

一种基于动态网格技术的 K-means 初始质心选取算法

张 真, 任贺宇

(河南交通职业技术学院, 河南 郑州 450000)

摘 要: 针对 K-means 算法随机选取中心点而无法得到全局最优误差平方和 SSE 的问题, 提出了一种基于移动网格技术的 K-means 中心点选取算法, 该算法在明显减少 K-means 算法的迭代次数的同时, 近似得到 SSE 的全局最优。

关键词: 聚类; 动态网格; K-means

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2013)06-0101-04

Anintial Centroid Selected Algorithm Based on Dynamic Grid Technology

ZHANG Zhen, REN He-yu

(Henan Transportation Vocational Technical College, Zhengzhou 450000, China)

Abstract: In order to solve the problem that K-means Algorithm cannot get global optimal SSE(sum of the squared error) by choosing centers randomly, this paper presents a new algorithm of choosing centers of K-means based on shifting grid, which decreases iterative time of K-means obviously and gets the global optimal of SSE approximately.

Key words: Clustering; dynamic grid; K-means

1 引言

K-means 算法在处理大数据集时, 具有可伸缩性和高效率。但该算法的执行效率和效果容易受初始质心点的选取影响。

针对该问题, 本文提出了一种基于动态网格^[1]的 K-means 中心点选取算法^[2-4], 该算法利用动态网格的密度可达性^[5], 确定出聚类的个数 K , 并选出比较合理的初始中心点位置, 从而提高 K-means 算法的效率和聚类精度^[6]。

2 相关概念与定义

假设 $A = \{A_1, A_2, A_3, \dots, A_k\}$ 是一个有界、全序域的集合, $s = A_1 \times A_2 \times \dots, A_k$ 是一个 k 维数值空间, 其中 $A_1, A_2, \dots, A_k$ 表示 s 的 k 个维或 k 个属性域, k 维数据集 $X = \{x_1, x_2, \dots, x_n\}$ 表示 s 上的一个点集, 其中 $x_i = \langle x_{i1}, x_{i2}, \dots, x_{ik} \rangle$ 表示一个数据点, x_{ij} 为数据点 x_i 的第 j 维的值。

定义 1: 动态网格

定义一个动态网格为超矩形区域: $\prod_{i=1}^d [x_i - \text{Eps}_i/2, x_i + \text{Eps}_i/2]$, 其中 $x = (x_1, x_2, \dots, x_d)$ 为中心点, $\text{Eps}_i = (\max_{x \in D} \{x_i\} - \min_{x \in D} \{x_i\})/k$, k 为按固定网格划分方法每一维上划分的区间个数。这样定义网格的方式就保证了生成的网格数远远小于数据点数 n , 且与数据点的维度无关, 也无空网格的出现。

定义 2: 网格密度相连和相离

如果网格 g_i, g_j 满足 $\text{density}(g_i \cap g_j) \geq \text{threshold}$, 则定义网格 g_i, g_j 为密度相连的。其中 $\text{density}(g_i \cap g_j)$ 定义了两个网格相交区域的密度, threshold 为密度阈值。

如果两个网格不是密度相连关系, 则称两个网格相离。

本文对两个网格之间只定义两种关系: 密度相连和相离, 如果两个网格在物理上相交, 但相交区域

收稿日期: 2012-09-10; 修回日期: 2012-10-27

基金项目: 国家自然科学基金项目(70701013)

无任何数据点或相交密度小于 threshold, 则两个网格均可能为聚类的边界, 因此, 本文中的定义是合理的。

定义 3: 网格划分无向图

以网格为顶点, 以两个网格是否密度相连为边所构成的无向图为网格划分无向图。

3 动态网格划分技术

动态网格划分技术是针对固定网格划分技术和自适应网格划分技术而提出了一种网格划分技术, 该划分技术的优点如下:

- (1) 不产生空网格, 减少了网格数量;
- (2) 网格数量与数据集维度无关, 只与数据集大小和分布有关;
- (3) 网格划分快。

动态网格的划分过程为: 算法读入一个数据点, 先判断该点是否属于已生成的网格。若是, 则将该点存入该网格中; 若一个点属于多个网格, 说明这些网格是相交的, 则记录下网格相交的信息和共享点信息; 若该点不能落入任何一个已有的网格中, 按照定义 1, 以该点为中心, 以 Eps_i 为第 i 维的边长创建一个网格。继续这一过程, 直到所有的点都处理完毕为止。图 1 中给出了划分算法的流程图。

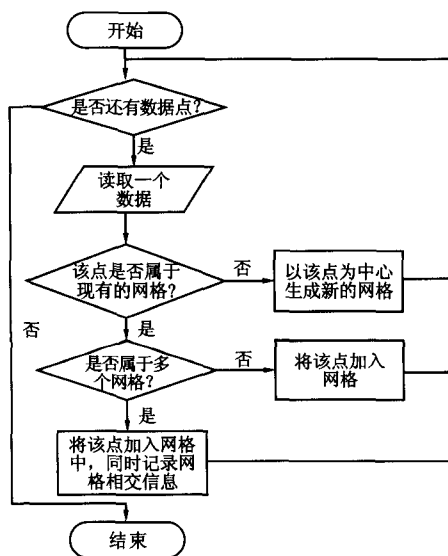


图 1 动态网格划分算法流程图

从网格的划分过程可以看出, 网格的数量独立, 只与数据集的大小以及数据点的分布有关。图 2 中以更直观的方式给予展现。

为了降低算法的复杂度空间复杂度, 在网格的划分过程中只需记录落入网格中的点数与网格间的相交信息, 而无需在网格中记录相应的数据点信息。

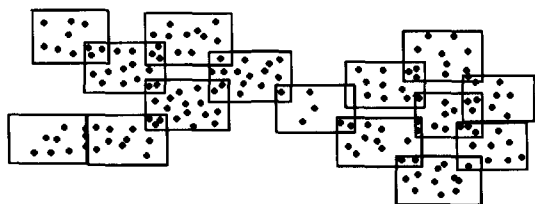


图 2 固定网格划分方式示意图图表

4 选取 K-means 聚类个数 K 和初始质心

4.1 计算聚类个数 K

完成网格划分后, 按照定义 3 生成网格划分无向图。在生成该无向图之前, 需计算任意两个网格之间的密度可达性, 对于图 3 所示网格划分示意图, 得到的网格划分无向图如图 3(b) 所示。

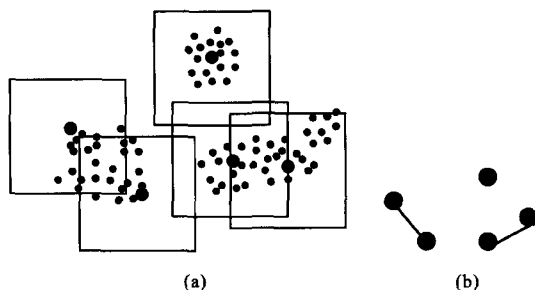


图 3 网格划分无向图

因此, 聚类 K 的个数即为网格划分无向图中连通子图的个数, 对于图 3 所示情况, 网格划分无向图中包含 3 个连通子图, 则聚类个数为 3。但本策略易受孤立点的影响, 因为对于孤立网格, 算法也会为其生成一个包含该个顶点的图, 对于此种情况, 可以通过为算法添加去除孤立网格的步骤加以消除, 本文旨在考虑如何合理地选取 K-means 算法的初始质心以提高算法的效率和性能, 因此就此问题不做深入讨论。

4.2 选取聚类初始质心

K-means 算法对初始质心的选取是敏感的, 初始质心的选取将影响算法的迭代次数和算法能否近似得到全局最优 SSE, 进而影响算法的执行效率和聚类精度。

基于动态网格划分的中心点选取策略为: 根据数据集的网格划分无向图选取 K-Means 算法的初始中心, 其选取依据如下:

- (1) 对于只有一个顶点的连通子图, 规定初始质心为该顶点的中心;
- (2) 对于包含多个顶点的连通图, 规定初始质心为连通图顶点的质心, 计算公式如下:

$$X_i = \frac{\sum_{i=1}^c g_i \cdot \text{Count} * g_i \cdot \text{Center} \cdot x_i}{\sum_{i=1}^c g_i \cdot \text{Count}} \quad (1)$$

式中, X_i 为初始质心在数据集第 i 维上的坐标, g_i 为网格 g_i 中包含数据点个数, $g_i \cdot \text{Center} \cdot x_i$ 为网格 g_i 的中心在数据集第 i 维上的坐标, c 为该连通子图所包含的顶点个数。

5 算法描述

基于动态网格划分技术的 K-means 中心点选取算法主要包括动态网格划分、构造网格划分无向图和初始化聚类质心。

算法 AGAR:

输入: 数据集, 网格密度相连阈值 threshold

输出: K-means 算法初始聚类质心

1. 动态网格划分

1.1 读取数据点, 划分网格

1.2 保存网格内数据点总数和网格相交信息

2. 构造网格划分无向图

2.1 读取网格密度可达信息

2.2 生成网格划分无向图并存储

3. 计算初始质心

3.1 遍历网格划分示意图, 得到连通子图

3.2 为每一个连通子图选取质心

读取网格密度可达信息的步骤可以采取深度优先搜索或者广度优先搜索算法, 对于无向图的存储通常采用索引结构以提高算法的性能。对于步骤 3, 在寻找连通自动的过程可以同时记录顶点信息, 在得到一个子图后即可直接算得初始质心, 且可减少一遍对图的遍历。

6 算法分析与实验结果

6.1 时间复杂度分析

设数据集大小为 n , 维度为 d , 生成网格的个数为 m , 读取数据点的时间复杂度为 $O(n)$, 判断数据点是否在一个网格中需要执行 d 次, 所以算法在第一阶段的平均时间复杂度为 $O(n * d)$; 在算法的第二阶段, 按深度或者是广度优先搜索的原则查找网格是否密度相连的信息, 则算法的平均时间复杂度为 $O(m + e)$, 假设有 e 对网格密度相连; 在第三阶段, 算法的平均时间复杂度为 $O(m)$ 。因此, 算法的整体时间复杂度为 $O(n)$ 。从时间复杂度上来看, 在执行 K-means 算法之前执行本算法并不会增加 K-means 算法的时间复杂度。

6.2 实验结果分析

实验环境为: Intel (R) Pentium (R) 4 CPU 2.40 GHz, 内存: 256 MB, 操作系统: Microsoft Windows XP Professional, 算法编写语言为 JAVA, JDK 1.5。

为了说明算法的有效性, 就 K-means 算法的迭代次数、最终 SSE 和聚类效果做了相关实验。

图 4 中“系列 1”表示在执行 K-means 算法之前, 采用本算法选择了初始质心后的算法性能, “系列 2”表示随机为 K-means 算法选择初始质心的算法性能。图 4(a) 反映的是 K-means 算法的迭代次数, 图 4(b) 反映的是最终 SSE 值的比较。X 轴为数据集大小。

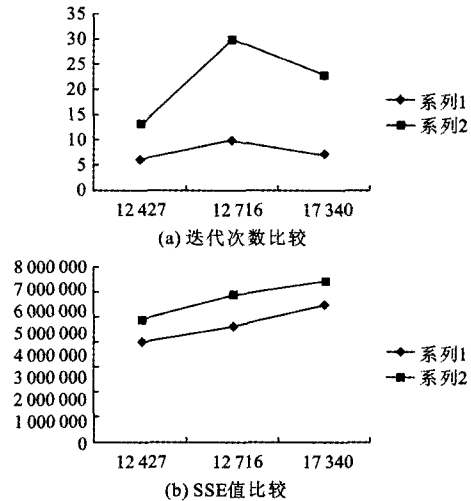


图 4 算法性能比较

对于图 5(a) 中的数据集, 数据集大小为 12 736, 包含 6 个聚类, 如果在执行 K-means 算法之前首先采用本算法选择初始质心点, 一次执行的聚类效果如图 5(b) 所示, 图 5(c) 则是通过为 K-means 算法随机选择质心点可能出现的聚类效果。



图 5 K-means 选择初始质心

通过实验可以说明, 在执行 K-means 算法之前采用 AGAR 算法初始化质心, 能够较大幅度地减少聚类的迭代次数和提高聚类的准确度。

7 结束语

本文基于动态网格划分技术提出了一种寻求

K-means 算法初始质心的新算法,该算法在不增加 K-means 算法时间复杂度的情况下减少了 K-means 算法的迭代次数,使 SSE 近似全局最优,并提高了聚类精度.

参考文献:

- [1] Baozhi QIU, Chang ZHANG, Junyi SHEN. Clustering based on overlapping grid[C]// The 10th China Conference on Machine Learning. Hai kou: IEEE, 2006.
- [2] Jiawei Han Micheline Kamber, Jian Pei. Data mining concepts and techniques[M]. Morgan kaufmann, 2001: 223-254.
- [3] Guha S, Rastogi R, Shim K. Cure: an efficient algorithm for clustering large databases[C]// Proceedings of ACM SIGMOD International Conference on Management of Data. NY, US: ACM, 1998: 73-84.

- [4] Martin Ester, Hans-Peter Kriegel, Jorg Sander, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining(KDD-96). OR, Protland: IEEE, 1996: 226-231.
- [5] Gilman M. White paper: nuggets and data mining. data mining technologies ins [C]// Improving inference through conceptual clustering proc AAAI Conf Seattle, WA: IEEE, 1987: 461-465.

作者简介:

张真女, (1982—), 硕士研究生, 讲师. 研究方向为数据挖掘.

任贺宇男, (1981—), 硕士研究生, 讲师. 研究方向为计算机网络、信息安全技术.

(上接第 100 页)

- [4] Vieira MAM, Vieira LFM, Ruiz LB, et al. Scheduling nodes in wireless sensor networks: a voronoi approach [J]. IEEE Computer Society, 2003: 423-429.
- [5] Tian Di, Georganas N D. A coverage-preserving node scheduling scheme for large wireless sensor networks [C]// Proc of ACM Workshop on Wireless Sensor Networks and Applications. New York, USA: ACM Press, 2002: 124-128.
- [6] Huang Chi-Fu, Tseng Yu-Chee. A survey of solutions to the coverage problems in wireless sensor network s

[J]. Journal of Internet Technology, Special Issue on Wireless Ad Hoc and Sensor Networks, 2004, 12(3): 2356-2359.

作者简介:

洪刚男, (1976—) 讲师, 硕士. 研究方向为计算机应用、传感器网络.

汤宝平男, (1971—), 硕士研究生. 研究方向为无线传感器网络方向的研究工作.