

一种基于离群指数的初始聚类中心优选算法

李 向¹, 刘素红^{2,3}

(1 郑州航空工业管理学院 计算机科学与技术系, 河南 郑州 450015;

2 北京师范大学 地理学与遥感科学学院, 北京 100875;

3 北京师范大学 遥感科学国家重点实验室, 北京 100875)

摘 要: 传统 K-均值聚类算法的初始聚类中心是随机选择的, 不同的初始聚类中心会得到不同的聚类结果, 聚类结果随机性较大、稳定性差。采用局部离群指数优化 K-均值聚类算法, 通过计算所有数据样本的局部离群指数, 选择 k 个相互距离最远的局部密集点作为初始聚类中心, 消除局部离群点对算法的影响。实验结果证明, 该算法能降低 K-均值聚类算法初始聚类中心选取的敏感度, 减少迭代次数, 得到更为准确的聚类结果。

关键词: 聚类; 离群指数; 初始聚类中心; K-均值聚类

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2013)06-0109-04

The Optimal Initial Centers Clustering Algorithm Based on Local Outlier Factor

LI Xiang¹, LIU Su-hong^{2,3}

(1 Department of Computer Science and Application, Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou 450015, China; 2 School of Geography and Remote Sensing Science, Beijing Normal University, Beijing 100875, China;

3 State Key Laboratory of Remote Sensing, Beijing Normal University, Beijing 100875, China)

Abstract: The initial clustering centers of traditional K-means clustering algorithm are generated randomly. Different initial clustering centers will lead to different clustering results. It means that unstable results were often obtained owing to random selection of initial centers while using traditional K-means. A new method about optimization of initial centers is brought forward based on K-means clustering algorithm which can gain high accurate results. The main idea of the improved algorithm is to choose the K-means initial clustering centers through calculating the local outlier factor of all sample data. The greatest distance dense points are chosen to eliminate the influence of local outliers. The experimental results show that the improved algorithm can reduce the selection sensitivity of the K-means initial clustering centers, and also get more accurate clustering results with less iterations.

Key words: cluster; local outlier factor; initial clustering center; K-means

1 引言

数据聚类(或聚类分析)是为了获得图案、点集或对象集本然的分组关系^[1]。聚类分析参照“物以类聚”的思想, 通过研究抽取样本数据的潜在结构, 将数据对象分组成为多个类(或簇), 使得同一个类中

的对象之间具有较高的相似度, 而不同类中的对象差别较大。由于类的形成完全是数据驱动的, 不需要任何的先验信息和假设, 因此聚类分析是一种非监督学习方法, 广泛应用于数据挖掘和数据分析^[2]。

K-均值聚类算法是基于划分的聚类算法, 该算法简单、快速、应用广泛^[3]。但是传统 K-均值聚类算

收稿日期: 2012-09-10; 修回日期: 2012-10-28

基金项目: 国家自然科学基金项目(41001235); 航空科学基金项目(2011ZC55005)

法的初始聚类中心是从数据集中随机产生的,而聚类结果的好坏又取决于聚类中心的选择. 所以该算法得到的通常是局部最优解而非全局最优解,并且聚类结果很不稳定. 针对这一缺点,学者们进行了广泛的研究,文献[4]中提出了基于竞争学习的 K 质心组合聚类算法,文献[5]中提出了一种基于密度算法优化初始聚类中心的改进算法,文献[6]中利用贪心算法思想构建 k 个数据集合,取集合中数据的平均值作为初始聚类中心. 这些研究都从聚类中心选取方法入手,本质上是对数据集按照某种策略初选,优先考虑那些内聚度较高的作为聚类中心的备选,从中选择并指定为 K-均值聚类算法的聚类中心.

针对随机选取初始聚类中心的缺陷,本文提出一种基于局部离群指数的改进 K-均值聚类算法,该算法首先采用 LOF(local outlier factor)^[7] 算法计算数据样本中每一个点的局部离群指数,然后将样本中的离群点排除在初始聚类中心的选取之外,选择相互距离最远的 k 个局部密集点作为初始聚类中心,消除聚类结果的波动性. 实验表明,与传统 K-均值聚类算法相比,这种优化初始聚类中心的方法能有效改进聚类算法,准确率较高.

2 传统 K-均值聚类算法

2.1 传统的 K-均值聚类算法

传统的均值聚类算法描述如下:

输入:具有 n 个对象的数据集 X 和聚类分组数 k .

输出: k 个聚类,使得目标函数 E 最小.

(1) 从样本数据集中任意选择 k 个对象作为初始聚类中心;

(2) 对数据集每一个对象,根据距离最近的原则,将其分配到相应的簇中;

(3) 针对每个簇,重新计算聚类中心,新的聚类中心为簇中所有对象的均值;

(4) 计算目标函数 E ;

(5) 若目标函数 E 收敛,算法终止,否则转(2).

2.2 传统的 K-均值聚类算法的局限性

传统 K-均值聚类算法中初始聚类中心的选择是随机选取的,初始聚类中心选取的不同会产生不同的聚类结果. 对 UCI 数据库中 Iris 数据采用传统 K-均值聚类算法分为 3 类,8 次实验结果,如表 1 所示.

从实验结果中可以看出,传统的 K-均值聚类算法受初始聚类中心的影响比较大. 根据初始聚类中

心选择的不同,分类准确率也不同,在 8 次实验中最好与最差相差 40.7%. 针对这一问题常用的方法是通过多次随机生成初始聚类中心,计算聚类结果,选择目标函数最小的一次计算结果作为最后的聚类结果,如 Matlab 中 K-均值聚类的实现. 该方法的缺点是耗时严重,试验效果不稳定,未必能够得到最优结果. 所以合理地选择初始聚类中心是 K-均值聚类算法改进的关键.

表 1 传统 K-均值聚类算法在 Iris 数据上的实验结果

初始中心	正确分类个数	错误分类个数	准确率
3,7,87	87	63	58%
9,18,101	86	64	57.3%
11,56,53	132	18	88%
35,23,88	72	78	48%
36,25,56	77	73	51.3%
14,11,97	86	64	57.3%
35,6,21	132	18	88%
100,105,144	133	17	88.7%

3 基于 LOF 的初始聚类中心优选

为了克服 K-均值聚类算法对初始聚类中心敏感的问题,本文在对聚类算法的改进中,引入计算数据对象局部离群指数的方法,避免初始聚类中心选择到局部离群点,使聚类中心更能够反映数据的分布特征.

3.1 局部离群指数

D. Hawkins 给出了离群点的定义^[8]:离群点是偏离其他对象的、由不同一般的机制原理生成的观察对象,通常离群点也被称为异常点、孤立点、奇异点. 离群点的检测方法很多^[9],本文采用 LOF 算法,该算法计算数据集中每一个对象的局部离群指数,通过比较该指数的大小,来确定局部离群点. 指数越大,表明该点临近区域的对象分布密度越小,该对象越离群.

LOF 算法的描述如下:

(1) 计算对象 x 的 k -距离. 对于给定的一个正整数 k ,对象 x 的 k -距离记为 $k\text{-dis}(x)$,则在样本空间中,存在对象 y ,它与对象 x 之间的距离记为 $\text{dis}(x,y)$.

如果满足以下两个条件,我们认为 $k\text{-dis}(x) = \text{dis}(x,y)$:

① 在样本空间中至少存在 k 个对象 z ,使得 $\text{dis}(x,z) \leq \text{dis}(x,y)$;

② 在样本空间中至多存在 $k-1$ 个对象 z , 使得 $\text{dis}(x, z) < \text{dis}(x, y)$.

如图 1(a) 中所示, 当 $k=3$ 时, 计算二维样本中所有点到数据对象 x 的距离, 距离最近的第三个数据对象为 y , 如图 1(b) 所示, 存在 m 和 n 两个数据对象距离 x 的距离小于 $\text{dis}(x, y)$.

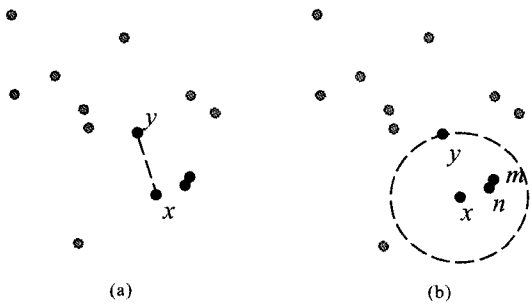


图 1 局部离群指数计算

(2) 找出对象 x 的 k -邻域. 计算出对象 x 的 k -距离之后, 与对象 x 之间距离小于等于 k -距离的对象的集合称为对象 x 的 k -邻域, 记为 $N_{k-\text{dis}(x)}(x)$. 如图 1(b) 所示, 即为对象 m, n 和 y .

(3) 计算对象 x 的可达距离. 可达距离记为 $\text{reachdis}_k(x, y)$. 对象 x 相对于对象 y 的可达距离用下面的公式计算:

$$\text{reachdis}_k(x, y) = \max\{k - \text{dis}(y), \text{dis}(x, y)\}.$$

(4) 计算局部可达密度. 局部可达密度是基于 x 的 k 最邻近点的可达距离的平均值的倒数.

$$L_{\text{den}}(x) = \frac{\sum_{y \in N_{k-\text{dis}(x)}(x)} \text{reachdis}_k(x, y)}{|N_{k-\text{dis}(x)}(x)|}$$

(5) 计算局部离群指数.

$$\text{LOF}(x) = \frac{\sum_{y \in N_{k-\text{dis}(x)}(x)} \frac{L_{\text{den}}(x)}{L_{\text{den}}(y)}}{|N_{k-\text{dis}(x)}(x)|}$$

以上是局部离群指数计算的步骤, 对象 x 的局部离群指数越大表明 x 的异常程度越高, 反之则可能越小. 按照数据集中对象的局部离群指数大小降序排列, 排列越靠前的对象离群的趋势越大, 排列越靠后的是密集程度越高.

3.2 基于局部离群指数的改进 K-均值聚类算法

基于局部离群指数的改进 K-均值聚类算法描述如下:

输入: 具有 n 个对象的数据集 X 和聚类分组数 k .

输出: k 个聚类, 使得目标函数 E 最小.

(1) 计算每个数据对象的局部离群指数;

(2) 将局部离群指数排在前 n 个的数据对象点删除, 得到处于密集区域的数据对象集合 D .

(3) 将局部离群指数最小的数据对象作为第一个聚类中心 o_1 , 将 o_1 加入初始聚类中心集合; 在集合 D 中找出距离 o_1 点最远的点 o_2 作为第二个聚类中心, 加入初始聚类中心集合, 并从 D 中删除 o_1, o_2 .

(4) 从集合 D 中找出 o_3 和 o_1, o_2 和 o_3 的距离之和最大, 然后将 o_3 加入到初始聚类中心集合, 将 o_3 从集合 D 中删除, 继续从集合 D 中找出到初始聚类中心对象集合中所有对象距离最远的点作为聚类中心, 直到第 k 个.

(5) 用得到的 k 个聚类中心, 运用 K-均值聚类算法对整个数据集进行聚类.

4 实验结果和分析

使用 UCI 数据库中的多个数据集对本文改进算法进行实验, 结果和分析如下.

4.1 稳定性分析

Iris 是人们广泛使用的数据集之一, 它含有 150 条数据, 可以分为三类, 每类 50 条. 每条数据用四个实数描述花的属性, 分别表示萼片 (sepal) 长度, 萼片宽度, 花瓣 (petal) 长度, 花瓣宽度.

采用本文方法对 Iris 数据进行聚类分析, 用三个属性数据作为坐标, 一次实验得到的结果如图 2 所示. 从图中可以看出, Iris 数据集中第一类 setosa 数据与其他数据离得较远, 第二类数据 versicolor 和第三类数据 virginica 有部分交叠.

根据多次实验的经验, 指定局部离群指数大于 1.5 的数据对象都应看作局部离群点, 在计算局部离群指数时, 参数 k 应设定为整个数据集样本个数的十分之一.

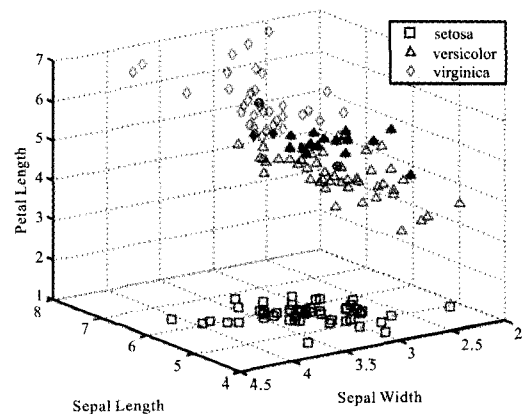


图 2 Iris 数据聚类结果

对 Iris 数据集的实验结果表明, 改进的聚类算

法正确率较高,达到 86.7%,与表 1 中传统 K-均值聚类算法结果对比可知,本文算法的分类精度优于传统算法的平均精度,此外,由于改进的算法通过局部离群指数计算优选了聚类中心,因此该算法聚类结果稳定性较好,算法的鲁棒性强,可以取得整体较好的聚类精度.此外,改进算法所分错的数据主要集中在 versicolor 和 virginica 两类之间,与传统 K-均值聚类算法分错的数据的分布情况一致,说明了 Iris 数据集中这两类数据的可区分性不强.

4.2 聚类精度及时间复杂度分析

在 UCI 数据库中的 Iris、Haberman 和 Hayes-Roth 进行实验,本文算法和传统 K-均值聚类算法运行 20 次后平均聚类精度、平均迭代次数和运行时间的对比如表 2 所示.

表 2 K-均值聚类算法和本文算法结果对比

算法	数据集	聚类精度	迭代次数	运行时间/ms
K-均值聚类	Iris	81.6%	9	16
	Haberman	51.96%	7	12
	Hayes-Roth	69.3%	12	20
本文算法	Iris	86.7%	4	125
	Haberman	73.82%	3	110
	Hayes-Roth	80.4%	5	164

从表 2 中可以看出,本文算法对 Iris、Haberman 和 Hayes-Roth 数据集的聚类精度比传统算法的平均精度都要高出许多,迭代次数也比传统算法要少,但是改进算法的运行时间较高.这说明本文所提出的改进算法聚类精度有较大提高,但是由于改进的算法中引入了局部离群指数计算,使算法的复杂度增加,综合考虑可以得出结论,对于较小数据个数的数据集和对时间要求不苛刻的条件下,可以使用本文算法.

5 结束语

本文提出一种新的初始聚类中心选择算法,局

部离群指数的引入,有效克服了 K-均值聚类算法对初始聚类中心的敏感性.

对多个数据集的实验证明,基于离群指数的改进 K-均值聚类算法,聚类结果稳定,并且保持了较高的准确率.但是由于引入局部离群指数计算,导致运算量较大,下一步要有针对性地研究筛选数据的方法,探讨如何减少算法的运算量,提高运算速度.

参考文献:

[1] Jain A K. Data clustering: 50 years beyond k-means [J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[2] HAN J W, Kamber Micheline, PEI J. Data mining concept and techniques[M]. 3rd ed. [S. l.]: Morgan Kaufmann-Elsevier, 2012.

[3] 冯超. K-means 聚类算法的研究[D]. 大连: 大连理工大学, 2007.

[4] 张宇, 邵良杉. 基于竞争学习的 K 质心组合聚类算法[J]. 计算机工程, 2011, 37(15): 40-42.

[5] 傅德胜, 周辰. 基于密度的改进 K 均值算法及实现[J]. 计算机应用, 2011, 31(2): 432-434.

[6] 全雪皎, 孟凡荣, 王志晓. 对 K-means 初始聚类中心的优化[J]. 计算机工程与设计, 2011, 32(8): 2721-2723.

[7] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, et al. LOF: identifying density-based local outliers [C]//Proceedings of ACM SIGMOD International Conference on Management of Data, Dallas, TX: ACM, 2000: 214-220.

[8] 徐翔, 刘建伟, 罗雄麟. 离群点挖掘研究[J]. 计算机应用研究, 2009, 26(1): 34-39.

[9] 薛安荣, 鞠时光, 何伟华, 等. 局部离群点挖掘算法研究[J]. 计算机学报, 2007, 30(8): 1455-1463.

作者简介:

李 向 女, (1982—), 硕士, 讲师. 研究方向为数据挖掘、遥感应用.

刘素红 女, (1967—), 博士, 副教授, 博士生导师. 研究方向为遥感与 GIS 集成、数据库知识发现.