数据挖掘中的异常值检测算法

静静熙

计算机科学与技术学院

中国矿业大学，江苏徐州221008 [xjk@cumt.edu.cn](mailto:xjk@cumt.edu.cn)

摘　　要

　　离群值定义为与其他观察值有太大差异的观察值。识别异常值可以导致发现有用和有意义的知识。在过去的几十年中，离群值检测已被广泛研究。但是，现有的研究大多集中在基于特殊背景的算法上，与之相比，离群值检测方法仍然很少。本文主要从数据挖掘的角度讨论和比较不同离群点检测的方法，可以将其分为两类：经典离群点方法和空间离群值方法。经典离群值方法基于交易数据集分析离群值，可以将其分为基于统计的方法，基于距离的方法，基于偏差的方法，基于密度的方法。空间离群值方法基于空间数据集分析离群值，即非空间和空间数据与交易数据有显着差异，可以将其分为基于空间的方法和基于图的方法。最后，本文总结了离群值检测的最新进展;

１　　简介

　　数据挖掘是从大型数据集中提取有效的，以前未知的，最终可理解的信息，并将其用于组织决策的过程[1]。但是，在大型数据集中挖掘数据存在很多问题，例如数据冗余，属性值不明确，数据不完整和异常[2]。

　　离群值被定义为与其他观测值相差太大的观测值，从而引起怀疑它是由与其他观测值不同的机制生成的[3]。异常值的识别可以导致发现有用和有意义的知识，并且在交通，生态，公共安全，公共卫生，气候学和基于位置的服务等领域具有许多实际应用。最近，已经进行了一些有关大型数据集离群值检测的研究[4]。但是，大多数现有研究都集中在基于特殊背景的算法上离群值检测方法仍然很少。本文主要从数据挖掘的角度讨论离群值检测方法。固有的想法是研究和比较这些方法的实现机制，以根据特殊的数据集和不同的背景来确定哪种方法更好。

本文的其余部分安排如下。第2节回顾了异常值检测中的相关工作。第三部分讨论和比较离群值检测方法，可以将其分为两种方法：基于事务数据集的经典离群值方法和基于空间数据集的空间离群值方法。经典的离群值方法可以分为基于统计的方法，基于距离的方法，基于偏差的方法，基于密度的方法。空间离群值方法可以分为基于空间的方法和基于图的方法。第4节提供了离群值检测的最新进展。最后，第5节总结了这些离群值检测算法

2　　准备工作

　　离群值的经典定义是由Hawkins [3]提出的，他定义“离群值是与其他观察值有很大差异，以至于引起怀疑的原因是它是由不同的机制产生的”。

在早期工作中，大多数关于异常值挖掘的方法都是基于统计数据，这些统计数据使用标准分布来拟合数据集。 离群值是根据概率分布定义的。 例如，Yamanishi等人。使用高斯混合模型来描述正常行为，并根据模型的变化为每个对象评分[5]。

克诺尔等。提出了一个基于距离概念的新定义，如果数据集中的k个点的距离不超过λ或小于p，则将数据集中的点p相对于参数K和λ视为离群值 [6]。

阿宁等 提出了一种基于偏差的方法，该方法通过检查数据集中的对象的主要特征和与这些特征“偏离”的对象来识别异常值[7]。

Breunig等。 引入了局部离群值的概念，这是一种基于密度的离群值，它根据每个邻域为每个数据分配一个离群值局部离群因子LOF [2]。仅当采用某些多维索引结构（例如R树和X树[8]）时，离群因子才可以非常有效地计算出来。在文献[9]中提出了一种基于top-n的局部离群值挖掘算法，该算法使用距离受限的微簇来估计密度。

Lazarevic和Kumar提出了一种叫做“特征装袋”的技术[10]。

Shekhar等。[11]提出了空间离群值的定义：“空间离群值是一个空间参考对象，其非空间属性值与其空间邻域中的其他空间参考对象的值明显不同”。

寇等。 开发了空间加权离群值检测算法，该算法在比较非空间属性时使用中心距离和公共边界长度等属性作为权重[12]。 阿达梅特·阿勒 提出了一种同时考虑邻居之间的空间关系和语义关系的算法[13]。Liu和Jezek提出了一种在不规则分布的空间数据集中检测离群值的方法[14]。

3　　利群检测方法

在过去的几十年中，离群值检测得到了广泛的研究，并且开发了许多方法。 这些方法主要可以分为两类：经典离群值方法和空间离群值方法。 经典离群值方法基于交易数据集分析离群值，可以将其分为基于统计的方法，基于距离的方法，基于偏差的方法，基于密度的方法。 空间离群值方法基于空间数据集分析离群值，可以将其分为基于空间的方法和基于图的方法，如图3-1所示。

3.1　经典离群值

经典离群值方法基于交易数据集分析离群值，交易数据集由项目集合组成。一个典型的例子是市场购物篮数据，其中每笔交易都是客户在一次交易中购买的商品的集合。还可以通过描述客户或交易环境的其他“项目”来扩充此类数据。通常，交易数据是相对于其他数据而言的，以简化异常值检测。因此，对交易数据研究了大多数异常方法。

（1）统计方法

统计方法是用于异常值检测的最早算法，该算法假定给定数据集的分布或概率模型，然后确定使用不一致检验的模型的异常值。实际上，Barnett和Lewis [15]以及Rousseeuw和Leroy[16]中描述的许多技术都是一维的。但是，随着维度的增加，为数据集建立模型变得更加困难且不准确。

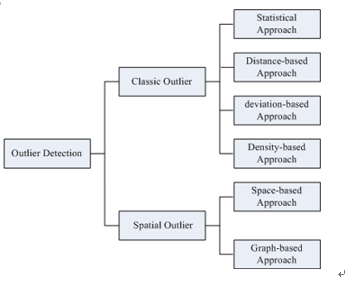


图3-1 利群检测方法

（2）基于距离的方法

基于距离的离群值的概念依赖于点的邻域概念，通常是k最近邻，最早由Knorr和Ng提出[6，17]。基于距离的离群值是输入数据集中距离内少于k个点的那些点。此定义未提供离群值的排名，需要确定参数的适当值。Ramaswamy等。[18]修改了Knorr和Ng引入的离群值的定义，并认为与第k个最近邻点的距离最大的前n个点p为离群值。为了检测异常值，提出了一种基于分区的算法，该算法首先使用聚类算法对输入点进行分区，然后对不能包含异常值的那些分区进行修剪。基于距离的方法在相当小的尺寸上是有效的，因为高维点的稀疏性，该方法对参数λ敏感，并且很难弄清先验。随着尺寸的增加，该方法的效果和准确性迅速下降。

（3）基于偏差的方法

阿宁等提出了一种基于偏差的方法，该方法通过检查数据集中的对象的主要特征和与这些特征“偏离”的对象来识别异常值[19]。

（4）基于密度的方法

基于密度的方法估计数据的密度分布，并将异常值标识为位于低密度区域的异常值。 Breunig等。[2]分配一个基于每个点附近的局部密度的局部离群值因子（LOF），该密度由用户指定的最小点数（MinPts）确定。 Papadimitriou等。 [20]提出了LOCI（局部相关积分），它使用基于数据本身的统计值来解决为MinPts选择值的问题。基于密度的技术的优势在于，它们可以检测具有单个全局标准的技术会遗漏的异常值。但是，数据通常在高维空间中稀疏，从而导致基于密度的方法存在问题。

3.2　空间离群值

对于空间数据，由于空间属性和非空间属性之间的质量差异，必须修改经典方法。空间数据集可以定义为空间参考对象的集合，例如道路，建筑物和城市。空间对象的属性分为两类：空间属性和非空间属性。空间属性包括位置，形状和其他几何或拓扑属性。非空间属性包括长度，高度，所有者，建筑​​物年龄和名称。空间参考对象的空间邻域是使用空间关系（例如，距离和邻接）基于空间维度的空间数据的子集。空间参考对象之间的比较基于非空间属性[21]。

空间离群值是空间参考对象，其非空间属性值与其空间邻域中的其他空间参考对象的非空间属性值显着不同。非正式地，空间离群值是局部不稳定性，或者是关于其邻近值的极端观察，即使它与整个总体可能没有显着差异。检测空间离群值在地理信息系统和空间数据集的许多应用中很有用[11、21、22]。

空间异常值的识别可以在许多应用中显示隐藏但有价值的信息，例如，它可以帮助定位严重的气象事件，发现高速公路拥堵路段，在卫星图像中确定军事目标，确定油藏的潜在位置以及检测水污染事件。

（1）天基方法

基于空间的离群值使用欧几里得距离来定义空间邻域。寇等。开发了空间加权离群值检测算法，该算法在比较非空间属性时使用中心距离和公共边界长度等属性作为权重[12]。阿达梅特·阿勒（Adamet al。）提出了一种同时考虑邻居之间的空间关系和语义关系的算法[13]。刘等。提出了一种在不规则分布的空间数据集中检测离群值的方法[14]。

（2）基于图的方法

  基于图的方法使用图连接性来定义空间邻域。寇玉峰等。提出了一组基于图的算法来识别空间离群值，该算法首先在空间域中基于k最近邻关系构造图，将非空间属性差异分配为边缘权重，然后连续切割高权重边缘以识别孤立点与它们的相邻对象非常不同的点或区域。与现有的空间离群值检测方法相比，该算法有两个主要优点：精确地检测点离群值和能够识别区域离群值[24]。

4　　离群值检测的最新发展

随着数据挖掘技术的飞速发展，大型数据集中离群值的识别受到越来越多的关注。传统的异常值检测方法可能无法有效地应用于大型数据集。因此，针对特定背景专门设计了一些新方法。

（1）基于高维的方法

高维空间是离群值检测的难题。根据文献[22]中提出的针对高维设计的技术准则，本文提出了一种基于投影概念的新方法ODHDP，可以很好地处理高维点的稀疏性。该方法的基本思想是通过对数据集的投影进行聚类来找到异常值。因此，首先，将每个维度上的数据集的投影聚类，并对每个维度赋予不同的权重；其次，依次为笛卡尔组合聚类选择其余维度中权重最大的维度，然后修剪点数小于阈值的候选聚类，直到扫描完所有维度为止。第三，基于残差中的点与聚类在全维度上的关系来计算残差的相似度，从而将异常值与残差区分开来[23]。

（2）基于SVM的方法

提出了一种基于支持向量机的离群值检测方法[25]。该方法基于从SVM模型获得的支持向量的特征，使用多种复杂程度不同的模型来检测离群值。这样做的优点是决策不取决于单个模型的质量，这增加了方法的鲁棒性。此外，由于这是一种迭代方法，因此将最严重的异常值先删除。这使下一次迭代中的模型可以从“更清洁”的数据中学习，从而揭示在初始模型中被“掩盖”的离群值。其他异常值检测工作包括支持向量方法[26]，使用复制器神经网络（RNN）[27]，以及仅针对少数几个固定参考点使用相对密度[28]。

5　　结论

本文主要从数据挖掘的角度讨论离群值检测方法。 首先，我们回顾离群值检测中的相关工作。 接下来，我们讨论并比较离群值检测算法，该算法可分为两类：经典离群值方法和空间离群值方法。 经典离群值方法基于交易数据集分析离群值，可以将其分为基于统计的方法，基于距离的方法，基于偏差的方法，基于密度的方法。 空间离群值方法基于空间数据集分析离群值，可以将其分为基于空间的方法，基于图的方法。 第三，我们总结了离群值检测的一些最新进展。

参考文献

1. Yu, D., Sheikholeslami, G. and Zang, “A find out: finding outliers in very large datasets”. *In Knowledge and Information Systems*, 2002, pp.387-412.
2. Breunig, M.M., Kriegel, H.P., and Ng, R.T., “LOF: Identifying density-based local outliers.” *ACM Conference Proceedings*, 2000, pp. 93-104.
3. D. M. Hawkins, “Identification of Outliers”. *Chapman and Hall*, London, 1980.
4. Aggarwal, C. C., Yu, S. P., “An effective and efficient algorithm for high-dimensional outlier detection, *The VLDB Journal*, 2005, vol. 14, pp. 211-221.
5. Yamanishi. K, Takeuchi. J ,and Williams. G On-line, “unsupervised outlier detection using finite mixtures with discounting learning algorithms”. *In Proceedings of the Sixth ACM SIGKDDOO*, Boston, MA, USA, pp.320-324.
6. Knorr, E.M., Ng, R.T., "Finding Intentional Knowledge of Distance-Based Outliers", *Proceedings of the 25th International Conference on Very Large Data Bases*, Edinburgh, Scotland, pp.211-222, September 1999.
7. Agarwal, D., Phillips, J.M., Venkatasubramanian, “The hunting of the bump: on maximizing statistical discrepancy”. *In: Proc. 17th Ann. ACM-SIAM Symp*. on Disc. Alg. pp. 1137–1146 (2006).
8. Berchtold, S., Keim, D., Kriegel, H.-P, “The X-tree: An efficient and robust access method for points and rectangles”. *In: VLDB (1996)*.
9. Jin, W., Tung, A.K.H., Han, J.W. “Mining Top-n Local Outliers in Large Databases”. *In: KDD (2001)*.
10. Lazarevic, A., Kumar” Feature Bagging for Outlier Detection”. *In: KDD (2005)*.
11. S. C. Shashi Shekhar, “Spatial Databases: A Tour. Prentice Hall”, 2003.
12. Y. Kou, C.-T. Lu, and D. Chen. “Spatial weighted outlier detection”. *In Proceedings of the Sixth SIAM International Conference on Data Mining*,pp. 614–618, Bethesda, Maryland, USA, 2006.
13. N. R. Adam, V. P. Janeja, and V. Atluri., “Neighborhood- based detection of anomalies in high-dimensional spatio- temporal sensor datasets”. *In Proceedings of the 2004 ACM，symposium on Applied computing*, Nicosia, Cyprus, 2004. pp. 576–583
14. H. Liu, K. C. Jezek, and M. E. O’Kelly, “Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and gis”. *International Journal of Geographical Information Science*, 15(8), 2001. pp.721–741.
15. Barnett, V. & Lewis, T. (1994).,”Outliers in Statistical Data”, *3rd edn*. John Wiley & Sons.
16. Rousseeuw, P. & Leroy, A. (1996).,”Robust Regression and Outlier Detection”, *3rd edn*. John Wiley & Sons.
17. E. Knorr, R. Ng, and V. Tucakov, “Distance-Based Outlier: Algorithms and Applications,” *VLDB* J., vol. 8, nos. 3-4 2000, pp. 237-253.
18. S. Ramaswamy, R. Rastogi, and K. Shim, “Efficient Algorithms for Mining Outliers from Large Data Sets,” *Proc. Int’l Conf. Management of Data (SIGMOD ’00)*, 2000, pp. 427-438.
19. A. Arning, R. Agrawal, and P. Raghavan, “A Linear Method for Deviation Detection in Large Databases,” Proc. Int'l Conf. Knowledge Discovery and Data Mining, 1996, pp. 164-169.
20. Papadimitriou, S., Kitawaga, H., Gibbons, P., Faloutsos, C., “LOCI: Fast outlier detection using the local correlation integral”, *Proc. of the Int’l Conf. on Data Engineering*, 2003.
21. Chang-Tien Lu, Dechang Chen,Yufeng Kou, “Detecting spatial outliers with multiple attributes”, *Tools with Artificial Intelligence, 2003.* Proceedings. 2003, pp.122– 128.
22. Aggarwal, C.C, Yu, P. "Outlier detection for high dimensional data", *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Santa Barbara, CA, 2001, pp. 37-47.
23. Ping Guo, Ji-Yong Dai, Yan-Xia Wang, “Outlier Detection in High Dimension Based on Projection”, Machine Learning and Cybernetics, 2006 International Conference, 2006,pp.1165 – 1169.
24. Yufeng Kou, Chang-Tien Lu, Dos Santos, R.F.” Spatial Outlier Detection: A Graph-Based Approach”, ICTAI 2007, Volume 1, 2007,pp.281 – 288.
25. E.M. Jordaan. Deployment of Robust Inferential Sensors, “Irtdusrriol application of Supper Vector Machines for Regression”, Ph. D. thesis. Eindhaven University of Technology, 2002.
26. Jordaan, E.M.;,Smits, G.F., ” Robust outlier detection using SVM regression”, Neural Networks, 2004. Proceedings. 2004,pp.2017 – 2022.
27. Harkins, S., He, H., Williams, G., Baster, R., “Outlier Detection Using Replicator Neural Networks”, *DaWaK’02*, 2002, pp. 170-180.
28. Pei, Y., Zaiane, O., Gao, Y., “An Efficient Reference based Approach to Outlier Detection in Large Dataset”, *IEEE Int’l Conference on Data Mining*, 2006.