

一种密度和划分结合的聚类算法

王玉雷, 李玲娟

(南京邮电大学 计算机学院, 江苏 南京 210003)

摘要:基于密度的 DBSCAN 聚类算法和基于划分的 k -means 聚类算法各有其优缺点。文中在 k -means 和 DBSCAN 聚类算法的基础上,以减少聚类算法对参数和数据点输入顺序的敏感性,发现任意形状的簇,提高聚类挖掘的质量为目标,提出了一种密度和划分结合的聚类算法—DDCA。该算法首先计算数据点的密度,以密度不小于给定阈值的中心点以及在其密度范围内的点组合成各个基本簇;再依据两个簇中心点之间的距离合并基本簇;最后把没有划分到任意簇的点划分到与其距离最近的簇中。理论分析和基于 KDD CUP 99 数据集的实验结果表明,提出的 DDCA 算法能够发现任意形状的簇,对数据点的输入顺序以及参数不敏感,在时间开销仅略有增加的情况下可获得更高的聚类准确度,其总体性能优于 k -means。

关键词:数据挖掘; k -means; DBSCAN; 聚类; 密度; 划分

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2015)09-0053-04

doi: 10.3969/j.issn.1673-629X.2015.09.011

A Clustering Algorithm of Combination of Density and Division

WANG Yu-lei, LI Ling-juan

(College of Computer, Nanjing University of Posts and Telecommunications,
Nanjing 210003, China)

Abstract: Both the density-based clustering algorithm DBSCAN and the division-based clustering algorithm k -means have their advantages and disadvantages. In order to reduce the sensitivity of clustering algorithm to the parameters and the input order of the data points, finding clusters of arbitrary shape and improving the quality of clustering mining, on the basis of DBSCAN and k -means clustering algorithm, propose a clustering algorithm combined density and division, named DDCA. This algorithm firstly calculates the density of data points, then combines the center point which has a density greater than a given threshold value and others point which in the density range of the center point to build basic clusters. Then merge two basic clusters according to the distance between their center points. Finally, divide point which is not belong to any cluster into its nearest cluster. Theoretical analysis and experimental results on KDD CUP 99 dataset show that this algorithm can find clusters of arbitrary shape, and is not sensitive to parameters and the input order of data points. It can get higher clustering accuracy with a little additional time cost. Its overall performance is better than k -means clustering algorithm.

Key words: data mining; k -means; DBSCAN; clustering; density; division

0 引言

数据挖掘(Data Mining)是指从大量的、不确定的、有噪声的数据中发现隐含的、未知的、有价值的信息的过程^[1]。聚类挖掘是指将给定数据集化分成若干簇,簇中数据对象之间相似,而不同簇之间的数据对象差异较大^[2]。聚类挖掘算法主要分为四大类:

(1)基于划分的聚类算法。按照数据对象之间的相似度,将给定的数据集划分成 k 个组,每一个组称为

一个聚类或簇。具有代表性的算法有 k -means^[3]、 k -medoids^[4]。

(2)基于层次的聚类算法。这种方法对给定的数据集进行层次似的合并或分解,直到满足某种条件为止。具体又可分为“自底向上”和“自顶向下”两种方案。具有代表性的算法有 BIRCH^[5]、CURE^[6]。

(3)基于密度的聚类算法。在数据集中,低密度的区域将高密度的区域隔离开,各个独立的高密度区

收稿日期:2014-06-17

修回日期:2014-10-15

网络出版时间:2015-08-26

基金项目:国家“973”重点基础研究发展计划项目(2011CB302903)

作者简介:王玉雷(1989-),男,硕士研究生,研究方向为数据挖掘、大数据、云计算;李玲娟,教授,通讯作者,研究方向为数据挖掘、信息安全、分布式计算。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20150826.1535.002.html>

域形成簇。具有代表性的算法有 DBSCAN^[7]、OPTICS^[8]。

(4) 基于网格的聚类算法。首先将数据空间划分成为有限个单元 (cell) 的网格结构, 所有的处理都是以单个网格为对象的。这样处理的一个突出优点就是处理速度很快, 与数据库中记录的多少无关, 只与把数据空间分为多少个单元有关。具有代表性的算法有 STING^[9]。

这些聚类算法都有各自的优缺点, 比如, 基于划分的 k -means 聚类算法具有算法简单、速度快的优点, 但它需要事先确定簇的个数 k , 且只能发现超球状的簇, 对初始点的选择很敏感; 基于密度的聚类算法 DBSCAN 具有可以发现若干任意形状簇的优点, 然而它对高密度区域的搜索是从随机对象开始的, 另外 DBSCAN 算法在处理密度差异不明显的数据集时效果不明显。

在分析总结 k -means 和 DBSCAN 算法的基础上, 文中提出了一种密度和划分相结合的聚类算法。它汲取 k -means 和 DBSCAN 算法的优点, 同时对 k -means 和 DBSCAN 的缺点进行改进。首先计算数据点的密度, 把密度不小于给定阈值的中心点以及在此中心点密度范围内的点组成一个基本簇; 其次合并满足条件的簇; 最后把没有划分到任意簇中的点划分到与其距离最近的簇中。

1 k -means 与 DBSCAN 聚类算法

1.1 k -means 聚类算法

k -means 聚类算法需事先确定 k 的大小, 随机选取 k 个初始中心, 然后把余下的对象分别划分到与其距离最近的初始中心所在的簇中; 计算新的中心点, 不断重复上述过程, 直到目标函数收敛, 达到较优的聚类效果为止。

算法的基本步骤如下:

输入: 簇的个数 k 以及数据集 D ;

输出: 满足条件的 k 个簇。

(1) 从数据集 D 中随机选取 k 个对象作为初始聚类中心;

(2) repeat;

(3) 计算对象与各个中心点的距离, 把对象划分到与其最近的中心点所在的簇中;

(4) 计算每个簇的均值, 作为新的聚类中心;

(5) until 簇中心点不再发生变化为止。

k -means 算法需要事先确定 k 的大小以及初始聚类中心, 只能发现超球状的簇, 对初始中心非常敏感。对 k -means 算法的改进^[10-11], 主要从确定 k 的大小以及 k 个初始中心的选择两方面进行。

1.2 DBSCAN 聚类算法

DBSCAN 聚类算法将簇定义为密度相连的点的最大集合, 它能把高密度的区域划分为簇, 并在含有噪声空间的数据库中发现任意形状的簇。算法的基本步骤如下:

输入: 半径 Eps, 最小数目 minPts, 数据集 D ;

输出: 满足条件的簇。

(1) repeat;

(2) 从数据库中取出一个未处理的数据点 p ;

(3) 如果 p 是核心点, 找出所有从该点密度可达的对象, 形成一个簇;

(4) 否则 p 点作为边缘点 (非核心对象), 跳出本次循环, 寻找下一个点;

(5) until 所有的点都被处理。

DBSCAN 的优点在于可以发现任意形状的簇, 能够识别噪声, 对数据的输入顺序不敏感; 其缺点是对输入参数敏感, 对密度差异不明显的数据处理不理想。

2 密度和划分结合的聚类算法

单纯地对 k -means 或 DBSCAN 聚类算法进行改进并不能弥补算法自身的不足, 因此有学者提出将密度和 k -means 进行结合^[12-13] 的聚类算法, 但他们在算法中引入了新的参数, 导致算法依然对参数敏感, 没有达到预期的效果。

为了克服 k -means 和 DBSCAN 聚类算法的缺点, 降低参数的影响, 提高聚类质量, 文中综合 k -means 和 DBSCAN 的优点, 提出一种密度和划分相结合的聚类算法 (Density and Division based Clustering Algorithm, DDCA)。

2.1 相关定义

与密度相关的定义如下:

(1) 距离: 用 $\text{dist}(p, q)$ 表示对象 p 和 q 之间的距离或相异度, 计算公式如下:

$$\text{dist}(p, q) = \sqrt{\sum_{i=1}^m (p^i - q^i)^2} \quad (1)$$

其中, m 表示数据的维度; $\text{dist}(p, C^i)$ 表示对象 p 和第 i 个簇的中心点之间的距离; $\text{dist}(C^i, C^j)$ 表示两个簇中心点之间的距离。

(2) Eps 邻域: 对象 p 的 Eps 邻域, 即以 p 为中心, Eps 为半径的超球状区域内数据点的集合, 计算公式如下:

$$N_{\text{Eps}}(p) = \{q \in D \mid 0 \leq \text{dist}(p, q) \leq \text{Eps}\} \quad (2)$$

其中, D 表示数据的集合; 用 $|N_{\text{Eps}}(p)|$ 表示对象 p 的 Eps 领域内包含对象的个数。Eps 参数的大小为所有点之间的距离的均值。

(3) 密度: 用 $\text{density}(p)$ 表示对象 p 的密度, 对象

p 的密度定义为 p 的 E_{ps} 邻域内对象的数量,其密度大小为 $|N_{Eps}(p)|$ 。

2.2 算法流程与描述

文中所设计的密度和划分相结合的聚类算法—DDCA 的主体流程如图 1 所示。

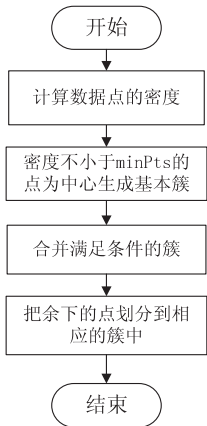


图1 DDCA 的流程

算法可具体描述如下：
输入：最小数目 minPts，数据集合 D ；
输出：若干个任意形状的簇。
(1)运用公式(1)、(2)依次计算数据点 p 的 E_{ps} 邻域 $N_{Eps}(p)$ ，将密度 $density(p)$ ，即 $|N_{Eps}(p)|$ 不小于 minPts 的各点作为初始中心点，将各中心点以及在其密度范围内的点组合成一个基本簇，同时计算簇中离中心点距离最远的点 q 与中心点之间的距离 $dist(q, C^i)$ ，并用 d^i 表示，其中 C^i 表示第 i 个簇；
(2)对于任意两个簇 i 和 $j(i \neq j)$ ，如果两个簇中心点的距离 $dist(C^i, C^j)$ 不大于合并阈值 $minD$ ， $minD$ 的取值为 $2 \times \max(D^i, D^j)$ ，则合并两个簇，更新簇中离中心点距离最远的点 q 与中心点的距离 d^i ；
(3)重复步骤(2)中的操作，直到没有簇被合并为止；
(4)把没有划分到任意簇中的点划分到与其距离最近的簇中。

这种密度和划分相结合的聚类算法能够发现若干任意形状的簇，不需要事先确定 k 的大小，对数据的输入顺序及参数不敏感；根据两个簇自身距离属性来确定合并阈值，合并阈值具有局部性，降低了全局参数 E_{ps} 和 $minPts$ 的影响。

3 实验

3.1 实验数据集介绍

采用 KDD CUP 99^[14] 高维混合属性数据集的子数据集 kddcup.data_10_percent 作为实验数据。10% 的数据子集中，包含 DoS、R2L、U2R 和 Probing 四大类异常数据，23 种攻击类型。总共 494 021 条数据，其中正

常数据 97 278 条，DoS 类型 391 458 条，R2L 类型 1 126 条，U2R 类型 52 条，Probing 类型 4 107 条。每条数据包含 41 个特征属性和 1 个类别属性，各属性之间用逗号分隔。

3.2 实验结果分析

实验平台：Windows 7 操作系统，Pentium(R) Dual-Core 2.0 G CPU，Matlab 2010b 编程软件。

从 10% 的数据子集中抽取 12 万条记录，构成 8 个数据子集进行实验，前 4 个是训练数据集，后 4 个是测试数据集；每个训练数据集包含 2.5 万条记录，每个测试数据集包含 0.5 万条记录。前 3 个训练数据集分别只包含 Dos、Probing 和 R2L 类型异常数据，第 4 个训练数据集包含全部 4 种类型的异常数据。将 DDCA 聚类算法和 k -means 聚类算法在运行时间、检测率 (True Positive Rate, TPR) 和误检率 (False Positive Rate, FPR) 三个方面进行比较验证。其中：

$$TPR = (\text{正确识别的异常数据数} / \text{异常数据数}) * 100\%$$
$$FPR = (\text{被认为异常的正常数据数} / \text{正常数据数}) * 100\%$$

实验结果如图 2 和图 3 所示。

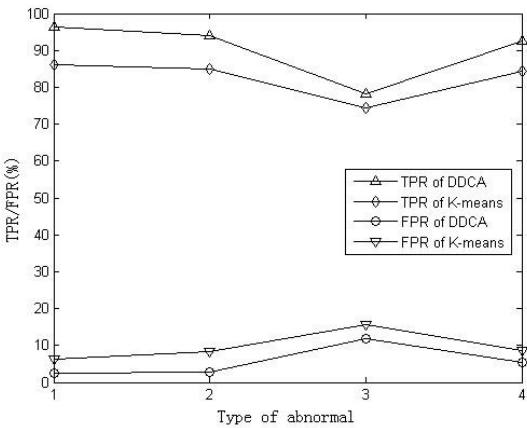


图2 两种算法 TPR 和 FPR 的比较

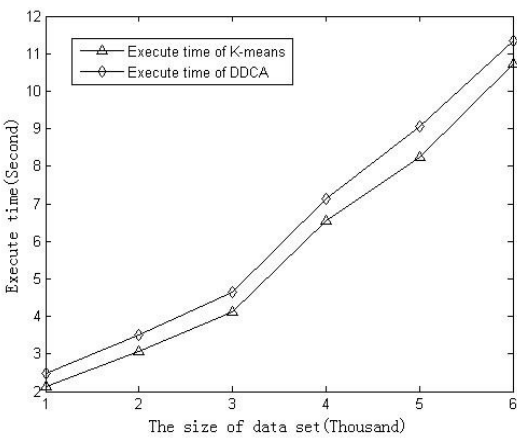


图3 两种算法运行时间比较

图2中 X 轴坐标的1、2和3分别代表数据集中只包含Dos、Probing和R2L类型异常数据,4表示数据集包含全部4种类型的异常数据。

图3中的实验数据集是从第4个训练数据集中抽取的6个数据集,6个数据集的大小分别从1千条记录递增至6千条记录。

从图2中可以看出,DDCA在TPR和FPR上都优于 k -means。从图3中可以看出,DDCA运行时间稍大于 k -means,且二者的运行时间差在可接受的时间范围内。但是 k -means需要事先确定 k 值的大小,而且对 k 个初始中心点的选择比较敏感。

总体来说,DDCA聚类算法对数据集的处理能力优于 k -means。

4 结束语

文中提出的DDCA算法结合了密度聚类算法和划分聚类算法的优点,利用簇自身的距离属性来判断是否进行簇的合并,不需要事先确定 k 的大小,同时提高了算法的准确度。算法能够发现若干任意形状的簇,并且算法对参数Eps和minPts不敏感。但是在合并簇的判断条件上还需进一步的改进,准确度有待于进一步提高。

DDCA聚类算法具有数据适应性、通用性和无需人工干预的特点,可应用于网络入侵检测、电信客户分类、网络话题识别跟踪等实际场景中,增强系统的自动化,并且可以提高系统对数据的适应能力。

参考文献:

- [1] Tan P N, Steinbach M. 数据挖掘导论[M]. 范明, 范宏建, 译. 北京: 人民邮电出版社, 2006.
- [2] 胡庆林, 叶念渝, 朱明富. 数据挖掘中聚类算法的综述[J].

(上接第52页)

- [7] 成谢锋, 姜炜, 刘子山. 一种新的人体运动强度检测方法的研究[J]. 仪器仪表学报, 2013, 34(5): 1153-1159.
- [8] 郑伟谋, 郝柏林. 实用符号动力学[J]. 物理学进展, 1990, 10(3): 316-373.
- [9] Xia J N, Shang P J, Wang J, et al. Classifying of financial time series based on multiscale entropy and multiscale time irreversibility[J]. Physica A: Statistical Mechanics and Its Application, 2014, 400: 151-158.
- [10] 马斌荣, 陈卉. 医学科研中的统计方法[M]. 第3版. 北


计算机与数字工程, 2007, 35(2): 17-20.

- [3] Kanungo T, Mount D M, Netanyahu N S, et al. An efficient k -means clustering algorithm: analysis and implementation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 881-892.
- [4] Park Hae-Sang, Jun Chi-Hyuck. A simple and fast algorithm for K -medoids clustering[J]. Expert Systems with Applications, 2009, 36(2): 3336-3341.
- [5] Zhang Tian, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases[C]//Proc of ACM SIGMOD. [s. l.]: Association for Computing Machinery, 1996: 103-114.
- [6] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases[J]. Information Systems, 1998, 26(1): 35-58.
- [7] 冯少荣, 肖文俊. DBSCAN聚类算法的研究与改进[J]. 中国矿业大学学报, 2008, 37(1): 105-111.
- [8] Ankerst Z M, Breunig M M, Kriegel Hans-Peter, et al. OPTICS: ordering points to identify the clustering structure[C]//Proc of ACM SIGMOD. [s. l.]: Association for Computing Machinery, 1999: 49-60.
- [9] Wang W, Yang J, Muntz R. STING: a statistical information grid approach to spatial data mining[C]//Proc of VLDB. [s. l.]: [s. n.], 1997: 186-195.
- [10] 邓海, 覃华, 孙欣. 一种优化初始中心的 K -means聚类算法[J]. 计算机技术与发展, 2013, 23(11): 42-45.
- [11] 谢秀华, 李陶深. 一种基于改进PSO的 K -means优化聚类算法[J]. 计算机技术与发展, 2014, 24(2): 34-38.
- [12] 王晶, 夏鲁宁, 荆继武. 一种基于密度最大值的聚类算法[J]. 中国科学院研究生院学报, 2009, 26(4): 539-548.
- [13] 张琳, 陈燕, 汲业, 等. 一种基于密度的 K -means算法研究[J]. 计算机应用研究, 2011, 28(11): 4071-4073.
- [14] 张新有, 曾华荣, 贾磊. 入侵检测数据集KDD CUP99研究[J]. 计算机工程与设计, 2010, 31(22): 4809-4812.

京: 科学出版社, 2005: 152-157.

- [11] Landis J R, Koch G G. The measurement of observer agreement for categorical data[J]. Biometrics, 1977, 33(1): 159-174.
- [12] 陈泓. 心血管系统仿真模型的研究[J]. 计算机技术与发展, 2014, 24(11): 222-225.
- [13] Wieser M, Gisler S, Sarabadani A, et al. Cardiovascular control and stabilization via inclination and mobilization during bed rest[J]. Medical and Biological Engineering and Computing, 2014, 52: 53-64.

一种密度和划分结合的聚类算法

作者：[王玉雷](#)，[李玲娟](#)，[WANG Yu-lei](#)，[LI Ling-juan](#)
作者单位：[南京邮电大学 计算机学院, 江苏 南京, 210003](#)
刊名：[计算机技术与发展](#)
英文刊名：[Computer Technology and Development](#)
年，卷(期)：2015(9)

引用本文格式：[王玉雷](#)，[李玲娟](#)，[WANG Yu-lei](#)，[LI Ling-juan](#) [一种密度和划分结合的聚类算法](#)[期刊论文]-[计算机技术与发展](#) 2015(9)