山东大学学报(工学版) 2012, 42(4) 35-40 DOI: ISSN: 0412-1961 CN: 21-1139



机器学习与数据挖掘

# 一种基于近似EMD的DBSCAN改进算法

张宏兵[1],陆建峰[1*],汤九斌[2]

1.南京理工大学计算机科学技术学院, 江苏 南京 210094; 2.中国电信江苏公司, 江苏 南京 210037

摘要:

DBSCAN(density based spatial clustering of applications with noise)算法是基于密度的经典聚类算法,但是该算法应用于高维数据时,常用距离函数不能很好地反映出数据点之间的关系, 从而可能导致聚类簇不够精确。如果能在高维空间中采用合适的距离度量,将会改善聚类结果。针对上述问题,提出利用近似EMD(earth mover' s distance,堆土机距离)作为距离测度,通过迭代搜索的方法找出所有直接密度可达对象实现聚类。实验结果表明:在高维文本数据的聚类中,和原来算法相比,改进算法的正确率提高了6%,两者在时间上相差不大;而对低维的Iris数据,改进算法通过EMD改善了实体间的相似性度量,减少了划分为噪声点的数据点个数,平均正确率提高了10%。实验结果表明了改进算法对高维数据的有效性,并可以改善聚类性能。

关键词: 聚类 DBSCAN算法 近似EMD 高维数据

# An improved DBSCAN algorithm based on the approximate EMD

ZHANG Hong-bing1, LU Jian-feng1*, TANG Jiu-bin2

1. School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China;
2. Jiangsu Corporation of China Telecom, Nanjing 210037, China

Abstract:

The DBSCAN algorithm is one of the classic clustering algorithms based on the density. When this algorithm was applied to high-dimensional data, the distance measures in common use could not reflect the relationships between instances well, which would lead to the inaccurate clustering. If appropriate distance measures were adopted in high-dimensional space, the clustering result would be improved. To solve the above problem, the approximate EMD (earth mover' s distance) instead of the common distance was used as the distance measure, and the clustering was achieved by finding all density reachable objects with the method of iterative search. The experimental results showed that the performance of improved algorithm was 6% higher than that of the original algorithm for the high-dimensional text clustering, while there is no obvious difference in time cost. For low-dimensional Iris data, the proposed algorithm could improve the similarity measure between the instances, reduce the number of data points classified as noise points, and boot the performance with 10%. The experimental results also indicated that the proposed algorithm could reveal its effectiveness for high-dimensional data, and could improve the clustering performance.

Keywords: clustering DBSCAN algorithm approximate EMD high-dimensional data

收稿日期 2012-05-06 修回日期 网络版发布日期

DOI:

基金项目:

江苏省自然基金资助项目(BK2009489);江苏省青蓝工程资助项目

通讯作者: 陆建峰(1969- ),男,江苏南京人,教授,博士生导师,主要研究方向为人工智能和图像图形技术等. E-mail:lujf@njust.edu.cn

作者简介: 张宏兵(1987- ),男,江苏东台人,硕士研究生,主要研究方向为文本挖掘. E-mail:iamzhanghongbing@126.com

作者Email: lujf@njust.edu.cn

PDF Preview

参考文献:

本刊中的类似文章

1．卜德云 张道强.自适应谱聚类算法研究[J]. 山东大学学报(工学版), 2009,39(5): 22-26

2．许延生,刘兴芳 .模糊聚类迭代模型在水资源承载能力评价中的应用[J]. 山东大学学报(工学版), 2007,37(3): 100-104

3．马志强,常发亮,田伟,赵瑶 .彩色图像中的人脸检测方法[J]. 山东大学学报(工学版), 2007,37(4): 19-22

4．牛新生,叶华,王亮 .彩色图像中的人脸检测方法[J]. 山东大学学报(工学版), 2007,37(4): 0-0

5．赵洪国,张焕水,张承慧 .基于RBF神经网络的交通流量预测算法研究[J]. 山东大学学报(工学版), 2007,37(4): 0-0

6．朱文兴,龙艳萍,,贾磊 .基于RBF神经网络的交通流量预测算法[J]. 山东大学学报(工学版), 2007,37(4): 23-27

7．王耘,穆勇,刘庆红 .基于灰关联分析的模糊聚类最优划分判定模型[J]. 山东大学学报(工学版), 2006,36(2): 86-89

8．孙宇清,赵锐,姚青,史斌,刘佳 .一种基于网格的障碍约束下空间聚类算法[J]. 山东大学学报(工学版), 2006,36(3): 86-90

9．雷小锋1,庄伟1,程宇1,丁世飞1,谢昆青2.OPHCLUS:基于序关系保持的层次聚类算法[J]. 山东大学学报(工学版), 2010,40(5): 48-55

10．杨立才,赵莉娜,吴晓晴 .基于蚁群算法的模糊C均值聚类医学图像分割[J]. 山东大学学报(工学版), 2007,37(3): 51-54