基于密度与最小距离的 *K*-means 算法初始中心方法

戚后林,顾 磊
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:为了克服在传统 *K* -means 聚类算法过程中因初始类簇中心的随机性指定所带来的聚类结果波动较大的缺陷,提出了一种基于密度与最小距离作为参数来确定初始类簇中心的算法。该算法根据一定的规则计算数据对象的密度参数,在计算完数据集中每条数据的单点密度之后,计算每个数据对象与较其密度大的其他数据对象的最小距离,以密度和最小距离作为参数,选取密度和最小距离同时较大的点作为 *K* -means 聚类过程的初始类簇中心。实验结果表明,在类簇数目确定的情况下,应用该算法确定的初始 *K* -means 类簇中心,在标准的 UCI 数据集上能够进行 *K* -means 聚类,且与随机选择类簇中心和其他使用密度作为参数的算法相比,基于改进后的初始中心方法的 *K* -means 聚类算法具有较高的准确率和更快的收敛速度。

关键词: *K* -means 算法;类簇中心;密度;最小距离;迭代次数
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2017)09-0060-04
doi:10.3969/j.issn.1673-629X.2017.09.013

An Initial Center Algorithm of *K*-means Based on Density and Minimum Distance

QI Hou-lin, GU Lei
(College of Computer, Nanjing University of Posts and Telecommunication, Nanjing 210003, China)

Abstract: In order to overcome a large fluctuation caused by the traditional *K* -means algorithm clustering with assignment of the random initial cluster centers, an algorithm taken the density and minimum distance as the parameters to determine the initial cluster centers is proposed, which calculates the density parameter of the data object according to certain rules and minimum distance between each data object and other data objects after having calculated single point density of each data in the data set. The larger one among the densities and minimum distances has been chosen as initial cluster center in the process of *K* -means clustering. Experimental results show that it has higher accuracy and faster convergence rate compared with ones using randomly selected cluster centers and using density as a parameter for *K* -means clustering on standard UCI data set.

Key words: *K* -means algorithm; cluster center; density; minimum distance; iteration number

1 概 述

近年来,随着大数据的兴起,如何从中总结出有价值的数据规律是一个重要任务。聚类作为一种数据分析法,在数据挖掘、图像处理等方面都有重要应用。聚类算法包括基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法和基于模型的方法。聚类分析的目的是数据集合应用不同的策略划分成相似的类簇的过程,从而使同一个类簇具有较高的相似度,而不同的类簇之间尽可能不同。同时聚类分析作为一种数据划分的方法,也可以作为其他数据挖掘方法的一个预处理步骤。

K -means 算法[1]是 MacQueen 提出的一种聚类方法。作为一种典型的基于划分的聚类算法,其特点为:几何意义直观,收敛速度快,资源消耗较少等。但缺点也显而易见:由于算法的初始点通常在算法开始时随机给出,算法的结果很不稳定;同时,算法对于初始类簇中心较为敏感,容易陷入局部最优,类簇中心的数目需要事先给定。

假设在 n 个数据点中找到 k 个聚类中心 $c_1, c_2, \cdots, c_k$,使得每个数据 x_i 与其最近的聚类中心 c_v 的平方距

收稿日期:2016-09-07 **修回日期**:2016-12-22 **网络出版时间**:2017-07-11
基金项目:国家自然科学基金资助项目(61302157)
作者简介:戚后林(1990-),男,硕士研究生,研究方向为中文文本分类;顾 磊,副教授,硕士生导师,研究方向为中文信息处理、机器学习。

离和最小化(该距离和称为偏差 Δ),也即 Δ 收敛。

输入:类簇的数目 k 以及 n 个记录的数据集;

输出:k 个类簇,Δ 最小或者收敛。

(1)初始化。指定 k 个类簇中心 $c_1,c_2,\cdots,c_k$。

(2)分配 x_i。找到距离 x_i 最近的类簇中心 c_v,并将其分配到其最近的类簇中心 c_v。

(3)修正 c_v。通过不断地计算簇的平均值,找到更加合适的聚类中心 c_v。

(4)计算偏差:

$$\Delta = \sum_{i=1}^{n} [\min_{r=1,\cdots,k} d(x_i,c_r)^2] \quad (1)$$

(5)Δ 收敛。如果收敛,返回 $c_1,c_2,\cdots,c_k$,并终止算法,否则返回步骤(2)。

为了克服 K-means 算法的缺陷,学者们试图从不同角度去改进 K-means 算法。文献[2]研究了确定 K 的算法,该算法假设数据集的子集服从高斯分布,然后在 K-means 聚类过程中不断增加 K 的大小,直到数据集满足假设,但是对于不服从高斯分布的数据集,该算法仍然存在缺陷。文献[3]通过 K-d 树来划分数据集,然后利用多个局部区域密度选择初始中心的方法,该算法依赖树节点的数量,当数据集的维度较大时,该算法的运行时间较长。文献[4]提出一种基于平均密度优化初始类簇中心的算法(adk-means),该算法首先找到数据集中的噪点,在算法执行过程中,噪点不参与聚类过程,但是该算法需要人为指出噪点,当数据集较大时,算法的主观性较强。文献[5]提出一种基于密度的算法,该算法通过缩小维度来加快初始化的过程,不断把可能的类簇中心移向高密度点,直到得到最大的 K 个最高密度点作为初始的类簇中心,但是该算法的运行迭代次数太高,运行时间较长。文献[6]提出利用二分搜索方法来寻找最佳的 K 个初始类簇中心。文献[7]提出一种基于密度的网格算法,该算法在 Map-Reduce 框架中确定 K 值的大小以及噪点的位置。文献[8]提出了密度峰值进行快速搜索的聚类方法,算法假设类簇中心主要有两个特征,一是具有较高的密度,二是距离比其密度大的类簇中心的最小距离较大。在计算出密度与最小距离决策图时,可以很直观地看到类簇中心和噪点,与文献[4]一样,该算法同样需要人为指出数据集中的噪点,同时,对于截断距离也有很大的主观性。文献[9]提出一种基于最小方差优化初始中心的算法,避免了重复计算数据集到类簇中心的距离,减少了迭代次数,缩短了聚类时间。文献[10]提出一种基于密度与特定阈值的改进 K-means 算法,该算法首先计算数据集中每个点的密度,然后利用阈值不断迭代较小的密度点加入到类簇中心集合,直到集合中的元素个数到达 K 为止。文献[11]提出一种基于数据集平均密度与最小距离的聚类算法,该算法计算密度较小的点距离较大点的最小距离,然后将最小距离与平均密度做乘积,由此来计算出离群点,反复迭代,不断剔除离群点,直到剩下 K 个点来作为初始的类簇中心。文献[12]为了减少 K-means 算法在数据量较大时运行时间较长的问题,提出在 MapReduce 平台下运行多路 K-means(Mux-Kmeans)算法。文献[13]提出一种基于 K-means 的聚类算法,该算法包括两个过程:利用 K-means 算法分割较为稀疏的子数据集,然后利用平均距离对已经分割好的子数据集进行合并。文献[14]提出一种基于密度与最佳距离的 K-means 初始中心选择方法,该算法利用配置函数的最优值选择初始的类簇中心,同时也可以减小迭代次数。文献[15]为了解决高维数据集聚类,提出了基于 K-means 算法的 K-String 算法。

为了解决传统 K-means 算法在聚类过程中初始中心随机选择的缺陷,在研究基于密度与最小距离的 K-means 初始中心选择算法原理和步骤分析的基础上,提出了改进的 K-means 初始中心算法,并进行了验证实验和对比分析。

2　基于密度与最小距离的 K-means 算法聚类

传统的 K-means 算法对于初始中心的选取比较敏感,因此,随机性地选取初始中心可能会影响聚类结果的速度和稳定性。而基于密度选取的初始中心则具有主观性,例如,当数据集中两个数据点的密度一样时,如何取舍会直接影响到聚类的结果。提出的算法综合数据集中数据点的密度参数和最小距离参数,引入其他参数作为初始类簇中心的选取指标。

2.1　基本定义

设数据集中含有 M 个样本数据,每个样本有 N 个属性,则任意数据可以表示为 $x=(x_1,x_2,\cdots,x_m)$。

定义 1:数据点 x 与 y 之间的距离为欧氏距离 $d(i,j)$:

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

定义 2:截断距离 d_c,定义为整个数据集中点与点之间距离的平均值:

$$d_c=\frac{1}{m^2}\times\sum d(i,j) \quad (3)$$

定义 3:样本数据点 p_i 的密度 ρ_i。表示以 p_i 为圆心,d_c 为半径的圆,包含在圆内其他数据点的数量之和。ρ_i 越大,p_i 成为初始中心的可能性越大。

$$\rho_i=\sum_j \aleph(d(i,j)-d_c) \quad (4)$$

其中，$\aleph(x)=\begin{cases}1,x\geqslant 0\\0,x\leqslant 0\end{cases}$。

定义 4：点 p_i 到其他更高密度点之间的最小距离为 δ_i，代表数据点 p_i 与 p_j 的不相似度，δ_i 越大，说明该点距离其他较大的类簇中心越远，即该点成为类簇中心的可能性较大。

$$\delta_i=\min(d(i,j)),j:\rho_j>\rho_i \tag{5}$$

对于密度 ρ_i 最大的点 p_i，其最小距离定义为：$\delta_i=\min_j(d(i,j))$。

定义 5：最终决定样本点 p_i 能否成为类簇中心的决定性参数 θ_i，综合点 p_i 的密度与最小距离，其值直接决定此点是否能成为类簇中心。θ_i 值越大，说明此点在拥有较大密度的同时，距离其他较高的类簇中心也较远，即此点成为初始中心的可能性较大：

$$\theta_i=\rho_i\times\delta_i \tag{6}$$

定义 6：假设在 K-means 算法迭代数次后，同一类簇中有且只有两点的坐标分别为 $(a_1,b_1,c_1,\cdots,n_1)$ 与 $(a_2,b_2,c_2,\cdots,n_2)$，则新的聚类中心的坐标为：$a=\frac{a_1+a_2}{2},b=\frac{b_1+b_2}{2},c=\frac{c_1+c_2}{2},\cdots,n=\frac{n_1+n_2}{2}$。

2.2 算法描述

在 K-means 算法中，数据集中点的距离可以使用欧氏距离来衡量，选取数据集的平均距离 d_c 作为截断距离，有利于算法收敛。两点之间的距离越小，说明两点越相似，但是在初始类簇数目一定的情况下，选取一点作为初始的类簇中心关系到聚类结果的准确性和迭代次数。假设类簇中心被其他较低密度的中心所包围，在计算出每个点 p_i 的密度 ρ_i 之后，同时计算 p_i 距离较高密度类簇中心的最小距离 θ_i，最后计算出 p_i 点的参数 θ_i。算法的详细描述如下：

(1)根据定义 1 计算出数据集中任意两点 p_i 与 p_j 的距离 $d(i,j)$，并根据定义 2 计算出数据集最大平均距离作为截断距离 d_c。

(2)通过定义 2 及定义 3 计算出数据集中每个点的密度 ρ_i 及距较高密度类簇中心的最小距离 δ_i。对于密度最大的点，δ_i 定义为其他点到此点的最大距离。

(3)根据定义 5 计算出 θ_i。

(4)选取 K 个具有较大 θ_i 的点作为 K-means 算法的初始类簇中心。

(5)利用 K-means 算法进行聚类。

如图 1 所示，假设数据集最终被划分为两个类簇。截断距离 d_c 由数据集的平均距离确定后，以任意一点 p_i 为中心，d_c 为半径的圆所包含其他数据点的个数为 p_i 点的密度 ρ_i。图中，密度较大的三点 p_1,p_2,p_3 的密度分别为 $\rho_1=4,\rho_2=5,\rho_3=6$。在计算出每个点的密度后，可以看出 p_1 距离密度较大的点为 p_2,p_3，但是与 p_2 的距离最小。密度较 p_2 较大的点只有 p_3，由于 p_3 是整个数据集中密度最大的点，所以在计算 p_3 点的最小距离时，只需在数据集中计算距离 p_3 最远的点。计算出 p_1,p_2,p_3 距离其他较高密度点的最小距离分别为 $\delta_1,\delta_2,\delta_3$。选取 $\theta_i=\rho_i\times\delta_i$ 较大者作为初始类簇中心，在图 1 的二维数据集中，假设要选取两个类簇中心，由于 p_2,p_3 的 θ_i 较大，故选取 p_2,p_3 为初始的类簇中心。

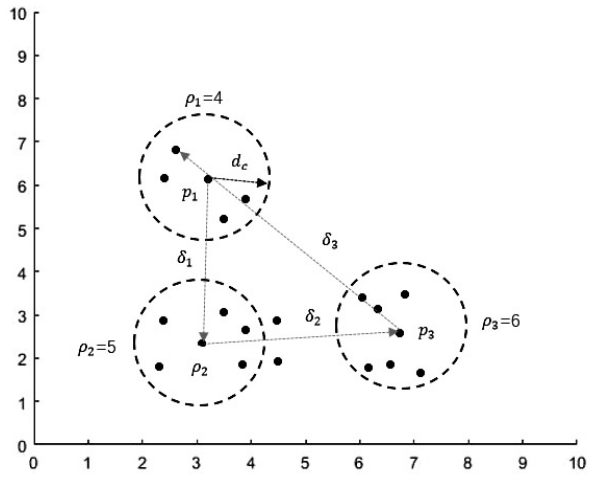

图 1 二维数据的密度和最小距离表示

3 实 验

3.1 数据集描述

为了验证上述算法选取初始类簇中心的有效性，选取了专用于测试聚类算法性能的 UCI 数据集。UCI 是一个专门用于测试机器学习与数据挖掘算法的数据库，库中的每个数据集都有明确的分类，因此可以直观表示聚类结果的质量。对 Iris，Wines，Seeds，Banlance-scale 四个数据集进行了测试，按照准确率、迭代次数来评价算法的性能。表 1 简单介绍了实验所用数据集的情况。

表 1 UCI 数据集描述

数据集	样本个数	数据属性个数	类数
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Balance-scale	625	5	3

3.2 实验结果及分析

用到的 K-means 初始中心选取算法不仅计算数据集中每个数据点的密度值 ρ_i，还将计算距离密度较高的点的最小距离 δ_i，然后将其与密度做乘积得到 θ_i。在四个数据集上随机选取 10 个点作为初始的类簇中心，对提出算法与其他初始中心选取算法进行了实验。表 2 为在 Iris 与 Wine 数据集上的实验结果，表 3 为在 Seeds 与 Balance-scale 数据集上的实验结果。

表 2　Iris 与 Wine 数据集的实验结果

算法	Iris		Wine	
	迭代次数	精度	迭代次数	准确率
随机十次初始类簇中心	9	0.733 3	6	0.7
	6	0.8	7	0.718 7
	8	0.666 6	5	0.68
	7	0.866 6	8	0.685 7
	7	0.68	5	0.709 5
	4	0.82	6	0.691
	5	0.806 6	7	0.696 6
	6	0.773 3	5	0.702 2
	5	0.82	6	0,719 1
	8	0.76	8	0.685 3
平均值	6.5	0.772 64	6.3	0.696 555 6
文献[10]算法	7	0.893 3	5	0.707 8
提出算法	5	0.906 6	6	0.747 1

表 3　Seeds 与 Balance-scale 数据集的实验结果

算法	Seeds		Balance-scale	
	迭代次数	精度	迭代次数	准确率
随机十次初始类簇中心	3	0.852 3	15	0.569 6
	10	0.666 6	12	0.579 2
	6	0.790 4	3	0.609 6
	8	0.671 4	12	0.580 8
	8	0.785 7	6	0.619 2
	7	0.7	4	0.617 6
	12	0.690 4	12	0.56
	4	0.828 5	7	0.563 2
	10	0.719	8	0.6
	11	0.723 8	10	0.620 8
平均值	7.9	0.742 81	8.9	0.592
文献[10]算法	6	0.833 3	12	0.550 4
提出算法	5	0.885 7	7	0.67

为了验证提出算法较其他初始中心选取算法具有较好的性能,分别选取了传统的初始中心算法与文献[10]算法进行对比。

在实验过程中发现,三种算法的类簇划分个数相同,但实验结果却不同。传统的随机化初始中心算法无论是准确率还是迭代次数都不稳定,无法计算出 θ_i 较大者作为初始的类簇中心。

Iris 数据集中,传统初始中心选择算法的准确率和迭代次数都波动较大,十次实验的平均准确率为 0.772 6,平均迭代次数为 6.5,文献[10]算法选择的初始中心准确率为 0.893 3,在迭代 7 次后算法收敛。在利用提出算法选取的类簇中心实验中,θ_i 较大的三个点为(50,16,123),对应的 θ_i 值为(950,930,1 276),算法的迭代次数为 5,准确率为 0.906 6。

Wine 数据集中,传统初始中心选择算法的准确率为 0.696 5,平均迭代次数为 6.3,而文献[10]算法选择初始中心的准确率为 0.707 8,迭代次数为 5。利用提出算法计算出的 θ_i 较大的点为(132,32,78),对应的 θ_i 值为(8 320,10 952,12 300),算法的准确率为 0.747 1,迭代次数为 6。

Seeds 数据集中,传统初始中心算法得到的平均准确率为 0.742 8,平均迭代次数为 7.9,文献[10]算法的平均准确率为 0.833 3,迭代次数为 6。而在提出的初始中心选取算法中,计算出 θ_i 较大的点为(22,184,53),对应的 θ_i 值分别为(693,396,440)。在选取了相应类簇中心后,利用 K-means 算法得到的聚类结果准确率为 0.885 7,迭代次数为 5。

Balance-scale 数据集是维度较大的数据集,传统初始中心选择算法的平均准确率为 0.592,平均迭代次数为 8.9,文献[10]算法得到的准确率为 0.550 4,迭代次数为 12。利用提出算法得到 θ_i 较大的三个点为(122,262,476),对应的 θ_i 值为(680,710,578),算法在迭代 7 次后收敛,聚类准确率为 0.67,可见初始中心选择算法对高维数据同样也有较高的准确率与迭代速度。

4　结束语

在分析传统的 K-means 初始中心选取算法和改进初始中心选取算法不足的基础上,提出了一种基于密度与最小距离优化的初始中心选择算法。该算法根据数据集中每个数据对象的密度与最小距离来确定数据集中类簇的位置。实验结果表明,该算法在消除 K-means 算法对于初始中心依赖性的同时,减小了迭代次数,提高了运行效率。

参考文献:

[1] MacQueen J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. [s. l.]:[s. n.],1967:281-297.

[2] Hamerly G,Elkan C. Learning the K in K-Means[C]//Advances in neural information processing systems. [s. l.]:[s. n.],2004.

[3] Zhang Xuanyi,Shen Qiang,Gao Haiyang,et al. A density-based method for initializing the k-means clustering algorithm [C]//International conference on network and computational intelligence. [s. l.]:[s. n.],2012:46-53.

[4] 邢长征,谷　浩. 基于平均密度优化初始聚类中心的 k-means 算法[J]. 计算机工程与应用,2014,50(20):135-138.

[5] Qiao J,Lu Y. A new algorithm for choosing initial cluster cen-

(下转第 69 页)

5　结束语

智能设备与移动互联网的发展普及,有效地促进了移动用户对智能设备功能的多方面需求,促使其生活方式发生诸多变化。智能设备的应用催生了基于用户兴趣的位置服务推荐系统。为此,围绕时间因素,提出了基于用户历史评分和兴趣消费情景的用户服务推荐系统。而作为推荐系统的算法,其效果多依赖于用户主观因素,仍缺乏其实际推荐效果的考察。为此,对提出算法进行了实验验证。实验结果表明,该算法能够有效地向移动用户提供个性化推荐服务。

移动互联网时代的到来,信息资源的获取和推送可以以任何方式发生在任何时间和任何地点,移动位置推荐系统利用移动环境在信息推荐方面的优势,克服其不利因素,基于用户偏好来预测和过滤不相关的信息,为移动用户提供个性化服务,为解决“移动信息过载”提供了一种有效方式。

参考文献:

[1] Jannach D,Zanker M,Felfernig A,et al. 推荐系统[M]. 蒋凡,译. 北京:人民邮电出版社,2013.

[2] 盛　珍. 基于 Android 平台的 LBS 应用系统开发技术研究[D]. 昆明:云南大学,2012.

[3] Razmerita L. An ontology-based framework for modeling user behavior-a case study in knowledge management[J]. IEEE Transaction on Systems and Humans,2011,41(4):772-783.

[4] 袁　柳,张龙波. 个性化检索中的用户特征模型研究[J]. 计算机工程与应用,2011,47(15):19-24.

[5] Wetzker R,Zimmermann C,Bauckhage C,et al. I tag,you tag:translating tags for advanced user model[C]//Proceedings of the third ACM international conference on web search and data mining. [s. l.]:ACM,2010:71-80.

[6] 任　磊. 推荐系统关键技术研究[D]. 上海:华东师范大学,2012.

[7] 孟祥武,王　凡,史艳翠,等. 移动用户需求获取技术及其应用[J]. 软件学报,2014,25(3):439-456.

[8] 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.

[9] Bobadilla J,Ortega F,Hernando A,et al. A collaborative filtering approach to mitigate the new user cold start problem[J]. Knowledge-Based System,2012,26:225-238.

[10] 孟祥武,胡　勋,王立才,等. 移动推荐系统及其应用[J]. 软件学报,2013,24(1):91-108.

[11] Basiri A,Lohan E S. Overview of positioning technologies from fitness-to-purpose point of view[C]//International conference on localization and GNSS. Helsinki, Finland: IEEE, 2014:1-7.

[12] Yu Z,Zhou X,Zhang D,et al. Supporting context-aw-are media recommendation for smart phones[J]. IEEE Pervasive Computing,2006,5(3):68-75.

[13] Moradeyo A,Teresa M. Supporting context-aware cloud-based media recommendation for smartphones[C]//2nd IEEE international conference on mobile cloud computing, service and engineering. Oxford,England:IEEE,2014:109-117.

[14] Lee T Q,Park Y,Park Y T. A time-based approach to effective recommender systems using implicit feedback[J]. Expert Systems with Applications,2008,34(4):3055-3062.

[15] Szomszor C,Cattuto H,Alani K,et al. Folksonomies,the semantic web,and movie recommendation[C]//Workshop on bridging the gap between semantic Web and Web 2.0. Innsbruck,Austria:Springer,2007:71-84.

(上接第 63 页)

ters for k-means[C]//Proceedings of the 2nd international conference on computer science and electronics engineering. Paris,France:Atlantis Press,2013:527-530.

[6] Kumar Y,Sahoo G. A new initialization method to originate initial cluster centers for K-Means algorithm[J]. International Journal of Advanced Science and Technology,2014,62:43-54.

[7] Ma L, Gu L, Li B,et al. An improved k-means algorithm based on MapReduce and grid[J]. International Journal of Grid and Distributed Computing,2015,8(1):189-200.

[8] Rodriguez A,Laio A. Clustering by fast search and find of density peaks[J]. Science,2014,344(6191):1492-1496.

[9] 张晓倩,曲福恒,杨　勇,等. 一种高效的基于初始聚类中心优化的 K-means 算法[J]. 长春理工大学学报:自然科学版,2015(4):154-158.

[10] 何佳知,谢颖华. 基于密度的优化初始聚类中心 K-means 算法研究[J]. 微型机与应用,2015,34(19):17-19.

[11] Yuan Q,Shi H,Zhou X. An optimized initialization center K-means clustering algorithm based on density[C]//IEEE international conference on cyber technology in automation,control,and intelligent systems. [s. l.]:IEEE,2015:790-794.

[12] Li C,Zhang Y,Jiao M,et al. Mux-Kmeans:multiplex kmeans for clustering large-scale data set[C]//Proceedings of the 5th ACM workshop on scientific cloud computing. [s. l.]: ACM,2014:25-32.

[13] Lin Y, Luo T, Yao S, et al. An improved clustering method based on k-means[C]//9th international conference on fuzzy systems and knowledge discovery. [s. l.]:IEEE,2012:734-737.

[14] Chu S Y,Deng Y N,Tu L L. K-means algorithm based on fitting function[C]//International conference on applied science and engineering innovation. [s. l.]:[s. n.],2015:1940-1945.

[15] Le V H,Kim S R. K-strings algorithm,a new approach based on Kmeans[C]//Proceedings of the 2015 conference on research in adaptive and convergent systems. [s. l.]: ACM, 2015:15-20.