文章编号:1672 - 058X(2012)08 - 0047 - 05

一种对 K-means 算法的改进

李光明,李 梁,张建刚
(重庆理工大学 计算机科学与工程学院,重庆 400054)

摘 要:数据聚类是一个功能强大的技术,它能够把数据特征相似的对象划分为一类,但是并不是所有的聚类算法的实现都能产生相同的聚类结果;并且 K 均值算法的结果很大程度上依赖它的初始中心的选择;提出了一种新颖的关于 K 均值初始中心选择的策略;该算法是基于反向最近邻(RNN)搜索,检索一个给定的数据集,其最近的邻居是一个给定的查询点中的所有点;使用这种方法计算初始聚类中心结果发现是非常接近聚类算法所需的迭代聚类中心;对提出的算法应用到 K 均值聚类中给予了证明;用几种流行的数据集的实验结果表明了该算法的优点。

关键词:聚类;最近邻查询;反向最近邻搜索;K 均值
中图分类号:TP305 **文献标志码**:A

0 引 言

数据挖掘是指从大量数据中提取或“挖掘”知识[1]。聚类是一种无监督学习,它把形式相似的对象聚集在一起。使得在同一簇中的对象之间相似度大,不同簇中的对象之间的相似度小。聚类算法已广泛应用在模式识别、图像处理、过程优化、配方设计等许多领域中,并取得了良好效果,受到了人们广泛重视[2]。目前最流行的聚类算法之一就是 K-means 算法。

部分成果讲述了对 K 均值初始聚类中心的研究。Duda etal[3]讲述了用递归的方法初始化簇均值,此方法的另一种形式包括整个数据集的平均值,然后随机的用微扰函数干扰它 k 次。张玉英[2]提出了基于密度和聚类对象方向的改进算法。算法采取聚类对象分布密度方法来确定初始聚类中心,然后根据对象的聚类方向来发现任意形状的簇。连凤娜[4]提出了一种改进的 K-means 算法,主要从数据预处理、初始聚类中心的选择方面进行了改进。顾洪博[5]对近年来 K-means 算法的研究现状与进展进行总结。对较有代表性的初始聚类中心改进的算法,从思想、关键技术和优缺点等方面进行分析。

提出了一种基于反向最近邻(RNN)搜索的 K 均值稳定的初始化方案,这是一种与上述研究不同的方法。本方法的主要优点是通过该方法得到的初始划分比较接近最终解决方案,根据几个评估标准 K 均值的性能也得到了显著的提高。该方法的计算复杂度相对较低,并且样本集中的离群点也可以通过该方法检测出来。

收稿日期:2011 - 12 - 10;**修回日期**:2012 - 02 - 15.

作者简介:李光明(1986-),男,山东临沂人,硕士研究生,从事数据挖掘与数据分析研究.

1 反向最近邻搜索初始化方案的基础

1.1 反向最近邻(RNN)搜索

最近邻(NN)搜索是在一个度量空间中寻找最近点的问题。最近邻查询[6](Nearest Neighbor Query, NNQ):给定一查询点 q 和一个对象集 O,找出 O 中距离 q 最近的对象 o,即 $\mathrm{NNQ}(q)=\{o \mid \mathrm{distance}(q,o) \leqslant \mathrm{distance}(q,p), o,p \in O\}$;其中距离是采用欧氏距离或者曼哈顿距离。反向最近邻搜索是最近邻搜索的一个变种。最近邻搜索返回距离查询点最近的 k 个对象,而反向最近邻则返回将查询点作为其最近邻的对象集[7]。反向最近邻搜索[6](Reverse Nearest Neighbor Search, RNNS):给定一查询点 q 和一个对象集 O,找出 O 中所有把 q 作为最近邻的对象 o,即 $\mathrm{RNNS}(q)=\{o \mid \mathrm{distance}(o,q) \leqslant \mathrm{distance}(o,p), o,p \in O\}$。

因此,反向最近邻(Reverse Nearest Neighbor, RNN)查询[11]是在数据集中找到以查询点为最近邻的对象点,它可被应用到知识发现、决策支持、设施定位、地理信息系统和多媒体数据库等多种领域。反向最近邻搜索是检索给定的数据集,其最近邻的点是一个给定的查询点中的所有点。图1所示的数据集包括4个点 p_1, p_2, p_3 和 p_4。假设用欧氏距离来表示两点之间的相似性,给定点 q(实心点)的反向最近邻搜索的返回值是 p_1 和 p_2。特别是 p_1 是一个结果,因为它是 q 的最近邻点,p_1 是数据集中最接近 q 的点。重要的是要注意 q 的最邻近搜索不一定是 q 的反向最近邻搜索。那么反向最远邻居(RFN)定义为如果查询点 q 是 p 的最远邻居之一,那点 p 就是查询点 q 的反向最远邻居。

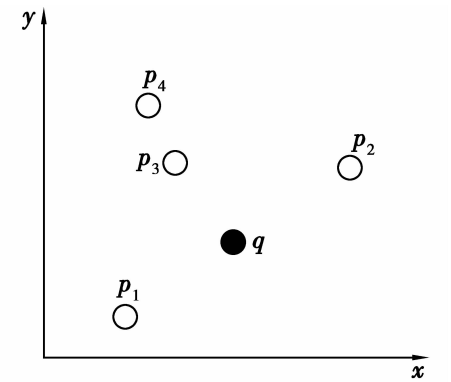

图1 反向最近邻搜索的定义

1.2 反向最近邻搜索的性质

定义[9](删除操作):设 p 是集合 D 中指定要删除的点,删除数据集 D 中所有的点,把 p 的反向最近邻搜索的点都放到数据集 D 中,重复上述过程。

引理1 迭代的删除反向最近邻搜索的点不会影响其他点的最近邻搜索。

引理2 每一个点都将被分配到一个集合中。

引理3 反向最近邻搜索可以删除离群点。

2 原始的 K-means 算法

2.1 算法步骤

K均值算法是以 k 作为输入参数,对 n 个连续值的对象样本集进行聚类的过程。初始从 n 个对象的样本集中随机的选取 k 各中心作为初始簇中心,然后将其余的数据点赋值到离它较近的簇中,然后再重新计算每个簇的均值,更新簇中心再计算剩余对象与各个簇均值的距离,将它指派到最相似的簇。不断地重复这个过程,直到每个簇不再变化或准则函数收敛为止。所以往往K均值的结果依赖于初始簇中心的选择,当初始簇中心不同时,聚类的结果也会不同。

原K均值算法伪代码如下:输入聚类个数 k,以及包含 n 个数据对象的数据样本集;随机选取 k 个对象

作为初始化 k 个聚类中心;设置迭代计数器 $t=0$;while($r\neq 0$)把样本点分到距离最近的聚类中心所代表的簇内;计算聚类目标函数 $J(t)$;$r=J(t)-J(t-1)$;重新计算各个聚类中心;$t=t+1$;输出聚类中心。

2.2　算法不足

算法对初始聚类中心以及样本的输入顺序敏感,不同的初始聚类中心及不同的样本输入顺序,导致不同的聚类结果。在用距离来衡量样本数据间的相似度时,该算法不适合有大小差别很大的簇存在的数据集。原算法对于噪声和离群点数据是敏感的[1]。

2.3　反向最近邻搜索初始化聚类中心算法

(1) 算法描述。首先初始化候选集(CS),计算候选集中的每个点的反向最近邻搜索,再按照每个点的反向最近邻搜索集的对象个数的多少降序排列;然后选择排序列表中的第一点作为候选质心,并从列表中级联的删除选定点和它的反向最近邻搜索集(根据删除操作)。如果该列表不为空,重做选择和删除的操作。每次迭代后让候选集是选定点的集合,重复上述过程,直到候选集中的对象数小于给定的 U 值(U 一般是分簇数 K 的 3 倍)。最后,在候选集(选定质心点的集合)中按照反向最远邻居(RFN)标准选定 k 个质心。

(2) 算法步骤。根据引理 1,第 3 步的时间复杂度会得到大大减小;并且离群点也会被检测出来。详细算法如下。

1. 初始化候选集 CS
2. While($|CS|<U$)
{
3. 计算出候选集中每个点的反向最近邻搜索;
4. 按照每个点的反向最近邻搜索集的对象个数的多少降序排列;
5. 选择排序列表中的第一点作为候选质心;
6. 从列表中级联的删除选定点(候选质心)和它的反向最近邻搜索;
7. 如果该列表不为空,回到 3;
8. 把所有的候选质心加入候选集中;
}
9. for($i=0;i<k;i++$)
{
10. 计算出候选集中每个点的反向最远邻居搜索;
11. 按照每个点的反向最远邻居搜索的对象个数的多少降序排列;
12. 选择排序列表中的第一点作为质心;
13. 从原始候选集中级联的删除选定点(质心)和它的反向最近邻搜索集;
}

3　实验结果

在这一节中,在此用 F_1 度量[10]等评价标准来比较反向最近邻搜索和传统的初始化方法对聚类性能的影响。

3.1　数据集来源

因为对不同的数据集聚类的性能有很大的差异,实验比较了不同的初始化方案对不同的数据集的结果。试验数据用从 UCI 机器学习库中下载的数据集。它们是“iris”,“wine”,“balance-scale”3 个样本集如表 1 所示。

表1 实验数据集的特征

数据集	样本的维数	样本的个数	分类的数量
Iris	4	150	3
Wine	13	178	3
Balance-scale	4	625	3

3.2 评估度量

用 F_1 度量度量方法来评估聚类的性能。用 N,$cnum'$ 和 $cnum$ 分别来表示给定数据集中实例的个数,实验得到簇的个数和原始数据集中簇的个数。因此,让 $cl_k(k=1,2,\cdots,cnum')$ 和 $class_l(l=1,2,\cdots,cnum)$ 分别表示获得的集群和实际的集群。并且每个实例的类标号是 $d_i \in cl_k, i=1,\cdots,|cl_k|$,它表示的值为 $c_j(j=1,2,\cdots,cnum)$。

F_1 度量措施已经是用来衡量聚类质量的方法,特别是聚类使用手动标记的时候。F_1 度量类似于精度度量措施(精度是在统计学习常用的方法),定义 F_1 度量如下:

$$F_1 = \frac{\sum_{k=1}^{cnum'}(|cl_k| \max F(cl_k, class_l))}{N} \tag{1}$$

其中,

$$P(cl_k, class_l) = \frac{|cl_k \cap class_l|}{|cl_k|} \tag{2}$$

$$R(cl_k, class_l) = \frac{cl_k \cap class_l|}{|class_l|} \tag{3}$$

$$F(cl_k, class_l) = \frac{2R(cl_k, class_l)P(cl_k, class_l)}{R(cl_k, class_l) + P(cl_k, class_l)} \tag{4}$$

3.3 实验结果分析

表2显示了实验数据用传统的初始化方法(InitRandomClusters,IRC)与改进后的初始化方法即反向最近邻搜索(RNN)的K均值聚类实验聚类结果性能的比较。

表2 F_1 度量结果

数据集	IRC	RNN
Iris	0.937	0.958
Wine	0.925	0.929
Balance-scale	0.453	0.520

表3 改进算法对 Iris,Wine 数据集的聚类结果

聚类数	数据集	准确率/%			
		传统	密度	距离	反向最近邻
$K=3$	Iris	84.69	89.33	78.40	95.97
	Wine	70.13	70.22	92.60	92.60

表3统计了 Iris 和 Wine 数据集分别用传统初始化方法(IRC),基于密度的初始化方法,基于两阶段最大最小距离法搜索出最佳初始聚类中心的方法和反向最近邻搜索(RNN)方法的聚类结果精确度比较情况。明显的可以看出反向最近邻搜索方法初始化聚类中心得到的聚类结果的准确率得到了明显的提高。此外,K均值算法用反向最近邻搜索比用其他方法运行到收敛需要的迭代次数较少。

4 结　论

K 均值算法计算速度快,资源消耗小,对于处理大数据集是相对可伸缩的和高效的,但是初始聚类中心的确定会直接影响聚类结果,如何取得有效的初始聚类中心是算法的关键。在此利用反向最近邻搜索提出了一种新的稳定的初始化 K 均值聚类方法,在实验中对于不同数据集合,用该方法总是较好的选择初始化中心数据。在实际应用中,应根据待聚类数据集的数据类型、聚类结构选择好的初始化簇中心的方法。在未来的应用中,该方法将用于更多的数据集中。对于大型的数据集,需要使用更复杂的数据结构,例如 M-tree结构来加速反向最近邻搜索,以更快取得聚类结果。

参考文献:

[1] JIAWEI H,MICHELINE K. 数据挖掘概念与技术[M].2 版.北京:机械工业出版社,2008
[2] 张玉英,孟海东. 数据挖掘技术中聚类方法的改进研究[J]. 包头钢铁学院学报,2005,24(4):338-341
[3] DUDA R O,HART P E. Pattern Classification and Scene Analysis[M]. New York: John Wiley and Sons,1973
[4] 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,2008,16(1):38-40
[5] 顾洪博,张继怀. 聚类算法初始聚类中心的优化[J].西安工程大学学报,2010,24(2):222-226
[6] 余海彦,郝忠孝.时空数据库中基于 TPR 一树的反向最近邻查询[J].哈尔滨理工大学学报,2007,12(3):87-90
[7] 王晓辉,曹泽文. 移动对象反向最近邻查询技术研究[J].计算机工程,2010,36(20):66-67
[8] 魏大刚,唐昌杰. 基于最优投影和动态阈值的最近邻搜索算法[J].四川大学学报,2006,43(4):777-782
[9] XU J L,XU B W. Stable Initialization Scheme for K-Means Clustering[J]. Journal of Natural Sciences,2009,14(1):24-28
[10] JASON D M. RENNIE. Derivation of the F-Measure[Z]. http://mathworld.wolfram.com/HarmonicMean.html
[11] 刘润涛,张佳佳. 基于 Voronoi 图的反向最近邻查询[J].计算机工程,2009,35(19):81-82,85

A Kind of Improvement for K-means Algorithm

LI Guang-ming, LI Liang, ZHANG Jian-gang

(School of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract: Data clustering is a powerful technology and can classify the objects with similar data characteristics into a class, however, the implementation of all clustering algorithms does not produce the same clustering results, moreover, the results of K-means algorithm largely depend on the selection of initial clustering center. This paper proposes a novel strategy about K-means initial clustering center selection, whose algorithm is based on reverse nearest neighbor (RNN) search and retrieves a given data set whose nearest neighbor is all points in a given inquiry point. The result by using this algorithm to calculate initial clustering center reveals that this center is very close to iterative clustering center needed by clustering algorithm. This paper also verifies the application of the proposed algorithm to K-means cluster and uses the experiment through several popular data sets to demonstrate the advantages of this algorithm.

Key words: cluster; nearest neighbor inquiry; reverse nearest neighbor search; K-means value

责任编辑:代小红