K-means 算法的初始聚类中心的优化

赖玉霞, 刘建平
LAI Yu-xia, LIU Jian-ping

浙江理工大学 信息电子学院, 杭州 310018
College of Electronic Information, Zhejiang Sci-Tech University, Hangzhou 310018, China
E-mail: lenslin2000@yahoo.com.cn

LAI Yu-xia, LIU Jian-ping. Optimization study on initial center of K-means algorithm. Computer Engineering and Applications, 2008, 44(10): 147-149.

Abstract: The traditional K-means algorithm has sensitivity to the initial centers. To solve this problem, an improved K-means algorithm based on density is presente. First it computes the density of the area where the data object belongs to; then finds K data objects all of which are belong to high density area and the most far away to each other, using these K data objects as the initial start centers. Theory analysis and experimental results demonstrate that the improved algorithm can get better clustering .and eliminate the sensitivity to the initial start centers.
Key words: clustering; K-means algorithm; density; clustering center; high density area

摘 要: 传统的 K-means 算法对初始聚类中心敏感, 聚类结果随不同的初始输入而波动, 针对 K-means 算法存在的问题, 提出了基于密度的改进的 K-means 算法, 该算法采取聚类对象分布密度方法来确定初始聚类中心, 选择相互距离最远的 K 个处于高密度区域的点作为初始聚类中心, 理论分析与实验结果表明, 改进的算法能取得更好的聚类结果。
关键词: 聚类; K-means 算法; 密度; 聚类中心; 高密度区域
文章编号: 1002-8331(2008)10-0147-03 **文献标识码**: A **中图分类号**: TP274

1 引言

随着数据库应用的普及, 人们正逐步陷入“数据丰富, 知识贫乏”的尴尬境地。而近年来互联网的发展与快速普及, 使得人类第一次真正体会到了数据海洋无边无际。而数据挖掘技术的出现, 使得人们能够利用智能技术将这巨大数据资源转换为有用的知识与信息资源, 从而能够科学地进行各种决策。

数据挖掘, 就是从大量的数据中提取出隐含的、以前不为人所知的、可信而有效的知识, 能够对数据进行再分析, 以期获得更加深入的了解, 并具有预测功能, 即可通过已有的历史数据预测未来。现有数据挖掘方法有多种, 其中比较典型的有关联分析、序列分析、分类分析、聚类分析等。其中聚类就是对大量数据进行分类, 使得同类内的数据相似度尽可能大, 相异度尽可能小, 而不同类间的数据的相似度尽可能小而相异度尽可能大。它可以发现不同数据的潜在特征, 实现对数据的分类。聚类分析作为数据挖掘系统中的一个模块, 既可以作为一个单独的工具以发现数据库中数据分布的深层信息, 也可以作为其他数据挖掘分析算法的一个预处理步骤, 在数据挖掘领域中, 是一项重要的研究课题。目前已经被广泛应用到许多领域, 如模式识别、数据分析、图像处理、市场分析、客户关系管理等。

K-means 算法是聚类分析中一种基本的划分方法, 因其理论上可靠、算法简单、收敛速度快、能有效地处理大数据集而被广泛使用, 但传统的 K-means 算法对初始聚类中心敏感, 从不同的初始聚类中心出发, 得到的聚类结果也不一样。因此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。

2 K-means 算法的基本思想

K 均值算法是一种得到最广泛使用的聚类算法。K 均值算法以 K 为参数, 把 n 个对象分为 K 个簇, 以使簇内具有较高的相似度, 而簇间的相似度较低。相似度的计算根据一个簇中对象的平均值来进行。算法首先随机地选择 K 个对象, 每个对象初始地代表了一个簇的平均值或中心。对剩余的每个对象根据其与各个簇中心的距离, 将它赋给最近的簇。然后重新计算每个簇的平均值。不断重复该过程, 直到准则函数收敛。准则函数如下:

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}|x-\bar{x}_i|^2$$ ($\bar{x}_i$ 为簇 C_i 的平均值)

K 均值算法描述:

输入: 簇的数目 K 和包含 n 个记录的数据库;

输出: K 个簇, 使平方误差准则最小。

(1) 任意选择 K 个记录作为初始的聚类中心;

(2) REPEAT;

(3) for j=1 to n do;

计算每个记录与 K 个聚类中心的距离, 并将距离最近的聚类作

作者简介: 赖玉霞(1979-), 女, 讲师, 主要研究方向计算机网络和信息通讯系统; 刘建平, 男, 副教授, 主要研究方向计算机网络与信息通讯系统。
收稿日期: 2007-07-16 修回日期: 2007-10-17

为该点所属的类;

(4) for i=1 to k do 对每个聚集,计算聚集的质心(聚集点的均值);

(5) compute $E=\sum_{i=1}^{k}\sum_{x\in C_i}|x-\bar{x}_i|^2$

(6) until E 不再明显地发生变化。

3 基于密度的 K-means 算法

传统的 K-means 算法对初始聚类中心敏感,选择不同的初始聚类中心会产生不同的聚类结果且有不同的准确率。本文的主要目的就是找到一组能反映数据分布特征的数据对象作为初始聚类中心,对数据进行划分,最根本的目的是使得一个聚类中的对象是相似的,而不同聚类中的对象是不相似的,也就是改进上述算法中的第(1)步。

在用欧氏距离作为相似性度量的 K-means 算法中,相互距离最远的 K 个数据对象比随机取的 K 个数据对象更具有代表性。不过在实际的数据集中往往有噪声数据存在,如果只是单纯地取相互距离最远的 K 个点来代表 K 个不同的类别,有时会取到噪声点,从而影响聚类效果。一般在一个数据空间中,高密度的数据对象区域被低密度的对象区域所分割,通常认为处于低密度区域的点为噪声点[1]。为了避免取到噪声点,取相互距离最远的 K 个处于高密度区域的点作为初始聚类中心。

3.1 概念

定义 1 密度参数:以空间点 x_i 为中心,包含常数 Minpts 个数据对象的半径称之为对象 x_i 的密度参数,用 ε表示。ε越大,说明数据对象所处的区域的数据密度越低。反之,说明数据对象所处区域的数据密度越高。

定义 2 空间两点的距离用公式:样本 X 和样本 Y 的距离公式:

$$d(i,j)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}$$

定义 3 一个样本点与一个样本集的距离定义为这个样本点与这个样本集中所有样本点当中最近的距离,则一个样本点 X 和一个样本集合 Z 之间的距离定义如下:

$$d(X,Z)=\min(d(X,Y)),Y\in Z)$$

定义 4 两个集合 S 和 V 之间距离定义为最近的两个数据对象 X 和 Y 之间的距离。

$$d(S,V)=\min(d(X,Y)),X\in S,Y\in V)$$

3.2 基于密度的 K-means 算法

通过计算每个数据对象的密度参数,就可以发现处于高密度区域的点,从而得到一个高密度点集合 D。在 D 中取处于最高密度区域的数据对象作为第 1 个中心 Z_1;取距离 Z_1 最远的一个高密度点作为第 2 个聚类中心 Z_2;计算 D 中各数据对象 X_i 到 Z_1,Z_2 的距离 $d(X_i,Z_1)$,$d(X_i,Z_2)$,Z_3 为满足 $\max(\min(d(X_i,Z_1)),\min(d(X_i,Z_2)))$ 的数据对象 X_i;Z_k 为满足 $\max(\min(d(X_i,Z_1)),\min(d(X_i,Z_2)),\cdots\min(d(X_i,Z_{k-1})))$ 的数据对象 $X_i,X_i\in D$。依此得到 K 个初始聚类中心。

基于密度的 K-means 算法描述如下:

输入:聚类个数 K 以及包含 n 个数据对象的数据集;

输出:满足目标函数值最小的 K 个聚类。

(1) 计算任意两个数据对象间的距离 $d(x_i,x_j)$;

(2) 计算每个数据对象的密度参数,把处于低密度区域的点删除,得到处于高密度区域的数据对象的集合 D;

(3) 把处于最高密度区域的数据对象作为第 1 个中心 Z_1,加入集合 Z 中,同时从 D 中删除。

(4) 从集合 D 中找距离集合 Z 最远的点,加入集合 Z,同时从 D 中删除。

(5) 重复(4),直到 Z 中的样本个数达到 K 个。

(6) 从这 K 个聚类中心出发,应用 K-means 聚类算法,得到聚类结果。

3.3 实验结果与分析

本论文使用我国部分省市年国际论文收录和国内专利申请量数据,原始数据点分布图如图 1 所示,运用工具 SPSS 调用 K-means cluster 过程来对我国的区域创新能力进行聚类分析,随机选取不同的初始聚类中心得出的结果如图 2 所示,不同的初始聚类中心产生不同的结果,改进后的基于密度的 K-means 算法和传统的 K-means 算法执行的结果如图 3 所示,北京创新能力最高,单独成为一类地区。上海、江苏成为第二类,创新能力次之。天津、浙江、安徽、山东、四川、辽宁、吉林、湖北、湖南、广东、陕西成为第三类地区,创新能力较强,其他地区创新能力较差,构成第四类地区,实验结果符合数据的初始分布情况。

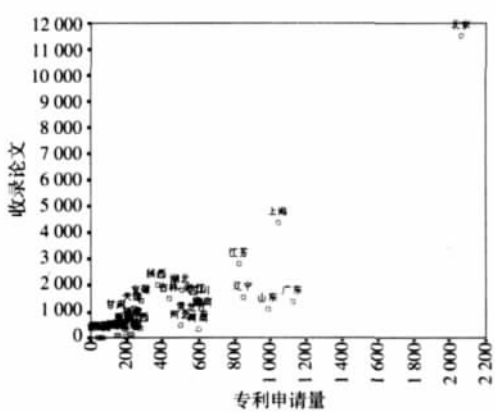

图 1 数据点散布图

Initial Cluster Centers

	Cluster			
	1	2	3	4
收录论文	11 541	4 391	2 003	10
专利申请量	2 062	1 047	373	39

Final Cluster Centers

	Cluster			
	1	2	3	4
收录论文	11 541	4 391	1 541	281
专利申请量	2 062	1 047	622	239

Initial Cluster Centers

	Cluster			
	1	2	3	4
收录论文	878	75	12	10
专利申请量	149	143	59	39

Input from FILE Subcommand

Final Cluster Centers

	Cluster			
	1	2	3	4
收录论文	11 541	2 514	1 086	148
专利申请量	2 062	720	531	198

图 2 K-means 算法产生的结果

从以上的实验的结果来看,随机选取初始中心的方法从不

	地区	收录论文	专利申请	改后簇	改后距离	改前簇	改前距离
1	北京	11541	2062	1	.00000	1	.00000
2	天津	1151	240	3	455.0391	3	347.4560
3	河北	472	503	4	325.6405	3	429.5485
4	山西	356	275	4	83.23013	4	251.4627
5	内蒙古	52	198	4	232.4729	4	88.09766
6	浙江	1469	587	3	46.59390	3	577.7512
7	安徽	1417	285	3	318.1985	3	552.0776
8	福建	610	240	4	329.2505	3	378.0300
9	江西	105	221	4	176.7145	4	65.84084
10	山东	1097	990	3	507.4068	3	545.1529
11	海南	13	68	4	317.9322	4	150.9022
12	重庆	314	191	4	58.75163	4	184.5605
13	四川	1368	609	3	57.57675	3	484.1449
14	贵州	85	154	4	213.5828	4	47.41517
15	云南	327	190	4	67.69881	4	197.2479
16	辽宁	1534	849	3	268.8734	2	595.1353
17	吉林	1496	440	3	177.7667	3	596.4316
18	黑龙江	755	568	4	576.9457	3	169.8025
19	上海	4391	1047	2	793.6599	2	2299.565
20	江苏	2819	827	2	793.6599	2	712.2838
21	河南	326	601	4	364.3830	4	480.9899
22	湖北	1822	506	3	408.4350	2	362.4335
23	湖南	936	628	3	489.9064	3	150.8174
24	广东	1385	1127	3	525.4547	3	806.9600
25	广西	102	232	4	178.9047	4	77.12633
26	陕西	2003	373	3	621.8606	2	364.1130
27	甘肃	878	149	4	604.0584	3	333.0950
28	青海	10	39	4	336.8690	4	172.4055
29	宁夏	12	59	4	323.7040	4	157.3647
30	新疆	75	143	4	227.2295	4	59.66054

Data View / Variable View

图3 改进的 K-means 算法和传统 K-means 算法产生的结果

同的初始中心出发产生不同的聚类结果,因而产生的聚类效果也不相同。这种随机选取初始聚类中心方法的聚类准确率不稳定,应用于实际的数据聚类,产生的聚类效果并不好,产生这样结果的原因就是随机的方法没有考虑到数据的分布情况,而只是给出了一个算法可以运行的必要条件(初始聚类中心)。而改进初值选取方法后的 K-means 算法,使用相互距离最远的 K 个高密度区域的数据点作为初始聚类中心的方法,保持了原有数据的实际分布情况,和原始的随机选取 K 个点的算法比较起来,最初多了一个优化初始聚类中心的过程,初始聚类中心不再是随机的选取,因此和原来算法比较起来更稳定了,也消除了孤立点对初始聚类中心的影响。

4 结束语

K-means 算法是一种应用非常广泛的聚类方法。关于 K-means 的算法的研究一直在深入,人们在原有基础上做了好多改进,比如:改变原有欧氏距离度量相似度的方式,在颜色提取中改进初始聚类中心选取方法,选取出现频率最高的方法等。另外还有许多其它改进算法。但是,很少有文章从数据本身分布来着手改进初始聚类中心的选取。本文改进的方法根据数据的自然分布来选取初始聚类中心,找出对象中分布比较密集的区域,这正是聚类的目的,从而摆脱了随机选取聚类中心对聚类结果产生的不稳定性,以及用质心代表一个簇所带来的"噪声"和孤立点数据对聚类结果的影响。

参考文献:

[1] 朱明.数据挖掘[M].合肥:中国科学技术大学出版社,2002:138-139.

[2] 荆丰伟,刘冀伟,王淑盛.改进的 K-均值算法在岩相识别中的应用[J].微计算机信息,2004(7):41-42.

[3] 刘立平,孟志青.一种选取初始聚类中心的方法[J].计算机工程与应用,2004,40(8):179-180.

[4] 张玉芳,毛嘉莉,熊钟阳.一种改进的 K-means 算法[J].计算机应用,2003:31-34.

[5] 毛国君,段立娟,王实,等.数据挖掘原理与算法[M].北京:清华大学出版社,2006:163-166.

[6] Guha S,Rastogi R,Shim K.Cure:an efficient clustering algorithm for large database[C]//Proc of ACM-SIGMOND Int Conf Management on Data,Seattle,Washington,1998:73-84.

[7] Gan W Y,Li D E.Hierarchical clustering based on kernel density estimation[J].Journal of System Simulation,2004,16(2):302-309.

[8] Ester M,Kriegel H P,Sander J.A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proc 2nd Int Conf on Knowledge Discovery and Data Mining,Portland,1999,20:226-231.

[9] 袁方,孟增辉,于戈.对 K-means 聚类算法的改进[J].计算机工程与应用,2004,40(36):177-178.

(上接 134 页)

[11] Chan K L,Misic V B,Misic J.Efficient polling schemes for bluetooth picocells revisited[C]//Presentation at HICSS-37-Mlnltrack on Wireless Personal Area Networks,Big Island,Hawaii,Jan 2004.

[12] Lee Y Z,Kapoor R,Gerla M.An efficient and fair polling scheme for bluetooth[C]//Proceedings MILCOM,2002,2:1062-1068.

[13] Kalia M,Bansal D,Shorey R.Data scheduling and SAR for bluetooth MAC[C]//2000 IEEE Vehicular Technology Conference(VTC-2000),May 2000.

[14] Kazemian H B.An intelligent video streaming technique in wireless communications[R].IFAWC2006,Mobile Research Center,TZI University Bremen,Germany,2006-03.

[15] Bluetooth SIG.Specification of the Bluetooth System.Audio/Video Distribution Transport Protocol[S].2003-05-22.

[16] Bluetooth SIG.Specification of the Bluetooth System.Audio/Video Control Transport Protocol[S].2003-05-22.

[17] Bluetooth SIG.Specification of the Bluetooth System.Audio/Video Remote Control Protocol[S].2003-05-22.

[18] Vilovic I,Zovko-Cihlar B.Performance of the bluetooth-based WPAN for multimedia communication[C]//4th EURASIP Conference Focused on Video/Image Processing and Multimedia Communications,July 2003,2:783-788.

[19] Prit A,Bilan P S.Streaming audio over bluetooth ACL links[C]//Proc 2003 International Conference on Information Technology:Computers and Communications(ITCC'03),2003:287-291.

[20] 房胜,梁永全.蓝牙与 802.11b 视频流传输性能研究和比较[J].计算机科学,2006,33(7):81-83.

[21] Chen L J,Kapoor R,Sanadidi M Y,et al.Enhancing bluetooth TCP throughput via link layer packet adaptation[C]//The 2004 IEEE International Conference on Communications(ICC 2004),Paris,France,2004.

[22] BlueCore™.CSR's Implementation of HCI on BlueCore,AN107,2002.

[23] 杨帆,王珂,钱志鸿.蓝牙分组传输性能分析与自适应分组选择策略[J].通信学报,2005,26(9):97-102.