系统工程与电子技术 2010, 32(12) 2721-2724 DOI: 10.3969/j.issn.1001-506X.2010.12.46 ISSN: 1001-506X CN: 11-2422/TN

软件、算法与仿真

一种基于马尔可夫链的高维离群点挖掘算法

唐志刚1，2, 杨炳儒1, 杨珺1

1. 北京科技大学信息工程学院， 北京 100083；
2. 南华大学数理学院， 湖南 衡阳 421001

摘要：

提出了一种基于马尔可夫链的离群点检测(outlier detection algorithms based on Markov chain， MRKFOD)算法。该算法把基本数据集看作一个加权无向图，数据集中的每个数据表示一个节点，用每条加权边表示节点之间的相似度；形成一个邻接矩阵，把邻接矩阵当作马尔可夫链中的概率转移矩阵；寻求概率转移矩阵的主要特征向量；把每个节点的主要特征向量值作为每个数据的离群度。实验结果表明，该算法与其他高维离群点挖掘算法相比，在效率及有效处理的维数方面均有显著提高。

关键词： 数据挖掘 离群点 高维数据集 马尔可夫链 加权无向图

New outlier detection algorithm based on Markov chain

TANG Zhi-gang1,2, YANG Bing-ru1, YANG Jun1

1. School of Information Engineering, Univ. of Science and Technology Beijing, Beijing 100083, China;
2. School of Mathematics and Physics， Univ.of South China, Hengyang 421001, China

Abstract:

An outlier detection algorithm based on Markov chain (MRKFOD algorithm) is presented. First, the basic data set is regarded as a weighted undirected graph, in which each datum represents a node, and each weighted edge denotes the similarity between nodes; so it forms an adjacency matrix, and then the adjacency matrix is regarded as a probability transition matrix in Markov chain. Secondly, the algorithm seeks the main feature vector of the probability transition matrix. Finally, the main feature vector of each node is looked upon as the outlier degree of each datum. The experimental results show that both the efficiency of MRKFOD algorithm and the maximum number of dimensions processed are obviously improved compared with other high- dimensional outlier mining algorithms.

Keywords: data mining outlier high dimensional data set Markov chain weighted undirected graph

收稿日期 修回日期 网络版发布日期

DOI: 10.3969/j.issn.1001-506X.2010.12.46

基金项目:

通讯作者:

作者简介:

作者Email:

参考文献：

本刊中的类似文章

1．齐照辉, 刘雪梅, 梁伟 .基于MCMC的导弹主动段突防仿真及灵敏度分析[J]. 系统工程与电子技术, 2010,32(4): 803-806

2．孙宇航，孙应飞.基于网络日志的数据挖掘预处理改进方法[J]. 系统工程与电子技术, 2009,31(12): 2994-2997

3．曾华，吴耀华，黄顺亮.非均匀类簇密度聚类的多粒度自学习算法[J]. 系统工程与电子技术, 2010,32(8): 1760-1765

4．王昊天，石健.基于可用度模型的故障预测与健康管理方法[J]. 系统工程与电子技术, 2010,32(12): 2584-2589

5．翟云,杨炳儒,曲武,隋海峰.基于新型集成分类器的非平衡数据分类关键问题研究[J]. 系统工程与电子技术, 2011,33(1): 196-0201