基于密度的孤立点检测算法分析

摘 要

近年来，随着科技的快速发展，信息技术发展的越来越迅速，信息数据的爆炸式增长已到了令人咂舌的地步，各式各样的商业活动产生了巨大的数据集，我们现今研究的状况是怎样从已知的不完整的数据中提取隐含在内的、人们无法从表象看到的数据。本文关心的重点是数据挖掘中孤立点的意义。孤立点(离群点)被定义为与其他观测值偏差过大的观测值，识别异常值可以发现有用和有意义的知识。近几十年来，专家们通过对孤立点的检测行了广泛的研究，提出了许多传统的分离点探测算法，包括距离检测法、密度检测法、空间检测法等。

本文主要针对基于密度的孤立点算法进行研究与分析，具体而言，本文所做工作包括以下几个方面：分析研究孤立点的背景和意义,探讨国内外研究背景及现状；该系统对传统的孤立点挖掘算法进行了综合分析，如基于统计分布、基于距离、基于密度和基于距离的孤立点检测方法；以及不同类型代表性算法的介绍与理论分析；重点是检测和分析基于度的 LOF 和 DBSCAN 算法，这些算法由模拟数据集和实际数据集的结果表示；在已有实验基础上,比较 LOF 与 DBSCAN 的实验结果,对其处理数据集的能力进行认定,分析验证算法的优点及不足之处,并且提出对算法的改进措施。

关键词：数据挖掘；孤立点检测；密度；LOF 方法；DBSCAN 聚类

Abstract

In recent years, according to increasingly science and technology and the increasingly developming of information technology, the explosive growth of information data has reached a surprising level, and various business activities have produced huge data sets, such as sales business, hospital information registration, stock trading records, product trading volume and so on. And how to extract the implicit data from a large number of noisy, incomplete and random data that people cannot see from the appearance has become the key of this paper. This paper focuses on the significance of outliers in data mining. Outliers (outliers) are defined as observations that deviate too much from other observations, and identifying outliers can lead to the discovery of useful and meaningful knowledge. Outlier detection has been widely studied in the past decades. Then, experts and scholars proposed many classical outlier detection algorithms, including distance-based detection method, density-based detection method, and space-based detection method. This paper mainly studies and analyzes the density-based outlier algorithm. Specifically, the work of this paper includes the following aspects:

The context and importance of Isolated point research will be analysed and the research context and the current situation at home and abroad will be discussed;This paper systematically analyzes various classical outlier mining algorithms, and expounds the detection methods based on statistical distribution, deviation, density, space and distance. All kinds of representative algorithms are evaluated and compared;The focus is to detect and analyze LOF and DBSCAN

algorithms based on outlier density, which are represented by the results of simulated data sets and actual data sets;On the basis of the existing experiments, the experiment results of LOF and DBSCAN were compared to identify its ability to process data sets, and the advantages and disadvantages of the algorithm were proposed, as well as the improvement measures for the algorithm

Keywords: Data mining; Outlier detection; Based on density; LOF method; DBSCAN clustering

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：
https://d.book118.com/578133142036006125