以看到, 图2所示的类内部数据对象相似度和类与类之间的相异度均高于图3所示, 最主要的体现是数据分布稠密. 因此合理地选择初始聚类中心是很关键的. 类似图3所示之类的选取聚类中心的k-means 算法的结果会导致聚类算法效率低, 算法迭代次数较多, CPU运行时间较长. 因此怎样找到一组初始中心点, 从而获得一个较好的聚类效果并提高聚类结果的精确度对k-means 算法具有重要意义.

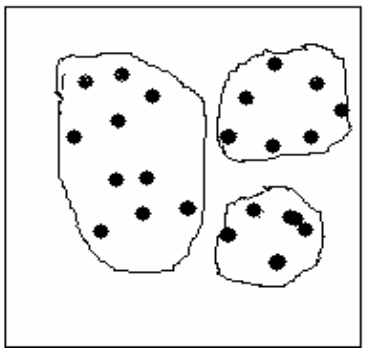
图 1　三个类的实际分布

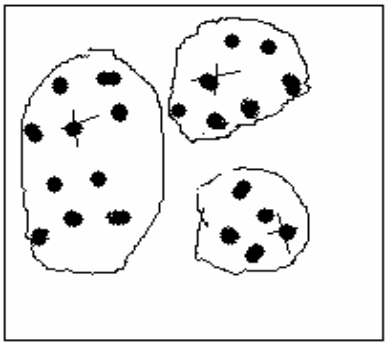
图 2　选取了较好中心的聚类结果

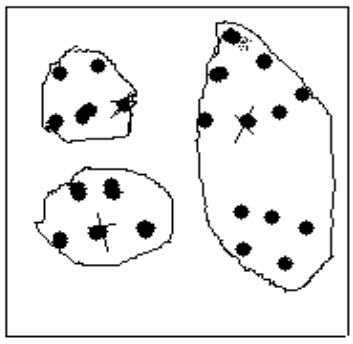
图 3　选取不好聚类中心的结果

本文提出了一种寻找初始聚类中心的方法, 使得初始聚类中心的分布尽可能体现数据的实际分布. 实验表明了这种算法的可行性和有效性.

2 原始的 k-means 聚类算法[4]及改进的算法分析

2.1 原始 k-means 聚类算法

收稿日期：2008-10-13
作者简介：步媛媛(1984-), 女, 成都信息工程学院计算机系在读硕士研究生; 关忠仕(1957-), 男, 成都信息工程学院网络中心高级工程师, 硕士生导师.

设待聚类的数据集: X=$\{x_1, x_2, \cdots, x_n\}$, k个聚类中心分别为 z_i, i=1, 2, ...n. 有如下定义:

定义1: 两个数据对象间的欧几里德距离为d(i, j)=$\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2}$

这里的i=($x_{i1}, x_{i2}, \cdots, x_{ip}$)和j=($x_{j1}, x_{j2}, \cdots, x_{jp}$)是两个p维的数据对象.

定义2: 准则函数E

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

这里的E是数据库中所有对象的平方误差的总和, p是空间中的点, 表示给定的数据对象, m_i 是簇 C_i 的平均值. 这个准则试图使生成的结果簇尽可能地紧凑和独立.

算法主要有三个过程组成: 首先是选取初始的聚类中心; 其次是样本点分类; 最后是聚类中心的调整. 其中后两个过程迭代交替进行. 下面是k-means 算法的流程描述:

输入: 簇的数目k和包含n个对象的数据库.

输出: k个簇, 使平方误差准则最小.

方法:

Step1　任意选择k个对象作为初始的簇中心;

Step2　repeat

Step3　根据簇中对象的平均值, 将每个对象重新赋给最类似的簇;

Step4　更新簇的平均值, 即计算每个簇中对象的平均值;

Step5　until不再发生变化

原始的k-means 算法对初始聚类中心的选择是随意的和盲目的, 这种选取方法很大程度上决定了算法的有效性和精确度. 因此对初始中心的选择进行改进既很有意义也很有必要. 本文的主要目的就是在欧几里德距离的意义下, 确定相隔最远的两个数据点之间的距离, 然后将数据集均分为k个段, 在每段内取一个中心作为初始的中心. 也就是改进上述算法中的step1.

2.2 改进的 k-means 聚类算法

定义1: 数据集中相隔最远的两个数据点之间的距离M

M=max{d(i, j)}

定义2: d=M/k

定义3: 假设一参照点为o, 数据集中与点o之间的距离最大的点记为数据集中的大者 m_i, i=1, 2, …, (k-1).

Step1　计算任意两个数据对象间的距离$d(x_i, x_j)$, 比较得出M.

Step2　计算d;

Step3　$X_1=X$;求出点 m_1;

C_1={ X_1 中与点 m_1 的距离小于d的点$\bigcup m_1$ };

Step4　X_2 =X- C_1 ;求出点 m_2;

C_2={ X_2 中与点 m_2 的距离小于d的点$\bigcup m_2$ };

……

Step5　X_{k-1} =X- C_{k-2} ;求出点 m_{k-1};

C_{k-1}={ X_{k-1} 中与点 m_{k-1} 的距离小于点d的点$\bigcup m_{k-1}$ };

Step6　C_K =X-($C_1\bigcup C_2\bigcup\cdots\bigcup C_{k-1}$);

Step7　计算中心 $z_i=\frac{1}{|C_i|}\sum_{x_j\in C_i}x_j$, i=1, 2, …, k.

Step8　从这k个聚类中心出发, 应用k-means聚类算法的步Step2, Step3, Step4, Step5, 得到聚类.

2.3 两种算法的比较分析

2.3.1 简单例子证明

为了说明改进的k-means聚类算法与原始的k-means聚类算法的不同, 举一简单例子进行分析比较.

例 设有一数据样本集合为X={1, 5, 10, 9, 26, 32, 16, 21, 14}，将X聚为3类，即k=3.分别用两种算法来执行.

(1)原始的k-means聚类算法如表1所示:

表 1 原始的 k-means 聚类算法

步	z_1	z_2	z_3	C_1	C_2	C_3	E
1	1	5	10	{1}	{5}	{10, 9, 26, 32, 16, 21, 14}	433.43
2	1	5	18.3	{1}	{5, 10, 9}	{26, 32, 16, 21, 14}	230.8
3	1	8	21.8	{1}	{5, 10, 9, 14}	{26, 32, 16, 21}	185.72
4	1	9.5	23.8	{1, 5}	{10, 9, 14, 16}	{26, 32, 21}	93.43
5	3	12.3	26.3	{1, 5}	{10, 9, 14, 16}	{26, 32, 21}	93.43

共迭代5次. 准则函数E一直是减小的.

(2)改进的k-means聚类算法如表2所示:

表 2 改进的 k-means 聚类算法

步	z_1	z_2	z_3	C_1	C_2	C_3	E
1	29	17	6.25	{32, 26}	{16, 14, 21}	{1, 5, 9, 10}	94.75

算法迭代次数为1，较传统k-means聚类算法明显减少. 聚类精确度较高.

2.3.2 实验结果

接下来对规模较大的数据进行数值实验以进一步说明改进的k-means聚类算法的有效性. 从CPU运行时间和迭代次数两个方面对它们的运行结果进行比较，说明改进的算法不仅是有效可行的，而且效率更高. 表3是实验结果的数值对照:

表 3 实验结果对照

数据维数	数据个数	聚类个数	传统算法的CPU耗时	传统算法的迭代次数	改进算法的CPU耗时	改进算法的迭代次数
20	100	5	0.29s	7	0.23s	4
40	200	10	1.972s	12	1.342s	4
80	1000	40	55.339s	29	33.828s	9

从表上可以得出结论: 改进的算法减少了迭代次数, CPU计算时间也明显减少.

3 结论

k-means 聚类算法是一种广泛应用的聚类算法，计算速度快，资源消耗少. 但是初始聚类中心选择的随机性决定了算法的有效性和聚类的精度. 本文提出了一种充分利用数据样本集的信息，对数据进行预处理，得出初始聚类中心从而进行聚类. 改进的算法为聚类节约了时间，而且聚类变得更加准确.

参考文献:

[1] 袁方，周志勇. 初始聚类中心优化的k-means 算法[J]. 计算机工程, 2007(5): 224-227.

[2] 连凤娜，吴锦林. 一种改进的k-means聚类算法[J]. 电脑与信息技术, 2008(4): 124-128.

[3] 徐义峰，陈春明. 一种改进的k均值聚类算法[J]. 计算机应用与软件, 2008(1): 27-31.

[4] JIAWEI HAN, MICHELINE KAMBER. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 北京: 机械工业出版社, 2007.

[5] 袁玉波，杨传胜. 数据挖掘与最优化技术及其应用[M]. 北京: 科学出版社, 2007.

Research of clustering algorithm based on K-means

BU Yuan-yuan[1], GUAN Zhong-ren[2]

(1. Department of Computer Sciences, Chengdu Institute of Information Engineering, Chengdu 610041, P.R.C.;
2. Internet Management Center, Chengdu Institute of Information Engineering, Chengdu 610041, P.R.C.)

Abstract: Original k-means clustering algorithm is the means that selects K centers randomly from the data sample cluster .This selection is blind and random, and to a certain extent the validity of algorithm lies on the selection. In order to avoid the blindness of selection, we should make full use of the information of existing data sample dot. We make pre-treatment of the data to choose the initial center. The experiment improves not only the calculation efficiency of algorithm, but also the precision of ultimate clustering.

Key words: data mining; clustering; K-means; clustering center