最近邻优化的 k-means 聚类算法

林 涛 赵 璨
(河北工业大学计算机科学与软件学院 天津 300401)

摘 要 传统的 k-means 算法不论其数据样本的分布情况,将簇边缘位置、簇中心位置、离群点的数据样本全部按照最小距离原则,划分到离它最近的聚类中心所在簇中,没有考虑数据样本与其他簇之间的关系。如果数据样本与另一簇中心的距离接近于最小距离,则此数据样本与两个簇的关系都很大,显然这样直接划分并不合理。针对此问题,文中提出了最近邻优化的 k-means 聚类算法。运用近邻的思想,将这些不"很属于"某簇的数据样本划分到其最近邻数据样本所在的簇中,实验结果表明,这种最近邻优化的 k-means 聚类算法有效地减少了算法的迭代次数,提高了算法的聚类准确度,得到了良好的聚类效果。

关键词 k-means,分布,关系,簇,最近邻
中图法分类号 TP181 **文献标识码** A

Nearest Neighbor Optimization k-means Clustering Algorithm

LIN Tao ZHAO Can
(School of Computer Science and Engineering, Hebei University of Technology, Tianjin 300401, China)

Abstract Traditional k-means algorithms usually ignores the distribution of the data samples, assign all of them in the cluster edge position, center position, outliers to the cluster which nearest clustering center locates, in accordance with the principle of minimum distance, without considering the relationsh1ip between the data sample and other clusters. If the distance between the data sample and the other cluster is close to the minimum distance, the data sample is very close to the two clusters, obviously, the direct division menthod is not reasonable. Aiming at this problem, this paper presented a clustering algorithm optimized nearest neighbor (1NN-kmeans). Using the ideas of neighbor, assign these samples that do not firmly belong to a certain cluster to the cluster that the nearest neighbor sample belongs to. The experimental results show that 1NN effectively reduced the number of iterations and improved the clustering accuracy and finally achieved the better clustering results.

Keywords K-means, Distribution, Relationship, Cluster, Nearest neighbor

1 引言

聚类分析在数据挖掘领域[1]中是一个研究热点,聚类是将对象分类的过程,同一个簇中的数据相似,不同簇中的数据对象差别较大,从而发现隐藏在数据背后的重要信息。k-means 是一种基于划分的硬聚类[2-3]算法,以最小误差平方和为目标函数,形成独立紧凑的簇,其运行速度快且适用于高维数据而被广泛应用。

针对 k-means 存在的缺陷,研究人员从不同角度提出了改进方案。文献[4]通过数据样本分层确定聚类数搜索范围,然后从中确定 k 值。文献[5]从方差的角度选择方差最小且有一定距离的数据样本作为初始聚类中心。文献[6]通过构造最小生成树,对其剪枝得到初始聚类中心。文献[7]提出了对离群点的处理办法,离群点不参与聚类中心的更新,最后将其直接划分到离其最近的聚类中心所在簇中。k-means 算法是一种硬聚类算法,严格地将数据样本划分到某一类中;模糊 c 均值(FCM)聚类[8]算法是一种软聚类算法,其并不直接将数据样本明确地归于某一簇,而是用隶属度表示属于某一簇的程度。文献[9]为了减小不同规则簇对聚类的影响,利用样本 k 近邻定义了一种新的隶属度。FCM 算法充分考虑数据样本属于簇的程度,根据隶属度更新聚类中心,按照最大隶属度划分样本。

对 k-means 的改进主要针对 k 值、初始聚类中心、离群点 3 个方面,很少考虑数据样本与簇的关系,将"很属于"某簇和不是"很属于"某簇的数据样本同等对待,全部按照最小距离划分。而软聚类的算法中只是应用定义的隶属度计算数据样本属于簇的程度,然后完全按照隶属度更新聚类中心,划分数据样本。本文通过设置阈值 ε,计算 r_i 界定数据样本属于簇的程度,对于不是"很属于"簇的数据样本按照最近邻划分,使数据样本的划分更加合理准确。

2 相关知识

待聚类的数据集 $X=\{x_i \mid x_i \in R^p, i=1,2,\cdots,n\}$。

定义 1 任意两个数据样本 x_i 和 x_j 的欧氏距离为:

本文受天津市自然科学基金重点项目(13jczdjc34400),河北省科技计划项目(17214304D),天津市科技重大专项(14ZCDZGX00818)资助。
林 涛(1970—),男,博士生,主要研究领域为物联网技术与数据分析,E-mail:1965511415@qq.com(通信作者)。

$$d_{ij}=\sqrt{(x_i-x_j)^{\mathrm{T}}(x_i-x_j)}$$

定义 2　数据样本 x_i 距离聚类中心第二小距离与最小距离的比值：

$$r_i=\frac{d(x_i,C_m)}{d(x_i,C_n)}$$

定义 3　数据样本集 X 除去一些数据样本后剩余的数据样本：

$$X-X'=\{x_i\mid x_i\in R^p,i=1,2,\cdots,n\text{ 且 }i\neq p,q,r\}$$

$$X'=\{x_p,x_q,x_r\}$$

定义 4　聚类结果的误差平方和：

$$E=\sum_{i=1}^{n}\sum_{k=1}^{K}|x_i-C_k|^2$$

定义 5　聚类算法的准确度：

$$r=\frac{n_1}{n}\times 100\%$$

其中，n_1 是正确聚类的数据样本个数，n 是总样本个数。

2.1　KNN 算法

K 近邻分类算法[10-12]是一种简单的有监督学习算法，其核心思想是计算样本数据的 k 个最近邻，将样本数据归于 k 个最近邻所属某类数量最多的类别中，因此 KNN 在分类决策上只依据最近邻的一个或者几个样本的类别，分类精度高且对异常数据不敏感，通常与其他算法结合使用。

当 $k=1$ 时，称为最近邻算法（1NN），分类的结果只依赖于最近邻的数据样本。

2.2　k-means＋＋初始聚类中心选择策略

随机选择初始聚类中心的方法带有很强的不确定性，往往达不到理想的聚类效果，k-means＋＋[13-14]选择初始聚类中心的策略因为思想简单、易于实现且效果良好，被广泛集成。

1NN-kmeans 算法整合了 k-means＋＋初始聚类中心选择策略，以消除初始聚类中心对聚类效果的消极影响，以下为 k-means＋＋选取初始聚类中心的具体流程：

(1)从样本集中随机选择一个数据对象作为第一个聚类中心 C_1。

(2)对于每一个数据对象 x_i，计算其与已选择的聚类中心中最小的距离：$D_x=\min d(x_i,C_k')(k'=1,\cdots,kselected)$。

(3)选择一个新的数据对象作为新的聚类中心，选择策略是：$D(x)$较大的点被选择聚类中心的概率较大。

(4)重复步骤(2)和步骤(3)直到选出 K 个初始聚类中心。

3　1NN-*km*eans

在数据集 $X=\{x_1,x_2,\cdots,x_n\}$ 中，计算数据样本 x_i 与所有聚类中心 $C_1,C_2,\cdots,C_K$ 的距离并升序排序 $d_{ik_1}^1,d_{ik_2}^2,\cdots,d_{ik_K}^K$，$d_{ik_1}^1$ 表示数据样本 x_i 与聚类中心 Ck_1 的距离，且是最小距离，$d_{ik_2}^2$ 表示数据样本 x_i 与聚类中心 Ck_2 的距离，且为第二小距离，若 $d_{ik_2}^2$ 与 $d_{ik_1}^1$ 的比值趋于无穷大，说明 x_i 明显靠拢 Ck_1 而远离其他聚类中心，x_i 偏于簇的中心位置，x_i“很属于”Ck_1 所在簇，可以直接将 x_i 划分过去，若 $d_{ik_2}^2$ 与 $d_{ik_1}^1$ 的比值小于某一阈值 ε，表示 x_i 靠近 Ck_1 的同时也靠近其他的聚类中心，x_i 可能是离群点也可能在空间位置上在两个或者多个聚类中心的中间，x_i 不是“很属于”Ck_1 所在簇，若依然按照最小距离原则划分 x_i，可能会造成分类错误。由此利用近邻思想，将 x_i 划分至最近邻所属簇中，以此减少迭代次数，提高算法准确度。

ε 的值界定数据样本是否“很属于”簇，因此 ε 对算法的影响很大，ε 没有一个固定的计算方法，通常设置在 1.5 左右，根据数据集的数据分布情况进行微小调节。

NN-Kmeans 算法如算法 1 所示。

算法 1　NN-Kmeans

输入：样本数据集 $X=\{x_1,x_2,\cdots,x_n\}$，聚类个数 K

输出：K 个聚类

1. 应用 k-means＋＋初始聚类中心选取策略选取 K 个初始聚类中心 $C_k=\{C_1,\cdots,C_k\}$。
2. 如果数据样本 x_i 所属类别为空，定义 ArrayList 集合 aList，计算 x_i 与所有聚类中心的欧氏距离 d_{ik}，将 d_{ik} 升序排序 $\{d_{ik_1}^1,d_{ik_2}^2,\cdots,d_{ik_K}^K\}$，将 x_i 的序号 i 添加到 aList 中，根据定义 2 计算 r_i。
3. 如果 $r_i>\varepsilon$，则直接根据 x_i 的最小距离原则划分 x_i。
4. 如果 $r_i<\varepsilon$，则根据定义 1 和定义 3 计算 x_i 在数据集 $X-\{x_p\mid p\in aList\}$ 中的最近邻 x_j，将序号 j 添加到 aList 中，如果 x_j 已经划分到 C_k 所在簇，则同样将 x_i 划分到 C_k 所在簇，如果 x_j 还未被划分，根据处理 x_i 的过程计算 x_j 与聚类中心的距离，如果 $r_j>\varepsilon$，则不需要计算 x_j 的最近邻，直接划分 x_i 和 x_j 到 x_j 要划分的簇中，如果 $r_j<\varepsilon$，则同样排除 aList 元素对应的样本计算 x_j 的近邻，重复以上过程，直到集合 aList 中最后一个元素对应的样本能够确定所属簇，aList 中第一个元素对应的 x_i 连同其他元素对应的样本都归于此簇。
5. 按照步骤 2－步骤 4 将数据集中的数据样本一一划分完毕后，根据定义 4 计算聚类结果的误差平方和 E′。
6. 如果 $E'-E<10^{-10}$，则算法收敛输出聚类结果；否则 $E=E'$，计算每个簇的质心作为聚类中心，循环执行步骤 2－步骤 5 直到算法收敛。

4　实验

4.1　数据集

为了验证基于近邻思想对 k-means 改进的 1NN-kmeans 算法的有效性，选用 UCI[15] 数据库中的 6 个数据集（iris，glass，wine，ionosphere，balance-scale，haberman）作为仿真实验数据集，所选用的数据集均不含缺失值。针对每一数据集运行算法 100 次，求 100 次结果的平均值作为有效性分析的数据依据，并与传统的 k-means、优化初始聚类中心的 k-means＋＋算法的聚类结果进行对比。表 1 列出了 6 个数据集的相关属性值。

表 1　UCI 数据集描述

数据集名称	样本个数	属性个数	类别数
iris	150	4	3
glass	214	9	6
wine	178	13	3
ionosphere	351	34	2
balance-scale	625	4	3
heberman	306	3	2

4.2　ε 值的设定

ε 是决定数据样本选择哪种划分方法的阈值，因此对 ε 的设定至关重要，其直接影响最后的聚类效果。当 $r_i<\varepsilon$ 时，看作此数据样本不是“很属于”某簇，不适宜用最小距离划分，参照最近邻的数据样本更合理，r_i 是第二小距离与最小距离的比值，由此可得 $\varepsilon\geqslant 1$，通常设置在 1.5 左右，不宜过大，ε 接近无穷大的情况便是将全部的数据样本划分到一个簇中的情况。因此 ε 过大会造成数据集中更多的数据样本按照最近邻划分，增强算法的聚集性，将更多的数据样本划分到一个簇

中，造成不能识别较为松散的簇，这不仅会增加更多的时间开销，还会造成分类错误。

由于每个数据集的簇规则不同，因此 ε 的最佳取值也不同。以 balance-scale 数据集为例，改进的算法的准确度能够达到 100%，但是在对 k-means++算法进行的 500 次实验中，未出现全部分类正确的情况。为了提取最佳的 ε 值以及查看不同的 ε 值对准确度的影响，将算法在 ε=1.1，…，1.9 时各运行 500 次，统计准确度为 100%的次数，结果如表 2 所列。

表 2 ε 与对应的次数

ε	准确度为 100%的次数
1.1	83
1.2	101
1.3	157
1.4	202
1.5	215
1.6	181
1.7	163
1.8	158
1.9	109

从表 2 可以看出，在 ε=1.5 时，准确度为 100%的次数为最大值，占总实验次数的 43%，并且在 ε<1.5 和 ε>1.5 时次数呈下降趋势，由此得出对于 balance-scale 数据集 ε 的最佳值为 1.5。按照上述确定 ε 的方法，确定其他数据集的 ε 值，结果如表 3 所列。

表 3 实验数据集对应的 ε 值

数据集	ε
iris	1.8
glass	1.1
wine	1.1
ionosphere	1.2
balance-scale	1.5
heberman	1.4

4.3 实验结果与分析

实验运行的 k-means，k-means++，1NN-kmeans 算法是在开源框架 weka 下实现的，修改 weka 源码集成算法 1NN-kmeans，ε 设置为最佳取值，在每一数据集下将每一算法运行 100 次，求 100 次结果的平均值作为有效性的分析数据，表 4 列出了迭代次数、误差平方和(E)、准确度的对应结果。为了更清晰地显示 1NN-kmeans 算法的有效性，将表 2 中的数据绘制成了图 1－图 3 的折线图，图 4 是运行时间的折线图。

表 4 实验结果数值

数据集	迭代次数			*E*			准确度/%		
	k-means	k-means++	1NN-kmeans	k-means	k-means++	1NN-kmeans	k-means	k-means++	1NN-kmeans
iris	3.32	2.92	2.33	143.9	109.3	71.25	73.51	92.95	94.85
glass	7.05	6.12	2.38	519.1	359.9	334.6	51.6	80.1	86.7
wine	6.66	6.35	4.41	455.1	433.7	351.3	70.08	94.4	91.6
ionosphere	7.24	6.82	3.95	6810	6331	4550	73.5	73.5	75.21
balance-scale	9.48	9.37	5.61	3193	2947	2084.26	71.53	75.8	91.79
haberman	5.42	5.36	2.93	2101	2067	1230	63.26	75.16	95.09

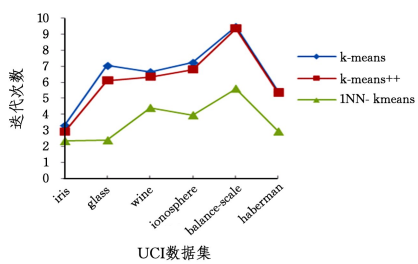

图 1 迭代次数

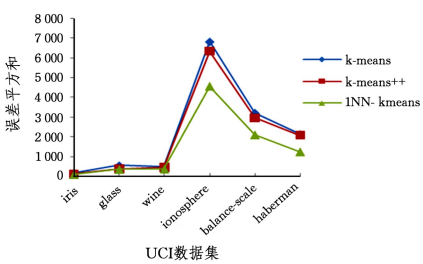

图 2 误差平方和

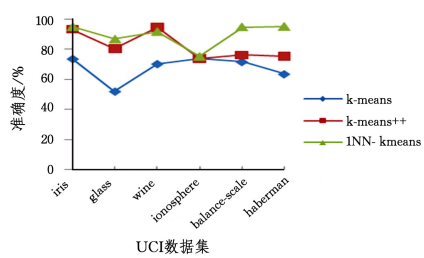

图 3 准确度

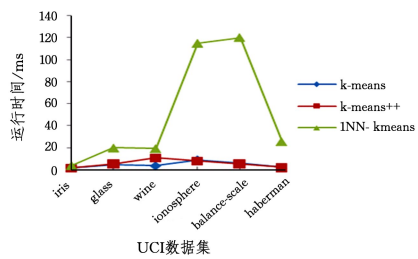

图 4 运行时间

在迭代次数方面，1NN-kmeans 算法的效果显著。由实验数值和图 1 可知，k-means 和 k-means++迭代次数相近，k-means++迭代次数稍小于 k-means，经典的 k-means 算法以随机的方式选择初始聚类中心，聚类效果含有很强的不确定性，k-means++初始中心选择策略是随机方法的优化，效果优于 k-means。1NN-kmeans 算法的优势明显，例如在 balance-scale 数据集上更是优于 k-means++算法 4 次，但聚类结果却很稳定。1NN-kmeans 算法的聚集性较好，能够在每次迭代中将数据样本稳定地归于某一簇；误差平方和通常作为目标函数衡量聚类效果的标准，误差平方和越小，簇内越紧凑，聚类效果就越好，由实验数据与图 2 看出，1NN-kmeans 误差平方和在 6 个数据集上都小于其他两种算法。用定义 5 计算准确度，1NN-kmeans 算法的准确度较高，在 haberman 数据集上 1NN-kmeans 算法的准确度能够达到 90%以上，而其他两种算法在 80%以下，wine 数据集的准确度虽不及 k-means++，但基本能与其持平，两者都保持在 90%以上的

较高准确度，充分验证了 1NN-kmeans 算法理论上的可行性和有效性。由图 4 可以看出，1NN-kmeans 运行时间稍长，这是因为计算 $r_i<\varepsilon$ 的数据样本的最近邻耗费了时间，但 3 种算法的运行时间都在毫秒级，均未超过 1 s。

1NN-kmeans 算法的聚集性较好，聚类的结果通常只是某一簇的某些数据样本错误地全部分到另一个簇中，而其他的簇能够全部分类正确，通常不会是原本一个簇中的数据被分到其他几个簇中去。1NN-kmeans 聚类结果很稳定，在数次实验中能够得到不变的结果，对某些数据集的准确度能够达到 100%。针对数据样本与簇的关系不同采用不同的划分方法更合理和有效。

结束语　传统的 k-means 算法将所有属于簇的程度不同的数据样本同等对待，一概按照最小距离原则划分。本文提出的 1NN-kmeans 算法通过计算第二最小距离与第一最小距离的比值是否大于某一参数 ε，来界定数据样本是否“很属于”某一簇，而与其他簇关系很小，对于不是“很属于”簇的数据样本，采用近邻思想将其划分到最近邻所在的簇中。实验选择 UCI 数据库中的 6 个数据集，对算法的有效性进行验证。通过与 k-means，k-means＋＋对比可以得出，1NN-kmeans 算法能够在较低的迭代次数中，达到较高的准确度和较低的误差平方和，聚类结果稳定，适应不同簇规则的数据集，是一种高效的 k-means 优化算法。

参 考 文 献

[1] 高曼，韩勇，陈戈，等. 基于 K-means 聚类算法的公交行程速度计算模型[J]. 计算机科学，2016，43(S1)：422-424，439.

[2] 赵建民，管国权，王红艳. 基于遗传算法的硬聚类算法改进[J]. 计算机工程与科学，2008(8)：83-85.

[3] 唐胡鑫. 电子商务客户忠诚度模型仿真研究[J]. 计算机仿真，2016，33(1)：413-415，424.

[4] 王勇，唐靖，饶勤菲，等. 高效率的 K-means 最佳聚类数确定算法[J]. 计算机应用，2014，34(5)：1331-1335.

[5] 谢娟英，王艳娥. 最小方差优化初始聚类中心的 K-means 算法[J]. 计算机工程，2014，40(8)：205-211，223.

[6] 郁启麟. K-means 算法初始聚类中心选择的优化[J]. 计算机系统应用，2017，26(5)：170-174.

[7] 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.

[8] 朴尚哲，超木日力格，于剑. 模糊 C 均值算法的聚类有效性评价[J]. 模式识别与人工智能，2015，28(5)：452-461.

[9] 马闯，吴涛，段梦雅. 基于 K 近邻隶属度的聚类算法研究[J]. 计算机工程与应用，2016，52(10)：55-58，117.

[10] 王超学，潘正茂，马春森，等. 改进型加权 KNN 算法的不平衡数据集分类[J]. 计算机工程，2012，38(20)：160-163，168.

[11] 华辉有，陈启买，刘海，等. 一种融合 Kmeans 和 KNN 的网络入侵检测算法[J]. 计算机科学，2016，43(3)：158-162.

[12] 苏毅娟，邓振云，程德波，等. 大数据下的快速 KNN 分类算法[J]. 计算机应用研究，2016，33(4)：1003-1006，1023.

[13] ARTHUR D，VASSILVITSKII S. k-means＋＋：the advantages of careful seeding[C]// Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics Philadelphia，PA，USA，2007：1027-1035.

[14] 余秀雅，刘东平，杨军. 基于 K-means＋＋的无线传感网分簇算法研究[J]. 计算机应用研究，2017，34(1)：181-185.

[15] ASUNCION A，NEWMAN D J. UCI machine learning repository[EB/OL]. [2009-12-23]. http://archive.ics.uci.edu/.

（上接第 203 页）

[5] DENG X Q，LIU T H，LI W Z，et al. A Latent Factor Model of Fusing Social Rregularization Term and item Regularization Term [J]. Physic A：Statistical Mechanics and its Applications，2019，525：1330-1342.

[6] LI H，DIAO X，CAO J，et al. Collaborative Filtering Recommendation Based on All-Weighted Matrix Factorization and Fast Optimization [J]. IEEE Access，2018，6：25248-25260.

[7] LIN C，WANG L，TSAI K. Hybrid Real-Time Matrix Factorization for Implicit Feedback Recommendation Systems [J]. IEEE Access，2018，6(10)：21369-21380.

[8] HU Y，KOREN Y，VOLINSKY C. Collaborative filtering for implicit feedback datasets[C]// Proceedings of the 8th IEEE International Conference on Data Mining. Pisa，Italy：IEEE，2008：263-272.

[9] LING G，YANG H，KING I，et al. Online learning for collaborative filtering[C]// The 2012 International Joint Conference on Neural Networks. Brisbane，Australia：ACM，2012：1-8.

[10] MA H，ZHOU D，LIU C，et al. Recommender systems with social regularization[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. Hong Kong：ACM，2011：287-296.

[11] MA H，YANG H，LYU M R，et al. Sorec：social recommendation using probabilistic matrix factorization [C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Velley，Califomia：ACM，2008：931-940.

[12] JAMALI M，ESTER M. A matrix factorization technique with trust propagation for recommendation in social networks[C]// Proceedings of the 4th ACM Conference on Recommendation Systems. Barcelona，Spain：ACM，2010：135-142.

[13] PAN R，ZHOU Y H，CAO B. One-class collaborative filtering [C]// Proceedings of the 8th IEEE International Conference on Data Mining. Pisa，Italy：IEEE，2008：502-511.

[14] TANG J，HU X，GAO H，et al. Exploiting local and global social context for recommendation[C]// Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Palo Alto，Califomia：AAAI，2013：2712-2718.

[15] BREESE S J，HECKERMAN D，KADIE C M. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[C]// Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence. Madison，wisconsin：ACM，1998：43-52.

[16] MINAS G，CARTER T B，MACIEJ K，et al. Multigraph Sampling of Online Social Networks [J]. IEEE Journal on Selected Areas in Communications，2011，29(3)：1893-1905.

[17] LIAN D，ZHAO C，XIE X，et al. GeoMF：Joint geographical modeling and matrix factorization for point-of-interest recommendation[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM，2014：831-840.