学校编码：10384

学号：24320111152278

厦门大学

硕士学位论文

数据挖掘中的离群点检测算法研究

Research on Outlier Detection Algorithm in Data Mining

胡婷婷

指导教师：陈海山

专业名称：计算机软件与理论

答辩日期：2014年5月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　月　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文(包括纸质版和电子版)，允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于:

(　　)1.经厦门大学保密委员会审查核定的保密学位论文，于　年　月　日解密，解密后适用上述授权。

(　　)2.不保密，适用上述授权。

(请在以上相应括号内打“✓”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。)

声明人(签名):

年　　月　　日

摘　要

离群点检测是数据挖掘中的一个分支，它的任务是识别其特征显著不同于其他数据的观测值。在我们平常的社会生活和自然界中，大部分的事件和对象，都是很寻常或者是平凡的。但是我们也不能因此忽视，在其中也有很多不寻常或者不平凡的对象存在的可能性。这些对象的事件背后可能蕴含着更大的研究价值，有着广阔的应用前景。因此，离群点检测是一个非常有意义的研究方向。

目前，研究者们已经提出了很多离群点检测方法，包括基于统计的离群点检测方法、基于频率的离群点检测方法、基于深度的离群点检测方法、基于距离的离群点检测方法和基于密度的离群点检测方法等。本文分析了离群点检测的研究背景、意义和国内外研究现状，研究基于距离的离群点检测方法和基于频率的离群点检测方法，并改进了传统的离群点检测方法。

属性通常可以分为两类，包括数值属性以及分类属性。本文详细分析了两种属性的区别，并做了以下工作：

针对数值数据，对传统的基于距离的检测算法进行改进。传统的基于距离的检测算法输入参数多，而且算法对参数比较敏感，因此选择基于平均距离的离群点检测算法。针对这种算法计算量大，在大数据集中不适用的问题，根据如果数据对象r邻域内数据的个数达到k个以上就不是离群点的规则剪去部分非离群点，通过聚类，将类别按照密度进行排序，先选择密度小的类别开始检测，离群度阈值可以迅速增大，利用剪枝规则再次剪枝。这样可以大大减小运算时间。

针对分类数据，分析了基于距离的算法的不足，介绍了针对分类数据的常用的离群点检测方法，包括基于频率的方法和基于信息熵的方法。指出基于频率的AVF算法的不足，提出改进的基于频率的检测算法。通过常用的针对分类属性数据的k-modes聚类算法对数据集进行聚类，去除相似度较高的对象，再通过基于频率的方法进行检测，以达到更好的检测效果。

关键词：离群点检测；平均距离；频率

Abstract

Outlier detection is a branch of data mining. Its task is to identify the observations whose characteristics are significantly different from other data. In field of nature, human society, or data sets, most of the events and objects are ordinary or usual. But there are also many unusual or extraordinary objects. Value may be behind these objects. Outlier detection has broad application prospects. So outlier detection is a very interesting research.

There are already a large number of methods of outlier detection, including method of statistic-based outlier detection, method of depth-based outlier detection, method of distance-based outlier detection and method of density-based outlier detection. In this paper, the background, significance and research status of outlier detection is introduced. The method of distance-based outlier detection and frequency-based method are analyzed. The paper analyzes the problems of traditional approach and improves the algorithm.

Attributes can usually be divided into two categories, including numerical attributes and categorical attributes. The paper analyzes the differences between the two attributes and does the following work:

For numeric data, the paper improves method of distance-based detection. The traditional distance-based detection algorithm has many parameters and is sensitive to the choice of parameters, so the average distance is chosen to detect outliers. This algorithm needs a lot of computations and is not suitable in the large data set. To solve the problem, some non-outliers are pruned by the rule that if the number of the data in the r-neighborhood is k or more than k it is not outlier. By clustering, clusters are sorted by the density of the clusters. The cluster whose density is low is firstly detected. The pruning threshold can increase quickly. Pruning rules are used again. This can greatly reduce the computing time.

For the categorical data, the paper analyzes the shortcomings of distance-based

method. The methods are introduced which are commonly used for categorical data including entropy-based method and frequency-based method. The paper points out the lack of frequency-based algorithm AVF and improves it. Data set is clustered by k-modes clustering algorithm which is used for categorical data to remove objects with high similarities, then frequency-based method is used to detect outliers in order to achieve better detection.

Keywords: Outlier Detection; Average Distance; Frequency

参考资料

[1]薛安荣.空间离群点挖掘技术的研究[D], 江苏大学,2008.

[2]Pang-Ning Tan, Michael Steinbach, Vipin Kumar. 数据挖掘导论[M]. 北京: 人民邮电出版社, 2006.

[3]H. Fan, O. Zaiane, A. Foss.J. Wu.Cooper. A Nonparametric Outlier Detection for Efficiently Discovering top-n Outliers from Engineering Data [C]. PAKDD Conference, 2006.

[4]R. Derrig. Insurance Fraud [J]. Journal of Risk and Insurance, 2002, 271-287.

[5]M. Deshpande, G. Karypis. Selective Markov Models for Predicting Web Page Accesses [C]. ACM Transactions on Internet Technology, 2004, 163 – 184.

[6]罗敏, 阴晓光, 张焕国等. 基于孤立点检测的入侵检测方法研究[J]. 计算机工程与应用, 2007, 43(13): 146-149.

[7]J. Laurikkala, M. Juholal, E. Kentala. Informal Identification of Outliers in Medical Data [C]. Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology, 2000, 20 – 24.

[8]H. Dutta, C. Giannella, K. Borne, H. Kargupta. Distributed top-k Outlier Detection in Astronomy Catalogs using the Demac System [C]. SDM Conference, 2007.

[9]D. Jackson, Y. Chen. Robust Principal Component Analysis and Outlier Detection with Ecological Data [J]. Environmentrics, 2004, 129-139.

[10]S. Ramaswamy, R. Rastogi, K.Shim. Efficient Algorithms for mining outliers from large data sets [C]. Proceedings of 2000 ACM-SIGMOD Intl. Conf. on Management of Data, 2000, 427-438.

[11]A.chaudhary, A. S. Szalay, A.W. Moore. Very fast outlier detection in large multidimensional data sets [C]. Proc. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2002.

[12]S.D.Bay, M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule[C]. Proceedings of the 9th Intl. Conf. on Knowledge Discovery and Data Mining, 2003, 29-38.

[13]Angiulli F , Pizzuti C. Outlier Mining in Large High Dimensional Data Sets [C]. IEEE Trans. Knowledge and Data Eng, 2005 , 2 (17) :203-215

[14]Aggarwal C C . Redesigning Distance Functions and Distance based Applications for High Dimensional Data [J]. SIGMOD Record Date, 2001 , 30 (1) :13218

[15]C.C.Aggarwal, P.S.Yu. Outlier detection for high dimensional data [C]. Proceedings of 2001 ACM-SIGMOD Intl. Conf. on Management of Data, 2001, 37-46.

[16]M.-L.Shyu, S.-C. Chen, K.Sarinnapakorn, L.Chang. A novel Anomaly Detection Scheme Based on Principal Component Classifier [C]. Proceedings of the 2003 IEEE Intl. Conf. On Data Mining, 2003, 353-365.

[17]Angiulli F, Basta S, Pizzuti C. Distancebased detection and Prediction of outlier [C]. IEEE Trans. Knowledge and Data Eng. ,2006 , 2 (18) : 1452160

[18]许龙飞,熊君丽. 基于粗糙集的高维空间离群点发现算法研究[J].计算机工程与应用,2004,40 (7):58-60.

[19]Kriegel HP, Schubert M, Zimek A. Angle-based outlier detection in high-dimensional data[C]. Proceedings of the Int ’l Conf. on Knowledge Discovery and Data Mining. China: Beijing,2008, 444-452.

[20]刘培奇,孙婧,段中兴. 高维空间中离群点检测算法研究[J].微电子学与计算机,2013,30(7):116-121.

[21]鞠可一,周德群,张玉强. 高维离群检测算法及其应用[J]. 系统工程,2008,26(11):68-71.

[22]S.Shekhar, C.-T.Lu, P.Zhang. A Unified Approach to Detecting Spatial Outliers [J]. GeopInformatica, 2003, 7(2):139-166.

[23]Shekhar S , Lu C T , Zhang P . Detecting Graph-based Spatial Outliers[J]. International Journal of Intelligent Data Analysis(IDA) , 2002 , 6 (5) :451-468.

[24]Lu C-T, Chen Dechang , Kou Yufeng. Detecting Spatial Outliers with Multiple Attributes[C]. Proceedings of the 15th International Conference on Tools with Artificial Intelligence. Sacramento ,2003 :122-128.

[25]文俊浩,吴中福,吴红艳. 空间孤立点检测[J]. 计算机科学,2006 , 33 (5) :185-187.

[26]薛安荣,鞠时光.基于空间约束的离群点挖掘[J]. 计算机科学,2007,34 (6) :207-210.
[27]Xue Anrong ,Ju Shiguang. Algorithm for Spatial Outlier Detection Based on Outlying Degree[C]. Proceedings of the WCICA 2006.Dalian ,12 (7) :6005-6009.
[28]Choy K. Outlier detection for stationary time series [J]. Journal of Statistical Planning and Inference, 2001, 99(2):111-127.
[29]Keogh E , Lonardi S , Chiu B . Finding surprising pat terns in a time series database in linear time and space[C].Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York , 2002 :550-556.
[30]Sun Pei , Chawla S , Arunasalam B. Mining for Outliers in Sequential Databases[C]. Proc. of the Sixth SIAM International Conference on Data Mining. Bethesda , 2006 : 94-105.
[31]张晓伟,谢强,陈伟. 基于划分和孤立点检测的审计证据获取研究[J]. 计算机应用研究,2009 , 26(7) :2495-2501.
[32]张继福,蒋义勇,胡立华. 基于概念格的天体光谱离群数据识别方法[J].自动化学报,2008, 34(9) :1060-1066.
[33]Maneesh K Singh, Narendra Ahuja. Mean-shift Segmentation with Wavelet-based Bandwidth Selection [C]. Proceedings of the 6th IEEE Workshop on Applications of Computer Vision (WACV ’02), 2002.
[34]吴国洋. GML时空离群点挖掘技术研究[D],江西理工大学,2011.
[35]毛国君,段立娟. 数据挖掘原理与算法[M]. 北京: 清华大学出版社, 2007.
[36]Karjalainen L P, SomaniM, C Porter. Regression and Solute Drag Models for the Activation Energy of Static Recrystallisation in Hot-worked Steels [J]. Materials Science Forum, 2003, 4262432 (2):1181-1188.
[37]Struyf A, Rousseeuw P J. High dimensional Computation of the Deepest Location[J]. Computational Statistics and Data Analysis, 2000,34: 415-426.
[38]Han Jiawei, Micheline K. Data mining: concepts and techniques. 2nd edition [M]. San Francisco : Morgan Kaufmann Publishers , 2006.
[39]Arning A, Agrawal R, Raghavan P. A Linear Method for Deviation Detection in Large Databases [C]. Proceedings of 1996 Int. Conf. Data Mining and Knowledge (Special Issue on High Performance Data Mining), 2000.
[40]Breunig M, Kriegel H P,Ng R, et al. LOF: Identifying density based local outliers [C].Proceedings of ACM SIGMOD Conference. Dallas, 2000 :93-104.
[41]薛安荣, 姚林.离群点挖掘方法综述[J].计算机科,2008,35(11):13-18.
[42]Jin Wen, Tung Ant hony K H, Han Jiawei ,et al . Ranking Outliers Using Symmetric Neighborhood Relationship [C].Proceedings of the PA KDD. 2006: 577-593.
[43]Angiulli F, Fabio F. Very efficient mining of distance-based outliers. Proceedings of 6th ACM Conference on information and knowledge management. New York, 2007, 791-800.
[44]杨茂林. 离群检测算法研究[D]. 华中科技大学, 2012.
[45]Vu NH, Gopalkrishnan V. Efficient pruning schemes for distance-based outlier detection[C]. Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases.2009:160-175.
[46]Amol Ghoting, Srinivasan Parthasarathy.et al. Fast mining of distance-based outlier sin high-dimensional datasets [J].Pattern Recognition Letters, 2001, 22(6-7):691-700.
[47]KDD CUP 99 Data Set [Z]. Available at: https://archive.ics.uci.edu/ml/machine-learning-databases/kddcup99-mld/kddcup99.html. Accessed 2014.
[48]He Z, Deng S, Xu X. A fast greedy algorithm for outlier mining [C]. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 567-576, Seoul-Korea, 2006.
[49]Koufakou A, Oritiz E G, et al. A Scalable and Efficient Outlier Detection Strategy for Categorical Data[C]. Proc. Of the 19th IEEE International Conference on Tools with Artificial Intelligence, Washington DC, 2007, 210-217.
[50]Michael K.Ng, Mark Junjie Li, et al. On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm [J]. IEEE Transaction On Pattern Analysis And Machine Intelligence, 2007, 29(3), 503-507.

[51]Liang Bai, Jiye Liang, et al. An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data [J]. Knowledge-Based System, 2011, 24,785-795.
[52]Breast Cancer Wisconsin (Original) Data Set [Z]. Available at: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original). Accessed 2014.
[53]Mushroom Data Set [Z]. Available at: https://archive.ics.uci.edu/ml/datasets/Mushroom. Accessed 2014.