基于划分的聚类个数与初始中心的确定方法

征 原,谢 云
(南京邮电大学 江苏省无线通信重点实验室,江苏 南京 210003)

摘 要:k 均值聚类算法在对数据进行聚类时需要以确定的聚类个数和初始聚类中心为前提,但聚类个数是难以准确给定的,通常随机选取 k 个样本作为初始聚类中心,由于不同的初始聚类中心可能导致不同的聚类结果,采用随机选取初始聚类中心的方法存在着较大的盲目性,造成聚类结果极不稳定。为此,提出了一种基于划分的聚类个数与初始中心点的确定方法。该方法通过对数据空间进行划分,统计每个网格空间中数据点数目作为网格的数据密度,同时计算局部密度极大值的网格个数;按照不同的分度值对数据集进行划分,当局部密度极大值的网格个数趋于相对稳定时,将局部密度极大值的网格个数作为聚类个数,并同时获得聚类初始中心。基于机器学习数据库数据集以及随机生成的人工模拟数据集进行了仿真实验,实验结果表明,所提出的算法有效可行,具有较高的准确性。

关键词:k 均值聚类;聚类个数;初始聚类中心;划分
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2017)07-0076-03
doi:10.3969/j.issn.1673-629X.2017.07.018

A Determination Method for Clustering Numbers and Initial Centers Based on Partitioning

ZHENG Yuan, XIE Yun
(Jiangsu Key Lab of Wireless Communications, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The k-means clustering algorithm needs the determined clustering number and initial clustering center before data clustering. However, the clustering number is difficult to be accurately given. Since different initial clustering centers may lead to distinct clustering results, the randomly selective method of initial clustering centers exists blindness to make clustering results very instable. Therefore, a new algorithm for determining optimal number of clusters and initial centers with partitioning has been proposed, in which partition of data space has been conducted to take the statistical number of data marker inside each grid as the data density in the grid and count the grid number with local maximum density. The data set has been partitioned according to the different index value. While the number of local maximum density grid tends to be relatively stable, it can be considered as cluster number and initial cluster centers can be acquired meanwhile. Simulation experiments for verification have been conducted with UCI data sets and random artificial data sets. The experimental results show that the proposed algorithm is effective and feasible with quite fine accuracy.

Key words: k-means clustering; number of clustering; initial clustering centers; partitioning

0 引 言

数据挖掘已是人工智能研究中的热点领域,而聚类作为该领域一项重要的分析手段,可以将数据集中的相似对象聚集成类,以便于进行相似对象的群体性研究[1]。迄今为止,众多的聚类算法被提出并应用,其中 k 均值聚类算法以其简单、快速之优势成为聚类分析中使用最为广泛的算法之一[2]。

k 均值算法是以确定的聚类个数和选定的初始聚类中心为前提,而在实际中,聚类个数是难以准确给定的,通常需要先确定一个搜索范围,搜索范围的确定通常需要用户根据对数据集的了解估算出一个范围[3-4];初始聚类中心则需要用户在数据集中随机选取 k 个数据点直接作为初始聚类中心,存在很大的盲目性[5-6]。

为了解决现有聚类算法无法给出聚类个数和初始聚类中心的问题,提出了一种基于划分的聚类个数与

收稿日期:2016-07-07 **修回日期**:2016-10-20 **网络出版时间**:2017-04-28
基金项目:国家自然科学基金资助项目(61471203,61101105);教育部博士点基金(20113223120001);江苏973项目(BK2011027)
作者简介:征 原(1992-),男,硕士研究生,研究方向为数据挖掘、人工智能、聚类分析。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20170428.1703.056.html
万方数据

初始中心的确定方法。在该方法中,考虑到数据集包含多个类,各个类中心附近的数据分布密度较大,而距离类中心相对较远时数据分布密度较小,对于某一聚类中心,距离其较近的空间中分布的数据点数目通常会大于距离其较远的空间中分布的数据点数目,通过对数据集进行划分可以形成多个局部极大网格。采用划分的方法不仅能计算出聚类的个数,同时还可以利用局部极大值网格内的数据有效地计算出接近于真实类心的初始聚类中心,并可以有效滤出噪声点,避免将噪声点选为初始聚类中心。

1　k 均值聚类基本算法

1.1　k 均值聚类算法

该算法首先选定 k 个类和 k 个初始聚类中心,按最小距离原则将各样本分配到 k 类中的某一类,之后不断计算类中心和调整各样本点所属的类别,最终使各样本到其所属类别中心的距离平方之和最小。算法步骤[7-8]如下:

(1)在数据集合 $U=(u_1,u_2,\cdots,u_m)$ 中任选 k 个数据点,将其作为初始聚类中心点 $Q=(q_1,q_2,\cdots,q_k)$。

(2)对于每一个样本点 u_i,计算其到每一聚类中心的距离,选取距离它最近的聚类中心 q_j,将其划分到聚类中心 q_j 标明的类中。

(3)采取求均值的方法计算重新分类后的各聚类中心。

(4)计算距离函数 $E(U,Q)=\sum_{j=1}^{k}\sum_{i=1}^{m}d(u_i,q_j)$,如果 $E(U,Q)$ 收敛,则输出最终的初始聚类中心和每个类中的成员;否则转向步骤(2)。

1.2　传统确定最佳聚类数的算法

传统的确定最佳聚类数的基本思想[9-10]是:首先估算出最佳聚类数所处的聚类个数搜索范围,利用聚类算法根据搜索范围产生不同聚类数目的聚类结果,选择合适的有效性指标[11-12]对聚类结果进行评估,根据评估结果确定最佳聚类数。

传统 k 均值算法的最佳聚类数确定算法归纳如下[13-14]所示:

(1)确定聚类个数的搜索范围 $[k_{\min},k_{\max}]$,通常选取 $k_{\min}=2$, $k_{\max}=\mathrm{int}\sqrt{n}$。

(2)对于 $k\in[k_{\min},k_{\max}]$,随机选取 k 个初始聚类中心,利用 k 均值聚类算法,当聚类结果符合终止条件时,利用聚类结果计算有效性指标。

(3)比较各聚类结果的有效性指标,选取有效性指标最优时所对应的 k 为最佳聚类个数 k。

(4)输出聚类结果:最佳聚类个数 k,聚类中心和每个类中的成员。

万方数据

2　基于划分的聚类个数与初始中心点的确定方法

在研究 k 均值聚类基本算法的基础上,提出了一种基于划分的聚类个数与初始聚类中心的确定方法。该方法通过对数据集进行划分后,统计每个网格中的数据点数目作为网格中的数据密度,通过计算局部密度极大网格的个数来确定聚类个数。考虑到在数据集中,对于某一聚类中心,距离其较近的空间中数据点的分布相对密集,而距离聚类中心较远的空间中数据点的分布相对稀疏,因此,距离每个聚类中心较近的单位空间分布的数据点数目一定会大于较远的单位空间。将数据集在每一维度上都划分出 x 个区间,从而将数据集划分为彼此独立的网格空间,每个数据点只属于一个独立的网格空间,计算出每个网格空间中数据点的数目,如果某一网格空间中含有的数据点数目多于与其相邻的网格空间,则称该网格空间为局部极大值网格。当按照不同的分度值进行划分,落入每个网格的数据点个数不相同,局部最大值网格的个数也不相同。当划分分度值较大,局部极大值网格的个数较少,可能使得两个局部极大值网格被划分在同一个网格,被认为是一个局部极大值网格。当划分分度值较小时,一些噪声点则被误划分为局部极大值网格,从而导致局部极大值网格的个数迅速变大,不符合实际情况。因此,在划分较为合理时,局部极大值网格的个数会处于一个相对稳定的状态,即可作为聚类的个数。而当划分不合理时,局部极大值网格的个数会随着划分的不同出现较大变化。现结合一个实例来详细描述。

图 1 中的数据集为利用 Matlab 随机生成的二维数据集,每个类中的数据点个数分别为 200、300、500;图 2 将数据集的每维等分为 15 个区间,编程获取每个网格中数据点的数目,并统计局部极大值网格的个数。通过对数据集进行不同的划分,可以得到其所对应的局部极大值网格的个数,做出划分区间个数与局部极大值个数曲线图,如图 3 所示。

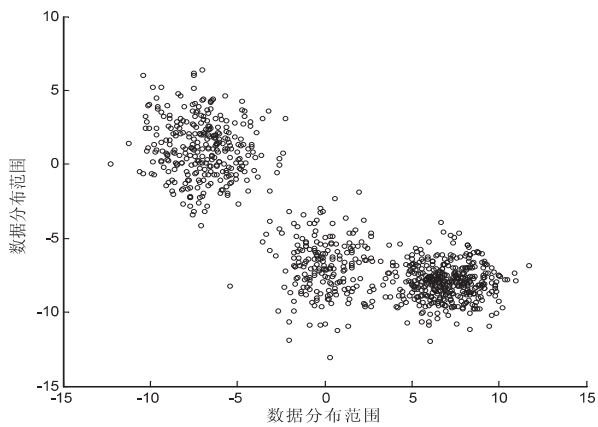

图 1　数据集样本分布图

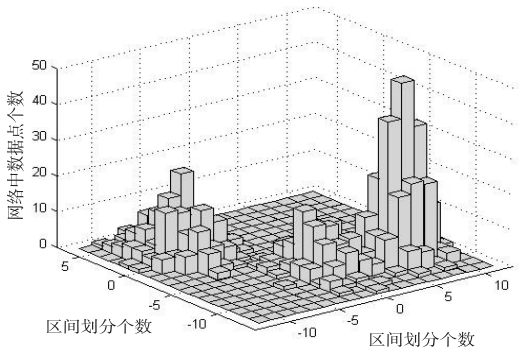

图 2　数据集样本分布立体图

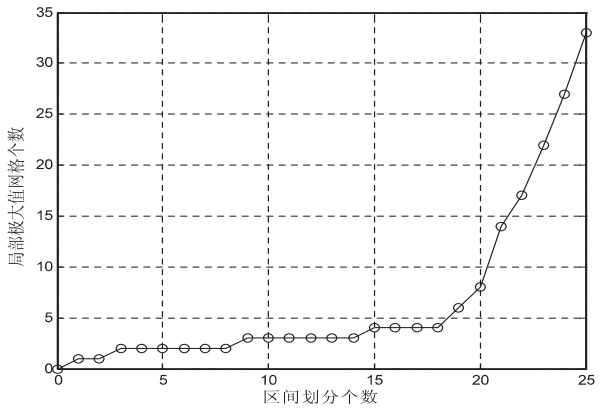

图 3　划分区间个数与局部极大值个数曲线图

基于上述考虑,认为在对数据集划分合理时(即局部极大值的个数趋于相对稳定时),用局部最大值的网格个数来估算聚类个数是比较合理的。通过将每一维等分为 x 个区间, x 从 1 开始计算每次划分后的局部极大值网格的个数,当局部极大值网格的个数趋于相对稳定时,即可估算得到聚类个数 k 。算法步骤如下:

输入:含有 m 个数据点的 n 维数据集 U 。

输出:数据集类的个数 k ,同时输出每个类的初始聚类中心。

(1)将数据集在每一维度上都划分出 x 个区间,数据集被划分为彼此独立的网格空间,计算出每个网格空间中数据点的数目。

(2)通过遍历所有网格,统计局部极大值网格的数目 k ,并利用点 (x,k_x) 做出划分区间数目与局部极大值网格数目的曲线图,计算曲线斜率,当斜率达到最小值时算法即可结束,转到步骤(3);否则取 $x = x + 1$,重复步骤(1)。

(3)将曲线斜率达到最小值时对应的局部极大值个数作为聚类个数 k, 并将每个网格内的数据取平均值作为每个类的初始聚类中心。

3　仿真验证

为了验证基于划分的聚类个数与初始中心的确定方法的有效性,采用 Matlab 仿真工具,测试了该方法在多种数据集下的效果,现以最具代表性的 6 组数据集进行说明,其中 4 组选自 UCI 数据库,另外 2 组利用 Matlab 仿真工具随机生成。6 组数据集的特征如表 1 所示。

表 1　数据集描述

数据集	样本数	维数	类目数	来源
Iris	150	4	3	UCI
Bupa	345	7	2	UCI
Wine	178	14	3	UCI
Cmc	1 473	10	3	UCI
DS1	1 000	2	3	人工
DS2	5 000	3	5	人工

把以上 6 组数据集运用于所提算法中,将得到的结果与传统的利用 $\mathrm{int}\sqrt{n}$ 来确定聚类个数的结果进行比较,如表 2 所示。

表 2　比较结果

数据集	样本数	正确类数	$\mathrm{int}\sqrt{n}$	所提算法
Iris	150	3	2 ~ 12	3 ~ 5
Bupa	345	2	2 ~ 18	2 ~ 5
Wine	178	3	2 ~ 13	2 ~ 6
Cmc	1 473	3	2 ~ 38	3 ~ 5
DS1	1 000	3	2 ~ 31	2 ~ 5
DS2	5 000	5	2 ~ 70	3 ~ 7

从以上实验结果可以看出:传统的利用 $\mathrm{int}\sqrt{n}$ 计算聚类个数的算法通常包含正确的聚类个数,但计算出的聚类个数范围较大,相对来说不够精确,同时较大的聚类个数范围也导致在实际的聚类过程中计算量较大,时间复杂度较高。而采用所提出的算法也能得到正确的聚类个数范围,且大大缩小了聚类个数范围,有效降低了聚类过程的时间复杂度,同时将每个局部极大值网格内的数据取平均值作为每个类的初始中心点,可有效避免选取聚类初始中心点的盲目性。

4　结束语

在当前的聚类算法中,聚类个数和聚类初始中心点等信息难以准确给定,导致在实际应用中,用户对于聚类得到的结果无法进行合理选择。为此,提出了基于划分的聚类个数与初始中心的确定方法。该方法通过对数据集进行不同尺度的划分,通过分析数据分布的密度来确定聚类个数与聚类初始中心。通过该方法获得的聚类个数范围较小并可以准确包含最佳聚类个数,有效避免了因较大的搜索范围造成聚类过程中较

(下转第 82 页)

万方数据

4 结束语

为了提高推荐算法的准确度和召回率,根据时间点不同反映用户行为不同,特别引入了时间参数,建立与时间相关的基于物品的推荐系统,实现推荐与其购买过的物品最相似的物品。实验结果表明,该方法极大地提高了推荐系统的准确度。

参考文献:

[1] 许海玲,吴 潇,李晓东,等. 互联网推荐系统比较研究[J]. 软件学报,2009,20(2):350-362.

[2] 刘建国,周 涛,汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展,2009,19(1):1-15.

[3] Adomavicius G, Tuzhilin A. Towards the next generation of recommender systems:a survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge & Data Engineering,2005,17(6):734-749.

[4] Sarwar B,Karypis G,Konstan J. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international world wide web conference on item-based collaborative filtering recommendation. [s. l.]:[s. n.],2001:285-295.

[5] Ding Y, Li X. Time weight collaborative filtering[C]//Proceedings of the 14th ACM international conference on information and knowledge management. [s. l.]:ACM,2005:485-492.

[6] 蒋 凡. 推荐系统[M]. 北京:人民邮电出版社,2013.

[7] Davidson J,Liebald B,Liu J. The YouTube video recommendation system[C]//Proceedings of the fourth ACM conference on recommender systems. [s. l.]:ACM,2010:293-296.

[8] 李 蕊,李仁发. 上下文感知计算及系统框架综述[J]. 计算机研究与发展,2007,44(2):269-276.

[9] 王立才,孟祥武,张玉洁. 上下文感知推荐系统[J]. 软件学报,2012,23(1):1-20.

[10] Wang L C. Understanding and using contextual information in recommender systems[C]//Proceedings of ACM SIGIR. [s. l.]:ACM,2011:1329-1330.

[11] 李 聪. 电子商务推荐系统中协同过滤瓶颈问题研究[D]. 合肥:合肥工业大学,2009.

[12] Zheng H,Wang D,Zhang Q,et al. Do clicks measure recommendation relevancy:an empiricaluser study[C]//Proceedings of the fourth ACM conference on recommender systems. [s. l.]:ACM,2010:249-252.

[13] Kahng M,Lee S. Ranking in context-aware recommender systems[C]//International conference companion on world wide web. New York:ACM Press,2011:65-66.

[14] 曹 毅. 基于内容和协同过滤的混合模式推荐技术研究[D]. 长沙:中南大学,2007.

[15] Conway D,White J. Machine learning for hackers[M]. [s. l.]:O'Reilly Media,Inc.,2012.

[16] 项 亮. 推荐系统实践[M]. 北京:人民邮电出版社,2012.

(上接第 78 页)

大的计算量。同时获取的初始聚类中心点更接近于真实的类中心,不会受到噪声点的干扰。在实际数据和合成数据上的实验结果表明,新方法具有较高的有效性和准确性。

参考文献:

[1] Arora S,Chana I. A survey of clustering techniques for big data analysis[C]//5th international conference on confluence the next generation information technology summit. [s. l.]:[s. n.],2014:59-65.

[2] Soua M,Kachouri R,Akil M. A new hybrid binarization method based on Kmeans[C]//6th international symposium on communications,control and signal processing. [s. l.]:[s. n.],2014:118-123.

[3] 胡 伟. 改进的层次 K 均值聚类算法[J]. 计算机工程与应用,2013,49(2):157-159.

[4] 魏建东,陆建峰,彭甫镕. 一种层次初始的聚类个数自适应的聚类方法研究[J]. 电子设计工程,2015,23(6):5-8.

[5] Zhang Jie,Dong Jianrui,Xiao Yiyong. A new method on finding optimal centers for improving K-means algorithm[C]//27th Chinese control and decision conference. [s. l.]:[s. n.],2015:1827-1832.

[6] Shao Xiongkai,Pi Jing,Liu Jianzhou. A method of dynamically determining the number of clusters and cluster centers[C]//8th international conference on computer science & education. [s. l.]:[s. n.],2013:283-286.

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

[8] 张红云,刘向东,段晓东,等. 数据挖掘中聚类算法比较研究[J]. 计算机应用与软件,2003,20(2):5-6.

[9] 周世兵. 聚类分析中的最佳聚类数确定方法研究及应用[D]. 无锡:江南大学,2011.

[10] 周炜奔,石跃祥. 基于密度的 K-means 聚类中心选取的优化算法[J]. 计算机应用研究,2012,29(5):1726-1728.

[11] 孟令奎,胡春春. 基于模糊划分测度的聚类有效性指标[J]. 计算机工程,2007,33(11):15-17.

[12] 周开乐,杨善林,丁 帅,等. 聚类有效性研究综述[J]. 系统工程理论与实践,2014,34(9):2417-2431.

[13] Frey B J,Dueck D. Clustering by passing messages between data points[J]. Science,2007,315(5814):972-976.

[14] Frey B J,Dueck D. Response to comment on "clustering by passing messages between data points"[J]. Science,2008,319(5864):726.

万方数据