

基于核密度估计的 K -means 聚类优化

熊开玲¹, 彭俊杰¹, 杨晓飞², 黄俊²

(1. 上海大学 计算机工程与科学学院, 上海 200444;

2. 中国科学院 上海高等研究院 公共安全中心, 上海 201210)

摘 要: K -means 聚类算法作为一种经典的聚类算法, 应用领域十分广泛; 但是 K -means 在处理高维及大数据集的情况下性能较差。核密度估计是一种用来估计未知分布密度函数的非参数估计方法, 能够有效地获取数据集的分布情况。抽样是针对大数据集的数据挖掘的常用手段。密度偏差抽样是一种针对简单随机抽样在分布不均匀的数据集下容易丢失重要信息问题的改进方法。提出一种利用核密度估计结果的方法, 选取数据集中密度分布函数极值点附近的样本点作为 K -means 初始中心参数, 并使用核密度估计的分布结果, 对数据集进行密度偏差抽样, 然后对抽样的样本集进行 K -means 聚类。实验结果表明, 使用核密度估计进行初始参数选择和密度偏差抽样能够有效加速 K -means 聚类过程。

关键词: K -means 聚类; 密度偏差抽样; 核密度估计; 数据挖掘

中图分类号: TP305

文献标识码: A

文章编号: 1673-629X(2017)02-0001-05

doi:10.3969/j.issn.1673-629X.2017.02.001

K -means Clustering Optimization Based on Kernel Density Estimation

XIONG Kai-ling¹, PENG Jun-jie¹, YANG Xiao-fei², HUANG Jun²

(1. School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China;

2. Public Security Center, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China)

Abstract: K -means clustering algorithm is classical and widely used in many fields, but it has poor performance in the case of processing high dimensional and large data sets. Kernel density estimation is a nonparametric estimation method to estimate the density function of unknown distribution, which can effectively obtain the distribution of the data set. Sampling is a common method for data mining in large data sets. Density biased sampling is an improved method for the problem of easy loss of important information when using the simple random sampling in the inclined data set. A method is proposed using result of kernel density estimation, which chooses sample points from neighborhood of peak of density function of dataset as the initial center parameters of K -means and uses result of kernel density estimation to perform density biased sampling on the dataset, then runs K -means clustering on the sample set. The experimental results show that using the kernel density estimation for selection of initial parameters and density bias sample can effectively accelerate the K -means clustering process.

Key words: K -means clustering; density bias sampling; kernel density estimation; data mining

0 引 言

随着互联网、物联网等产业的发展, 各种各样包含高维和海量的大规模数据集被生成。针对大规模数据的数据分析也变得越来越普遍^[1]。 K -means 聚类算法作为一种应用广泛的经典聚类算法, 在面对大规模结构复杂的数据时, 与其他数据挖掘方法一样, 表现得不太理想, 主要集中在面对大数据时计算开销和时间开销成倍的增长和选择初始参数时变得极为困难两个问

题上^[2]。针对这样的情况, 通常应用特定的数据挖掘方法时(如聚类、关联规则等), 往往引入抽样技术先从数据集抽取出一个样本, 然后在这个样本数据集上进行处理, 最后将处理结果推广到整个数据集^[3]。简单随机抽样(Simple Random Sampling, SRS)是一种简单高效的抽样方法。SRS 应用广泛但在不均匀或者倾斜严重的数据集中效果得不到保证^[4]。由于许多自然现象都遵循如 Zipf 定律、二八定律等不均匀分布,

收稿日期: 2016-03-28

修回日期: 2016-07-05

网络出版时间: 2017-01-04

基金项目: 国家自然科学基金资助项目(61201446)

作者简介: 熊开玲(1992-), 男, 硕士研究生, 研究方向为数据挖掘; 彭俊杰, 博士, 副教授, 硕士生导师, CCF 高级会员, 研究方向为云计算。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20170104.1039.074.html>

因此简单随机抽样在许多数据集中并不适用。为此 Palmer 等提出了密度偏差抽样方法 (Density Bias Sampling, DBS)^[5-6]。该方法被证明在分布不均匀的数据集中有较好的适应性,数据简约性能较好^[7],但 DBS 需要预知数据集的分布。核密度估计 (Kernel Density Estimation, KDE) 是一种非参数估计方法,不需要有关数据分布的先验知识,是一种获取数据集未知分布的有效方法。因此可以使用核密度估计解决密度偏差抽样需要知道数据分布的问题。此外,针对数据集较稠密区域更容易成为类簇中心的特点^[8],使用核密度估计的结果选择 K -means 的初始中心参数,并在大数据集时使用核密度估计的结果进行密度偏差抽样,然后使用抽样样本集进行聚类分析。实验结果表明,该方法可以在不影响聚类结果的情况下有效减少聚类过程的时间。

1 K -means 聚类算法

K -means 聚类算法解决的问题是将含有 n 个数据点的集合 $X = \{x_1, x_2, \dots, x_n\}$ 划分为 k 个类簇 C_1, C_2, \dots, C_k , 使所有的数据点到其所在类中心的距离和最小,即 $\arg\min \sum_{i=1}^k \sum_{x \in C_i} \|x - c_i\|^2$ 。其中, c_i 是 C_i 的中心, $\arg\min \sum_{i=1}^k \sum_{x \in C_i} \|x - c_i\|^2$ 称之为聚类准则函数或目标函数。经典的 K -means 算法首先随机选取 k 个数据点作为 k 个类的初始类中心,然后遍历数据中所有的点,将它们分配到距离最近的中心点所在的类中,并重新计算各个类的中心点;使用新的中心点重复迭代上述过程,直到所有数据点到其所在类的中心点的距离和达到最小值,这时聚类过程结束^[9]。

K -means 算法流程如下:

Step1: 随机指定 k 个点作为初始化类中心点;

Step2: 对每一个样本 x , 将其分配到离它最近的聚类中心;

Step3: 重新计算各类中心;

Step4: 使用新的类中心, 进行 Step2 并计算偏差 J ;

Step5: 如果 J 值收敛, 则返回类中心 C 并终止算法, 否则回到 Step2。

作为一种基于划分的聚类算法, 由于算法简单, 容易理解, 所以 K -means 算法应用广泛。但是, K -means 聚类算法也有自身的局限性。主要缺陷表现如下^[10]:

(1) K -means 聚类算法需要预先给出簇的数目 K 。而在实际应用中, 聚类数据集究竟需要分成多少类往往是未知的。

(2) K -means 聚类算法的聚类结果受初始中心点选取的影响较大, 初始中心点选取不当很容易造成聚类结果陷入局部最优解甚至导致错误的聚类结果。

(3) K -means 聚类算法不适用于大数据集的聚类问题。 K -means 聚类算法迭代过程中每次都需对所有的数据点进行计算, 因此面对大数据集时算法的计算开销巨大。

2 密度偏差抽样

随着互联网、物联网、大数据等领域的发展, 大规模数据集越来越普遍, 这些数据集的数据量对 K -means 算法所带来的计算复杂度可以说是灾难级的。由于 K -means 聚类算法在处理大数据集时的效果不能令人满意, 并且该算法的计算开销较大。另一方面, 由于大规模数据集本身的复杂度较高, 因此对给定的大数据集进行数据挖掘时, 通常都是先抽取一个样本数据集, 然后在该样本数据集上进行处理, 最后将样本集处理的结果推广到整个数据集。

简单随机抽样作为所有抽样方法的基础是目前应用最广泛的抽样方法, 该方法虽然原理简单但效率较高。不过在不均匀或者倾斜严重的数据集中效果得不到保证。Palmer 等提出的密度偏差抽样方法被证明在分布不均匀的数据集中有较好的适应性, 数据简约性能较好。

密度偏差抽样是一种相对较新的抽样策略。其主要流程是, 先将原始数据集分成不同的组, 各组大小 (所含数据点的数量) 表示该组的密度, 然后按以下原则进行抽样^[11]:

(1) 一个分组内各个数据点被抽到的概率相同;

(2) 最终获取的数据样本集的分布特征与原始数据集的分布一致;

(3) 各组抽样概率的偏差依据各组大小的偏差;

(4) 期望的样本集的大小值已知。

综合上述原则, 当数据集的分布为均匀分布时, 各个分组的密度也应该是一致的, 这时密度偏差抽样的结果与简单随机抽样的结果应该是一样的, 因此, 简单随机抽样可视为数据集在均匀分布密度偏差抽样的特例。相对于简单随机抽样, 密度偏差抽样的优势主要是简约效果好、适应性强^[12]。

图 1 中的原始数据是有 20 000 个数据点 (x, y) 的数据集, 其中 4 个组都使用高斯分布生成。当对其进行 2% 的抽样时, 发现进行 DBS 抽样两个较小的数据集能够保留在抽样结果中, 但是使用 SRS 抽样时, 较小的数据集完全消失了。通过结果对比可知, 当数据集属于不均匀分布时, 使用密度偏差抽样的结果较使用简单随机抽样时的结果更优。因此为了更好地保证

抽样所得的样本集能够反映整个数据集的特征,需要知道数据集的密度分布。

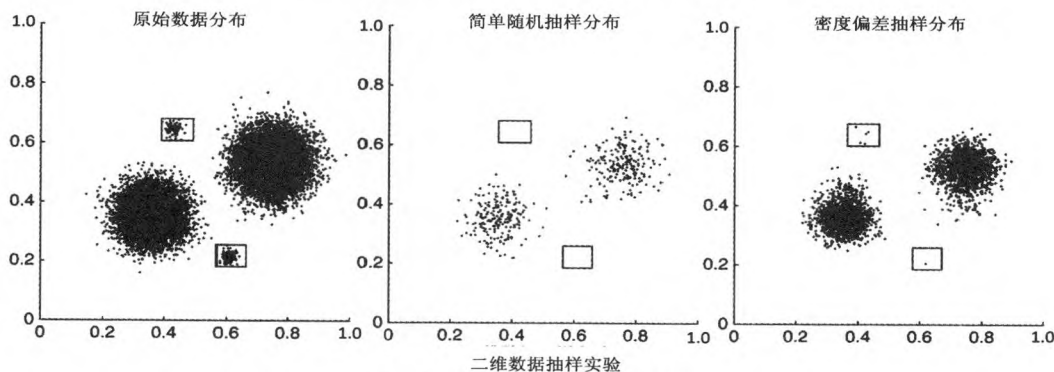


图1 DBS与SRS的实验对比

3 核密度估计

核密度估计是求解给定的随机变量集合分布密度问题的非参数估计方法之一,用来解决与之相对的参数估计方法所获得结果性能较差的缺陷问题。在参数估计方法中需要先对数据集做相关假定数据集的分布,然后求解假定函数中的未知参数。但是通常假定的密度函数模型与实际密度函数之间相差较大。由于核密度估计不需要对数据集做相关假设,只是从数据集本身出发研究数据集的分布特征,所以 KDE 自提出以来就在各个领域广泛应用。

核密度估计定义如下:

假设 x_1, x_2, \dots, x_n 为取值于 R 的独立分布随机变量,其所服从的分布密度函数为 $f(x)$, $x \in R$ 。定义函数:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right), x \in R \quad (1)$$

其中, $\hat{f}_h(x)$ 称为密度函数 $f(x)$ 的核密度估计; $K(\cdot)$ 称为核函数; h 通常称为窗宽或光滑参数,是一个预先给定的正数。

令 $K_h(u) = K(u/h)/h$, 则式(1)可以表示为:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K_h(x_i - x), x \in R \quad (2)$$

由上述定义可知,分布密度函数 f 的核密度估计 \hat{f} 不仅与给定的数据样本集有关,还与核函数的选择和窗宽参数 h 的选择有关^[13]。

理论上,任何函数均可以做核函数,但是为了密度函数估计的方便性与合理性,通常要求核函数满足以下条件^[14]:

$$K(-u) = K(u) \quad (3)$$

$$\sup |K(u)| < \infty, \int_{-\infty}^{+\infty} K(u) du = 1$$

常用的核函数有:

(1) 高斯核函数 (Gaussian kernel):

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

(2) 矩形窗核函数 (boxcar):

$$K(u) = \begin{cases} 1, & |u| \leq 0.5 \\ 0, & |u| > 0.5 \end{cases}$$

(3) Epanechnikov 核函数:

$$K(u) = \begin{cases} \frac{3(1-u^2)}{4}, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

(4) BiWeight 核函数:

$$K(u) = \begin{cases} \frac{15(1-u^2)^2}{16}, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

另外,窗宽参数 h 控制了求点 h 处的近似密度时不同距离样本点对点密度的影响程度。当 h 选择过大时会出现过光滑情况 (OverSmoothed), 当 h 选择过小时会出现不够光滑 (UnderSmoothed) 的情况^[15]。

如图2所示,使用正太分布随机产生100个随机数,虚线是其实际概率密度函数曲线,使用不同窗宽 h 进行核密度估计时,得出如图所示结果。当 $h = 0.05$ 时的 KDE 密度函数曲线波动频繁,当 $h = 2$ 时,曲线较为光滑但与实际相差甚远。在实际概率密度曲线与 $h = 2$ 之间的是 $h = 0.337$ 时的 KDE 密度函数曲线,该

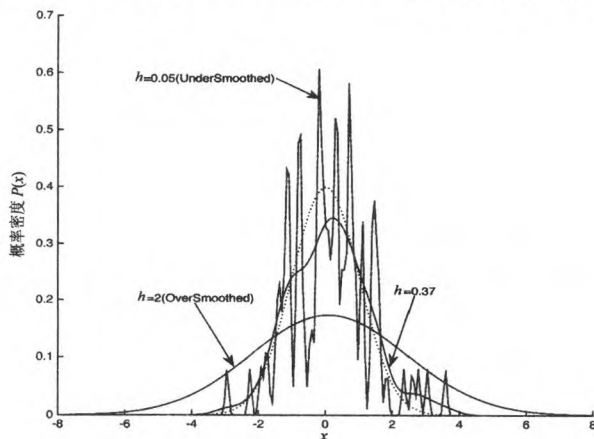


图2 窗宽 h 值的选择对核密度估计的影响

曲线与实际曲线十分接近。

通常在实际应用中,需要分析的数据集是多维的,因此需要分析多维空间中的核密度估计问题。假设 $x_1, x_2, \dots, x_n \in R^d, x_i = [v_1, v_2, \dots, v_d]$, 其中 x_i 是一个 d 维向量。为了方便起见,认为 d 维空间 R^d 的各个维度间相互独立。设 $f_j(v), j = 1, 2, \dots, d$ 是第 j 维的概率密度函数。那么数据的多维密度函数可表示为 $f_d(x) = \prod_{j=1}^d f_j(v)$ 。特别地,采用 d 维同分布的高斯核函数构造的 R^d 上的核密度函数可以表示为:

$$K_{d,\sigma}(x) = (\sqrt{2\pi}\sigma)^{-d} \prod_{j=1}^d e^{-\frac{v_j - x_j}{2\sigma^2}} \tag{4}$$

当对图 1 中的原始数据集进行高斯核估计时,可以得到如图 3 所示的密度分布。

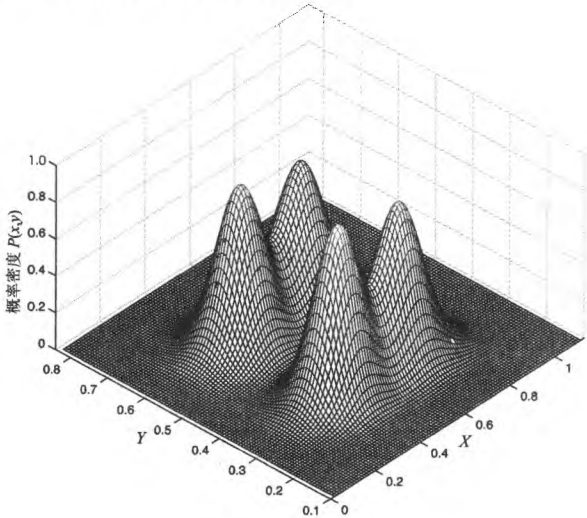


图 3 核密度估计

4 基于核密度估计的 K-means 聚类优化

由图 3 显示,原始数据集中数据点越稠密的地方密度函数值越大,而密度函数的极大值点也与原始数据集的稠密区域中心十分接近。因此为了减少 K-means 聚类的迭代次数,可以选择核密度函数的极大值点作为 K-means 初始迭代中心。另外,可以通过设置半径阈值,将距离小于半径阈值的极大值点归并到一个类;设置密度阈值,将密度小于密度阈值的极大值区域作为噪声去除,因为这些区域在整个数据集所占的权重过低。通过半径阈值和密度阈值的过滤,就可以确定一个聚类数目 K 。当数据集较大时,可以使用核密度估计函数对数据集进行密度偏差抽样,以此来约简数据集。

步骤如下:

- (1)对数据集进行高斯核密度估计获取数据集密度分布;
- (2)根据密度分布设定半径阈值和密度阈值;

- (3)根据半径阈值和密度阈值获取 K-means 算法的参数 K 和初始聚类中心;
- (4)对数据集进行密度偏差抽样获取样本集;
- (5)对样本集进行 K-means 聚类。

实验 1 使用来自 UCI KDD Archive 中的 Synthetic Control Chart Time Series 的数据集(600 条、维度为 60 维),数据样本较小,但维度较高。使用核密度估计选择的初始聚类中心和使用随机选择的初始聚类中心的 K-means,在未使用密度偏差抽样的情况下进行多次实验。表 1 是两种情况下迭代次数对比以及样本点的类内总距离的对比,其结果表明,在聚类效果差不多的情况下,核密度估计初始参数选择的迭代次数较随机初始参数更少。

表 1 利用核密度估计选择迭代初始参数与随机初始参数聚类对比

随机初始参数		核密度估计初始参数	
迭代次数	类内总距离	迭代次数	类内总距离
8	953 657	6	953 657
8	1.04E+06	6	1.04E+06
10	1.04E+06	7	1.03E+06
14	953 647	9	953 878
8	951 537	8	944 191
10	1.04E+06	6	1.03E+06

对图 1 所示的数据集,进行 K-means 和使用了核密度估计优化的 K-means 聚类分析。实验结果如表 2 所示。

表 2 使用了核密度估计初始参数选择和密度偏差抽样的 K-means 结果对比

迭代次数		类中心
无核密度估计	核密度估计	距离偏差
71	19	0.021 47
54	20	0.019 83
74	21	0.017 44
40	21	0.032 01
68	22	0.020 59
18	21	0.321 22
40	19	0.016 47

由于两次聚类结果一定会有差异,所以无法将两次结果中的类一一对应,因此为了方便对比实验结果,引入如下定义:

假设聚类结果为 k 个类簇,其类簇中心分别为 C_1, C_2, \dots, C_k , 那么其中心距离总和为 $Sum_c = \sum_{i=1}^{k-1} dist(C_i, C_{i+1}) + dist(C_1, C_k)$ 。其中, $dist(x, y)$ 表示 x, y 之间的距离。那么同一组实验内,直接采用 K-means 聚类

和使用核密度估计优化的K-means聚类的结果可以使用中心距离总和偏差 $|\text{Sum}_{cK\text{-means}} - \text{Sum}_{cKDE}| / \text{Sum}_{cK\text{-means}}$ 来验证使用核密度估计优化聚类的准确性。

表2中的结果表明,使用核密度估计优化后的K-means迭代次数较少,结果与直接使用K-means的相差不大,并且迭代次数较为稳定,相比随机参数初始化的K-means受初始参数的影响较小。如第6组实验中,使用随机参数初始化的K-means就陷入了局部最优解,因此结果差距较大。

5 结束语

提出了一种利用核密度估计结果的方法。通过实验结果表明,使用核函数密度估计所选取的K-means初始参数,可以在尽量不影响聚类效果的基础上有效减少K-means的迭代次数,同时在数据集较大时,可以使用核密度估计的结果对数据集进行密度偏差抽样,有效地简约数据量,从而加速聚类过程。

参考文献:

- [1] 程学旗,靳小龙,王元卓,等. 大数据系统和分析技术综述[J]. 软件学报,2014,25(9):1889-1908.
- [2] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61.
- [3] 张春阳,周继恩,钱权,等. 抽样在数据挖掘中的应用研究[J]. 计算机科学,2004,31(2):126-128.
- [4] 盛开元,钱雪忠,吴秦. 基于可变网格划分的密度偏差抽样算法[J]. 计算机应用,2013,33(9):2419-2422.
- [5] Palmer C R, Faloutsos C. Density biased sampling: an improved method for data mining and clustering[J]. ACM SIGMOD Record,2000,29(2):82-92.
- [6] Hoti F. On estimation of a probability density function and the mode[J]. Annals of Mathematical Statistics, 2003, 33(3): 1065-1076.
- [7] 李存华,孙志挥,陈耿,等. 核密度估计及其在聚类算法构造中的应用[J]. 计算机研究与发展,2004,41(10):1712-1719.
- [8] 倪巍巍,陈耿,吴英杰,等. 一种基于局部密度的分布式聚类挖掘算法[J]. 软件学报,2008,19(9):2339-2348.
- [9] MacQueen J B. Some methods for classification and analysis of multivariate observations[C]//Proceedings of 5th Berkeley symposium on mathematical statistics and probability. California: University of California Press, 1967:281-297.
- [10] Rosenblatt M. Remarks on some nonparametric estimates of a density function[J]. The Annals of Mathematical Statistics, 1956,27:832-837.
- [11] 余波,朱东华,刘嵩,等. 密度偏差抽样技术在聚类算法中的应用研究[J]. 计算机科学,2009,36(2):207-209.
- [12] Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset[J]. Genome Biology,2002,3(7):1-21.
- [13] Jones M C, Marron J S, Sheather S J. A brief survey of bandwidth selection for density estimation[J]. Journal of the American Statistical Association,1996,91(433):401-407.
- [14] Buch-Larsen T, Nielsen J P, Guillén M. Kernel density estimation for heavy-tailed distributions using the champemowne transformation[J]. Statistics,2005,39(6):503-518.
- [15] Park B U, Marron J S. Comparison of data-driven bandwidth selectors[J]. Journal of the American Statistical Association, 1990,85(409):66-72.