计算机科学
COMPUTER SCIENCE

离群点检测算法综述

孔翎超, 刘国柱

引用本文

孔翎超, 刘国柱. 离群点检测算法综述[J]. 计算机科学, 2024, 51(8): 20-33.

KONG Lingchao, LIU Guozhu. Review of Outlier Detection Algorithms[J]. Computer Science, 2024, 51(8): 20-33.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600052

离群点检测算法综述

孔翎超　刘国柱
青岛科技大学信息科学技术学院　山东 青岛 266061
(kk1392567492@163.com)

摘　要　离群点检测作为数据挖掘领域的一个重要研究方向,其目的是发掘隐藏在数据集合中与众不同且具有潜在分析价值的数据,辅助研究人员甄别数据源可能存在的问题。目前,离群点检测已被广泛应用于欺诈识别、智慧医疗、入侵检测、故障诊断等诸多领域。文中在总结前人经验的基础上,首先讨论离群点的定义、产生原因以及典型应用领域,综述了DBSCAN和LOF等离群点检测经典算法及其改进算法的优势和局限,分析了深度学习方法在离群点检测领域的优势;其次结合当前互联网背景下海量、高维、时序数据处理需求,对离群点检测算法在新环境下的发展状况做进一步研究;最后介绍离群点检测算法的评价指标、代价因子在离群点检测评价中的作用以及常用工具包和数据集,总结展望了离群点检测面临的挑战和未来的发展方向。

关键词:离群点;异常检测;深度学习;时序数据;数据挖掘
中图分类号　TP301

Review of Outlier Detection Algorithms

KONG Lingchao and LIU Guozhu
School of Information Science and Technology,Qingdao,Shandong 266061,China

Abstract　Outlier detection,as an important research direction in the field of data mining,aims to discover data points in a dataset that are different from the majority and have potential analytical value,assistresearchers in identifying potential issues in the data source. Currently,outlier detection has been widely applied in various domains such as fraud detection,smart healthcare,intrusion detection,and fault diagnosis. This study,based on summarizing previous experiences,first discusses the definition of outliers,their causes,and typical application domains. It reviews the advantages and limitations of classical outlier detection algorithms such as DBSCAN and LOF,as well as their improved algorithms. Additionally,it analyzes the advantages of deep learning methods in the field of outlier detection. Secondly,considering the requirements for processing massive,high-dimensional,and temporal data in the current internet context,further research is conducted on the development status of outlier detection algorithms in new environments. Finally,the evaluation indicators of outlier detection algorithms,the role of cost factors in outlier detection evaluation,as well as commonly used toolkits and datasets,are introduced. The challenges and future development directions of outlier detection are summarized and prospected.

Keywords　Outliers,Anomaly detection,Deep learning,Time-series data,Data mining

离群点,也称为异常点或新奇点。关于离群点的定义,被大多数研究者所认同的是由Hawkins提出的解释:离群点是一种与数据集合中其他数据非常不同的数据,从而使人们怀疑它是由一种不同的机制所产生的[1]。离群点虽然在数据集合中所占比例很小,但蕴含着非常重要的信息,尤其是在欺诈检测、工业设备监控等大数据应用领域具有重要的参考意义和使用价值。表1列出了离群点的应用领域以及相关技术,因此离群点检测(Outlier Detection)成为数据挖掘领域的研究热点之一[2]。

表1　应用领域及相关文献

Table 1　Application field and related literatures

应用领域	相关技术
网络入侵检测	粗糙集[3]、模式匹配等[4]、神经网络[5]、随机森林[6]
欺诈检测	深度学习[7]、图模型[8]、SVM等[9]、GAN[10]
医学异常检测	深度学习[11]、GAN[12]、深度自编码器[13]
传感器异常检测	PCA+局部投影回归[14]、神经网络[15]
视频监控异常检测	LSTM[16]、深度学习[17]

基于离群点检测在数据挖掘领域的重要作用,目前已有多篇有关离群点检测算法的研究综述发表。Chandola等[18]

到稿日期:2023-06-06　返修日期:2023-12-13
基金项目:国家自然科学基金(61973180)
This work was supported by the National Natural Science Foundation of China(61973180).
通信作者:刘国柱(lgz_0228@163.com)

介绍了离群点检测背景和研究意义，描述了各类经典算法的特点，重点分析了离群点检测在各个领域的应用并对未来发展方向进行了展望。国内 Xu 等[19]、Xue 等[20]也对离群点检测各种经典算法进行了总结。但是上述文献的年份较老，缺乏对近十年方法的研究。近年来，Mei 等[21]在详细分析传统经典方法的基础上提出了基于分布式的离群点检测方法；Wu 等[22]从有监督和无监督的角度重点分析了深度学习在离群点检测中的应用；Lei 等[23]介绍了新奇检测的相关概念，并对经典算法以及相关数据集进行了总结阐述。本文在上述文章的基础上对传统算法的优势与缺陷进行了详细总结，分析了离群点检测在新环境下的应用，介绍了离群点检测算法的评价指标以及常见工具包和数据集，并对今后的发展方向进行了总结与展望。

1 离群点检测概述

1.1 离群点产生的原因

1）离群点来源于新的类别

离群点发生在数据期望之外，偏离数据产生的正常轨道，属于本不可能发生的事件，例如新型疾病的爆发、网络入侵以及电信诈骗等。

2）数据的小概率事件

离群点的出现是由于数据自然变化导致的，属于同一数据集中的小概率事件，并且符合数据集的分布特征，例如气候的突然变化等。

3）测量误差

由于人为错误或测量设备故障导致的异常需要在数据清洗阶段剔除。

1.2 离群点分类

根据数据异常出现的不同情况可离群点划分为：点异常、上下文异常、集合异常以及其他异常[24]。

1）点异常

点异常（Point Anomalies）指在数据集合中与大部分数据存在显著不同的样本。将数据集合投影到高维空间中，异常点与正常点一般相距较远。如图 1 所示，点 O_1、O_2 以及簇 O_3 位于正常数据区域边界之外，此类异常属于点异常。该类异常发生最为普遍，也是离群点检测关注的重点。

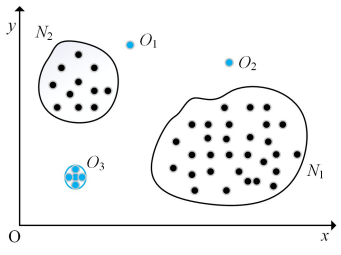


图 1 点异常

Fig. 1 Point anomalies

2）上下文异常

上下文异常（Contextual Anomalies）也称为条件异常，指某一点 p 虽然在全局范围内属于正常点，但在特定条件下属于异常点。这里的上下文代指数据的结构和关系，数据集中的每个数据属性均由上下文特征和行为特征构成，对于同一点 p，在不同数据环境下可能存在不同的结果。因此，在离群点检测的过程中，不能只参考数据的数值大小，还要结合数据产生的环境条件。该方法常被用于时间序列数据或空间数据的检测问题中。图 2 给出了北半球某地的全年温度变化情况，t_1 和 t_2 时刻温度相同且均属于全年最低值，t_2 时刻为冬季，出现该温度是正常情况，但 t_1 时刻为夏季，不符合夏季温度的特征，因此属于上下文异常点。

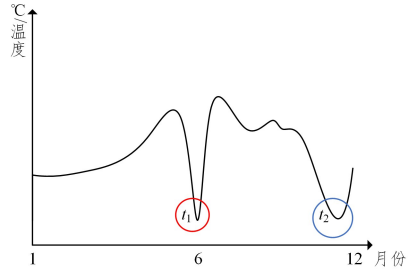


图 2 上下文异常

Fig. 2 Contextual anomalies

3）集合异常

针对数据集合中的某一部分数据，对它们进行单独检测时不属于异常点，但是这些点及其周围的点作为一个集合出现时可能存在异常，这些相关联的点组成的集合被称为集合异常（Collective Anomalies）。如图 3 所示，脑电图中低值本身不属于异常，但是曲线红色位置出现长时间的低值，该区域为集合异常。

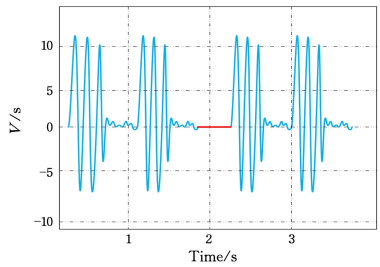


图 3 集合异常（电子版为彩图）

Fig. 3 Collective anomalies

4）其他异常

除上述异常外，根据特定的应用环境与检测数据，还存在向量异常（Vector Anomalies）、序列异常（Sequence Anomalies）、轨迹异常（Trajectory Anomalies）和图异常（Graph Anomalies）等。

1.3 离群点检测输出结果

1）分数

算法根据数据的离群程度，为数据中的每个实例分配一个离群概率分数，分析人员可以通过设定合适的阈值，对离群点与正常点进行划分。

2）标签

由于离群点检测的结果只有两类，因此可以直接输出正常或异常标签。但是，该种输出结果使得研究人员难以对样本的离群程度进行观测。

2 经典离群点检测算法

本章对离群点检测相关算法（包括基于统计学、基于聚类

等经典方法，以及在深度学习背景下基于自编码器和生成对抗网络的算法）进行总结，讨论各算法及其改进算法的优势与缺陷，分析各算法适合的应用场景以及数据类型。

2.1 基于统计模型的方法

基于统计学的离群点检测方法的核心思想是为数据集合拟定一个数学分布模型，离群点一般只占数据集合的小部分。基于统计学的方法可分为参数和无参数两类。

2.1.1 参数方法

参数方法假定数据服从某个以 λ 为参数的概率分布模型，数据点 x 的分布可能性由概率密度函数 $f(x,\lambda)$ 给出，该值越小表示数据点 x 越不符合该分布模型，是离群点的可能性就越大。常见的方法有高斯模型和回归模型。

2.1.2 无参数方法

无参数方法无需提前假定数据分布模型，其通过对数据进行分析学习得到分布特征，常用的方法有直方图和核密度估计法（Kernel-density Estimators，KDE）。

2.2 基于距离的方法

基于距离的离群点检测方法[25]是数据挖掘中最常用的方法之一，该类方法于 1998 年由 Knoor 和 Ng 提出，算法的核心思想是正常数据点与周围的点距离较近，而距离周围点大于某一预先设定阈值的数据即为离群点，通过计算样本点与周围数据点的欧氏距离或曼哈顿距离来检测异常。K-近邻算法（K-nearest neighbor，KNN）[26]是常见的基于距离的离群点检测算法，该算法计算每一个数据点 p 到其第 k 个最近邻数据点的距离 $Dk(p)$，根据距离大小将数据点进行排序，将距离大于预先设定阈值的点判定为离群点，但是 KNN 算法要计算数据集合中全部点的 $Dk(p)$，计算效率较低。后续研究人员提出了多种距离检测改进算法。基于索引（Index-based）的算法通过构建空间索引结构（如 X-树、R-树、KD-树等）来计算样本点指定半径领域数据点的个数，根据距离范围查找离群点。该算法受数据维度变化影响较小，在数据维度增加时具有良好的扩展性，但是构造索引结构的过程本身十分复杂，这在一定程度上限制了基于索引算法的应用。基于嵌套-循环（Nested-loop Based）的算法使用嵌套循环结构计算数据点之间的距离来检测离群点，在处理过程中将数据集划分为多组逻辑块分批次进行处理，提高了算法的数据处理效率。基于单元（Cell-based）的算法将数据集合分割成彼此相对独立的单元，通过对每一个单元的检索来实现快速异常检测。但是，随着数据维度的递增，单元数呈指数级增长，对于维度为 4 或更高维的数据集，该算法相比嵌套循环算法的性能降低。

基于距离的异常检测算法不受数据集分布状态约束，能够在未知数据集分布状态的情况下检测离群点。但是，该类算法存在以下不足：1）在处理海量高维数据时执行效率较低；2）距离计算函数选择和算法相关参数选择对检测效果影响较大；3）只能检测全局离群点，难以发现局部离群点。

2.3 基于聚类的方法

基于聚类的离群点检测的核心思想是通过聚类分析（Clustering Analysis）方法，根据不同数据的潜在特征属性将数据集合划分为多个簇，其中不属于任何一簇或者远离簇中心的数据便是离群点。使用聚类算法对异常数据进行检测时，对离群点出现的情况有以下 3 种假设：1）如图 4（a）所示，异常是不属于任何簇的点，将正常的簇删除掉，剩余的数据即为离群点，该类算法有 DBSCAN 算法[27]、SNN 算法[28]、ROCK 算法[29]等；2）如图 4（b）所示，将数据聚类成簇，远离簇中心的点为异常点，该类算法有 K-means 算法[30]、SOM 算法[31]等；3）如图 4（c）所示，经过聚类分析之后的数据集会形成大小、密度不一的簇，那些体积较小并且内部数据稀疏的簇被认为是异常簇，簇中包含的数据为异常数据，该类算法有 CBLOF 算法[32]、LDCOF 算法[33]、CMGOS 算法[34]等。

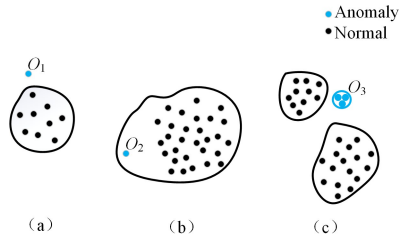


图 4 3 种聚类异常

Fig. 4 Three types of clustering anomalies

2.3.1 异常不属于任何集群

DBSCAN 算法（Density-Based Spatial Clustering of Applications with Noise）[35]是经典的基于密度聚类离群点检测算法，该算法的核心思想是通过计算某一点邻域的密度判断该点是否为离群点。该算法有两个重要的参数：密度阈值（Minpts）和邻域半径（Eps）。如图 5 所示，若数据集合中某一样本点 P 在其邻域半径范围内有大于等于密度阈值数量的样本点，则 P 点为核心点；若某样本点不符合核心点要求，但其在某一核心点的邻域半径范围之内，例如图中的 M 点，则该样本点称为边界点；经过迭代更新之后，剩余的既不属于核心点也不属于边界点的数据就是离群点，这些数据一般位于集合的稀疏区域。该算法的检测效果取决于 Minpts 和 Eps 两个超参数以及距离度量方法的选择。

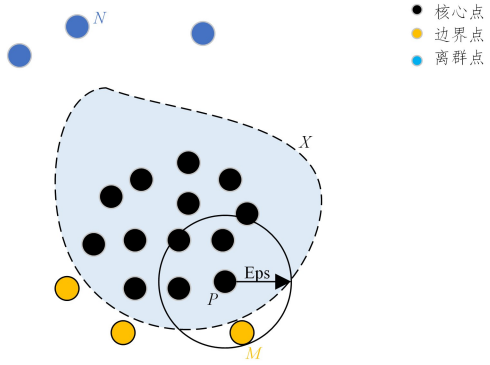


图 5 DBSCAN 算法

Fig. 5 DBSCAN algorithm

DBSCAN 具有以下优势：1）可以根据样本之间不同特征聚类合适数量的簇，不需要预先设定聚类簇的数量；2）能够在聚类的同时较好判别离群点；3）可以聚类任意形状的数据集，发现不同大小、不同稠密度的簇，对簇划分效果好。但是，DBSCAN 也存在一些局限：1）使用欧氏距离作为度量方法时，对于高维数据可能产生“维数灾难”；2）数据集合密度差异

很大时，聚类效果较差；3）受 Minpts 以及 Eps 两个超参数影响较大。Ankerst 等[36]提出了 OPTIC 算法，解决了 DBSCAN 算法处理密度不均匀导致的数据效果差以及对超参数敏感的问题。

2.3.2 异常远离集群中心

K-means 是一种简单、高效的划分聚类算法。如图 6 所示，该算法的基本原理是根据预先给定的 k 值随机选取 k 个质心，经过迭代更新将数据集合划分为 k 个簇，实现每个簇内数据高度相似而簇间数据相异，簇内距离质心较远的数据以及远离大簇的稀疏簇内数据就是离群点。

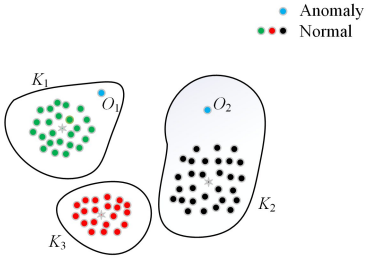


图 6 *K*-means 算法

Fig. 6 *K*-means algorithm

K-means 算法的数据处理效率较高，且算法通俗易懂，因此在数据挖掘领域中被广泛应用。但是该算法也存在以下局限：1）在进行数据处理前，k 值是未知的，需要依据先验知识给定合适的 k 值；2）*K*-means 算法对数据样本的初始分布很敏感，由于使用欧氏距离计算数据点之间的距离，因此算法的聚类结果偏向于球形，与 DBSCAN 算法相比不适合发现非凸状的簇或大小差别较大的簇；3）算法的聚类效果受制于初始质心的选择，在迭代更新的过程中，当质心位于局部最小值附近时，算法容易陷入局部最优而无法得到全局最优。

2.3.3 异常属于小集群或稀疏集群

前文描述的算法对离群点检测的表现突出，但检测由离群点聚合成的小簇较为困难，因此衍生出基于聚类的局部异常因子算法（Cluster-Based Local Outlier Factor，CBLOF）。该算法首先通过聚类生成大小不一的数据簇，按照给定阈值区分大小簇，根据簇中样本点到簇中心的距离、不同簇之间的距离以及簇内样本点个数的乘积计算每一个簇的 CBLOF 得分，以此来区分异常簇。但是 CBLOF 算法只计算簇中的样本数量，没有考虑簇的空间大小，因此没有真实表达簇的局部概率密度。Amer 等针对 CBLOF 算法的局限提出了 LDCOF 算法，首先得出簇中全部样本点到簇中心的平均距离，将待测样本点到簇中心的距离除以平均距离得出 LDCOF 得分，该方法兼顾了簇的密度特征。Goldstein 提出了改进算法 CMGOS，该算法提出了新的观点：1）约简；2）正则化；3）协方差矩阵，使用多维高斯模型实现局部概率密度估计，并引入 Mahala Nobis Distance 计算数据点之间的距离，基于马氏距离的性质可以较好区分异常簇。

2.3.4 小结

聚类算法属于无监督学习，训练过程中数据不需要完整的标签，可以通过数据特征之间的关系进行有效聚类并发现数据异常。*K*-means 算法的实时性好，可以实现在线检测。但是聚类算法大多存在以下缺陷：1）算法需要针对特定的数据分布情况进行选择；2）数据维度高并且分布稀疏时，正常数据与异常数据之间的距离差异会减小，影响检测算法的聚类效果；3）算法性能受超参数选择影响较大。

2.4 基于密度的方法

如图 7 所示，基于密度的离群点检测算法是距离算法的进一步优化，该算法认为正常样本的邻域数据密度与其周围样本的邻域数据密度相似，而离群点的邻域密度会明显低于其周围样本点，因此该类算法有一个显著的特点，即能够发现在全局数据中无明显异常，但在局部数据集合中与邻近数据相比存在显著异常的样本点。经典的基于密度异常的检测算法有 LOF 算法、COF 算法、INFLO 算法、LOOP 算法等。

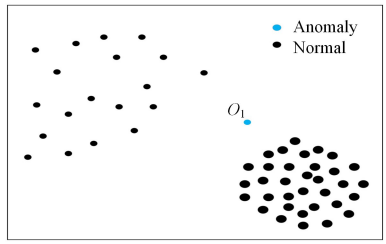


图 7 局部离群点

Fig. 7 Local outliers

局部离群因子算法（Local Outlier Factor，LOF）[37]是基于密度的经典无监督异常检测算法。该算法的基本原理是，计算某一点 p 在其 k 近邻集合内与邻居点的局部密度偏差，将其作为该点的离群程度 LOF，LOF 越大，点 p 的离群程度越大。LOF 算法[38]的优点是：1）无需预先知道数据集的分布情况，通过计算 LOF 值来找出其中的异常点，能够量化数据点的离群程度；2）解决了对局部离群点的检测问题。但是，LOF 算法也存在一定的局限性：1）算法的效果容易受 k 值选取的影响；2）数据集合包含不同密度的簇时，稀疏簇边界的点容易被算法误判为离群点；3）算法中涉及大量的数据排序、查找和距离计算，面对庞大数据集时算法的时间开销较大。

针对 LOF 算法不能有效检测序列数据和低密度数据的缺陷，COF（Connective-based Outlier Factor）算法[39]引入最小延伸树（MST）使用链式距离方法来计算最短路径，以衡量局部近邻密度。但是，由于 COF 算法引入最小延伸树计算近邻距离，因此时间复杂度比 LOF 算法更高，并且对数据分布的间接假设会导致不正确的密度估计。INFLO（Influenced Outlier Ness）算法[40]在计算密度时既计算了数据点的 k 近邻，又兼顾了反向 k 近邻对数据点离群程度的影响，改进了 LOF 算法对于不同密度簇彼此接近时，无法精确检测边界点这一缺陷。但在现实情况下，离群点只占数据集的极小部分，INFLO 需要计算每个数据点的反向 k 近邻，这无疑会增加算法的复杂性。局部异常概率算法（Local Outlier Probabilities，LoOP）[41]改进 LOF 计算密度的方式是，将距离的均方根作为样本点处的密度，该方法相比直接使用距离均值的倒数作为密度更加健壮。MDEF（Multi-Granularity Deviation Factor）算法[42]引入 r-邻域和 αr-邻域的概念，该算法可以根据实际要求设置多级邻域，并用邻域中包含的数据数目代替距离计算，降低算法的计算复杂度。但是超参数很难确定，需要

经过反复修改，因此 MDFF 的算法效果取决于参数的选择。Tang 等[43]提出了一种基于局部 KDE 的离群点检测方法 RDOS(Relative Density-Based Outlier Score)，它使用一种基于核密度估计的有效相对密度计算方法来度量离群值，从而使孤立点的计算更加稳健，同时该算法不像大多数算法那样只考虑 k 近邻，而是兼顾检测样本点的反向近邻和共享近邻。

表 2 列出了基于密度的相关算法的性能与优缺点。

表 2 基于密度的算法对比

Table 2 Comparison of density based algorithms

算法	解决的问题	性能	缺点
LOF	为每一个数据点分配一个异常因子，使得离群点检测不再只是一个二分类问题；解决了局部密度异常的检测问题	有标签数据：$O(n)$ 无标签数据：$O(n^2)$ 中低维度数据：$O(n\log n)$ 高维数据：$O(n^2)$。	无法处理多粒度问题，对超参数选择敏感；时间复杂度较高；对于邻域密度非常接近的簇误检率高
COF	解决了 LOF 算法不能有效检测序列和低密度数据的问题；可以独立检测样本密度偏离问题	中维度数据：$O(n)$ 中维度数据：$O(n\log n)$ 高维数据：$O(n^2)$	时间复杂度高于 LOF 算法
INFLO	能够很好地反映数据集的密度特征，检测效果更好	对于低维数据时间复杂度较低；当维度大于 12 时，时间复杂度和计算复杂度很高	加入反向 k 近邻使得计算复杂度较高；只适用于基于对称邻域关系的局部异常检测
LoOP	为每个数据点提供离散概率分数	计算性能显著增强，对于高维数据计算复杂度更稳定	只重视异常检测的精确度，而忽视了计算效率
RDOS	对于大规模数据集，该算法对局部异常点的检测性能优于 LOF 算法	采用基于核函数的方法，计算效率较高	只使用欧氏距离度量方法，缺乏对其他距离度量方法的使用

2.5 基于深度学习的方法

随着数据采集和数据存储技术的发展，需要更准确、高效率的方法发掘数据的不同特征，检测高维海量数据中的离群点。深度学习(Deep Learning)在处理复杂数据(高维数据、时序数据、空间数据和图形数据)方面表现出了巨大的潜力，因此，基于深度学习的离群点检测方法逐渐引起了研究人员的重视。在监督的深度异常检测方法中，一般把离群点检测问题作为分类问题来解决，利用正常和异常数据实例的标签来训练二元或多元分类器。但是，在真实环境下，训练数据往往缺乏大量标签，并且异常值往往只占数据集的极小部分，存在类不平衡问题，在这种情况下使用该方法的检测效果较差；半监督的方法通过学习正常样本特征来区分离群点，Kiran 等[44]对基于深度学习的半监督异常检测技术进行了综述；无监督方法可以使用无标签数据进行训练，根据数据之间的不同特征以及内在属性发现离群点。基于深度学习的离群点检测算法有自编码器、生成对抗网络等。表 3 列出了深度学习的经典算法。

2.5.1 Autoencoder

自编码器(Autoencoder，AE)[45]是一种经典的无监督学习神经网络，被训练用于学习重建接近其原始输入的数据特征，如图 8 所示。AE 由编码网络(Encoder)和解码网络(Decoder)组成，编码器通过非线性映射将原始数据映射到低维特征空间，而解码器尝试将数据从投影的低维空间映射回原始输入空间，基于反向传播算法与最优化方法(如梯度下降法)对数据进行重建，力求最小化整体重建误差，该算法使用正常数据进行训练，如果自编码器重构的某一输出节点与原始输入的差异超出一定阈值(threshold)，则可判定该节点是离群点。之后，研究人员提出诸多 AE 的变体[46]，以学习更加丰富、更具表现力的数据特征。去噪自编码器(Denoising Autoencoder，DAE)[47]实现对带有噪声的输入数据进行特征重建。变分自编码器(Variational Autoencoder，VAE)[48]通过学习隐藏空间的先验分布对输入数据进行编码，将正则化方法引入表示空间，确保生成数据包含输入数据的重要特征，防止训练过程过度拟合。Zhang 等[49]基于 AE 模型需要使用正常数据进行训练的局限，结合 PCA 方法采用交替方向乘子训练模型，实现了一种基于深度自编码器(Deep Autoencoder，DAE)的离群点检测模型。

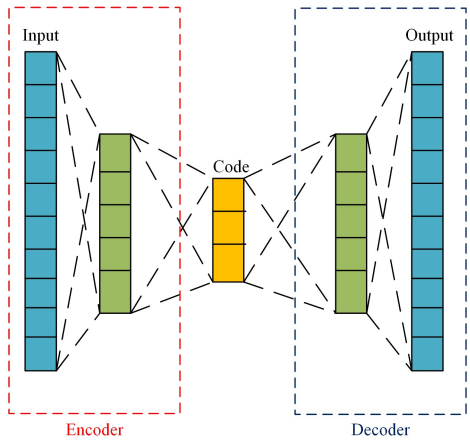


图 8 自编码器结构

Fig. 8 Autoencoder structure

2.5.2 生成对抗网络

生成对抗网络(Generative Adversarial Networks，GAN)和对抗性训练框架已成功实现对高维数据分布的拟合，表明其可以成功应用于离群点检测领域[50]。GAN 进行离群点检测任务的核心思想是，使用对抗性训练过程对正常数据进行建模，学习生成网络的潜在特征空间，使得潜在空间很好地捕捉给定数据的正态表示，然后将真实样本和生成样本之间的残差定义为异常分数。首次将生成对抗思想应用于异常检测领域的是 AnoGAN 算法[51]，该算法使用标准的 GAN 模型学习正常数据样本的空间分布特征，测试阶段随机选取一个数据样本 x，使用训练的生成模型尽可能拟合该样本 x，如果生成数据与原始数据 x 差异较大，则样本 x 为异常数据。但是该算法在迭代搜索过程中的计算效率较低，解决该问题的

一种方法是添加额外的网络，学习从数据实例到潜在空间的映射，即生成器的逆映射，如 EGBAD 算法[52] 和 fast AnoGAN 算法[53]。EGBAD 将 BiGAN 架构[54]引入异常检测领域，通过学习一个编码器 E，在对抗性训练过程中将输入样本映射到它的潜在空间表示。受上述经典算法的启发，GANomaly[55]改进生成器模型，将生成器网络修改为（编码器-解码器-编码器）结构，并添加额外的损失函数来提高检测效率。近年来，又有研究者提出了其他生成对抗模型，例如 Wasserstein GAN[56] 和 Cycle GAN[57] 等，进一步提高了异常检测算法的性能。

2.5.3 其他模型

循环神经网络（Recurrent Neural Network，RNN）[58] 能够处理长序列数据，兼顾数据之间的上下文联系，因此在时间序列的离群点检测过程中具有较大优势。自监督学习的离群点检测方法基于交叉特征分析，通过建立自监督分类模型学习数据的正态特征，将与分类模型不一致的数据识别为离群点。

表 3 深度学习算法的对比

Table 3 Comparison of deep learning algorithms

算法类型	变体	算法特点	优势	局限	适合数据类型
Autoencoder	DAE，SAE，VAE	Encoder+Decoder 结构；特征提取与数据重建思想	无需数据标签；可以通过多层堆叠提高算法性能；存在较多变体模型	需对模型进行预训练，收敛速度慢	数值型、图像
生成对抗网络	AnoGAN，EGBAD，GANomaly	生成模型＋对抗模型；博弈论，生成对抗思想	无需完整数据标签；可以拟合任意数据分布；存在较多变体模型	离散数据处理效果较差；需使用平滑滤波方法提高生成样本质量；训练具有不稳定性	图像
RNN	LSTM，GRU，GRNN	有向无环的循环结构	解决数据的长依赖问题；可以学习时间序列的上下文信息；获取的序列长度可变	对于过长时间轴存在梯度消失问题；难以实现堆叠提取高维特征信息	时间序列数据

2.6 基于孤立森林方法

离群点检测中大部分算法的核心思想是尽可能寻找数据集合中的正常点，并将不符合正常点定义的数据归类为离群点。孤立森林算法（Isolation Forest，IForest）[59] 与上述算法的思想有所不同，它将异常定义为“容易被孤立的离群点（more likely to be separated）”，该算法由 Liu 等于 2008 年在第八届 IEEE 数据挖掘国际会议上提出，如图 9 所示。

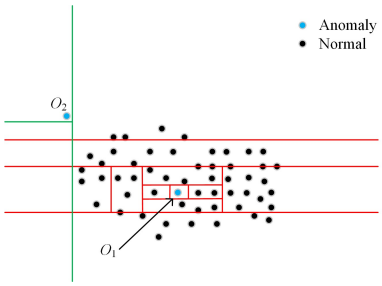


图 9 孤立森林算法

Fig. 9 Isolated forest algorithm

在数据空间中，正常点多位于密度高的簇中，需经过多次切割才能分离，而离群点一般远离其他样本点，更容易被分离，因此该算法可以专注于隔离与多数样本存在显著差异的数据，在检测过程中通过随机选定的特征递归分割数据集，根据各数据点的路径长度计算异常分数。因为离群点更容易被分离，所以路径长度相对较短，对得分进行归一化得出 0～1 的异常概率估计，越接近 1 的点越可能是离群点。

孤立森林算法具有以下优点：1）属于无监督方法，可以使用无标签样本进行训练并且准确率较高；2）属于集成学习算法，每棵孤立树单独处理数据，可以实现分布式部署，时间复杂度为 $O(n)$，数据处理效率较高。但是，该算法存在以下局限：只适用于检测全局离群点，不擅长处理局部的离群点。之后，Liu 等在 IForest 的基础上提出了 SCiForest[60]，该算法利用方差做特征选择并且加入随机超平面概念，使其更加适用于聚类异常。

2.7 小结

本章详细介绍了经典离群点检测算法的原理，分析了各类算法及其改进算法的优势与不足。表 4 列出了各类算法的优缺点、适合数据类型以及时间效率，“—”表示暂无相关数据。

表 4 离群点检测经典算法的总结

Table 4 Summary of classic outlier detection algorithms

算法分类	经典算法	优点	适合数据类型	缺点	高维空间	时间效率
基于统计	参数方法：高斯模型、回归模型、高斯混合模型等非参数方法：核密度估计法、直方图等	简单、高效	单变量且服从特定概率模型的数据集	适合单变量数据；需要了解数据分布等先验知识	不适合	受数据集和概率分布模型影响，一般不超过 $O(n\log n)$
基于深度	DEEPLOC 等	与统计方法相比数据无需服从特定数学分布	适合分层的 3 维以下数据集	当数据维度大于 3 时，计算复杂度较高	不适合	维度小于 3 时为 $O(n^{d/2})$，当数据维度增加时，算法效率变得很低，时间复杂度趋向无穷
基于距离	KNN，Index based，Nested-loop based，Cell-based 等	数据无需服从特定分布；方法简单、易理解	—	处理高维海量数据时执行效率很低；距离计算函数选择和算法的相关参数选择对异常检测效果的影响较大；难以发现局部离群点；需要预先设定超参数	高维情况下，数据稀疏，难以检测离群点	Index based 算法、Nested-loopbased 算法的时间复杂度为 $O(kn^2)$，Cell- based 算法时间复杂度为 $O(c^k+n)$

（续表）

算法分类	经典算法	优点	适合数据类型	缺点	高维空间	时间效率
基于密度	LOF,COF,INFLO 等	检测数据中的局部离群点；对数据点的异常程度进行量化	数据集聚类特征较明显	对高维数据的检测效率下降；需要预先设定超参数	可以检测，但是性能较低	—
基于聚类	K-mean,DBSCAN,CBLOF 等	无需标签和先验知识；对数据类型以及数据分布情况要求低	由特定聚类算法决定	聚类算法需要针对特定的数据情况进行选择；数据维度高并且分布稀疏时，算法的聚类效果较差；对超参数的设定敏感	高维情况下簇之间的边界区域难以识别	与具体的聚类算法有关
基于深度学习	GAN,AE 等	泛化能力强；精度高	有监督方法需要数据标签	对超参数敏感；模型复杂	可以处理	与具体算法有关
基于孤立森林	IForest,SCiForest 等	可以分布式部署实现并行运算；计算复杂度低	异常样本占比较低；正常样本与异常样本差异较大	难以检测局部离群点	不适用	$O(n)$

3 离群点检测在新环境下的应用

在数据量飞速发展的新环境下，离群点检测面临诸多传统算法难以解决的新问题，包括高维海量数据的处理以及时序数据的实时性处理。同时，基于分布式的并行处理算法在解决数据处理量以及处理速度方面具有天然优势、基于图模型的离群点检测方法也为新冠疫情背景下病毒传播渠道和传染源分析提供了新的思路。

3.1 高维数据离群点检测

随着信息采集技术的快速发展，数据维数呈指数级增长[61]，传统的离群点检测算法对高维数据处理效果差，主要原因有以下两点：1）传统的离群点检测算法大多基于数据之间的相似程度划分数据，在高维空间中，数据之间的相似程度被弱化，距离更加稀疏，基于距离或密度的度量方式在高维空间中没有明显效果；2）高维数据一般复杂度较高，采用传统检测算法直接处理高维数据执行效率极低。目前针对高维数据中的离群点检测可以使用降低维度方法[62]以及使用高鲁棒的度量函数寻找全维离群点方法[63]。降低数据维度方法的思想核心是将数据从高维映射到低维空间，然后在低维子空间中对数据进行处理。Kriegel 等[64]提出了子空间异常度评价方法（Subspace Outlier Degree，SOD），通过探索样本点邻域所跨越的平行子空间，来确定该样本点与子空间中邻域偏离的程度，以进行离群点检测。Aggarwal 等[62]提出采用遗传算法寻找最优子空间，解决了高维数据子空间组合模式多的问题。由于不同子空间可能存在不同的异常值，因此可以采用集成的方法，先检测各个子空间的异常点，再将检测结果进行综合评价得到最终结果。Keller 等[65]提出了一种高对比度的子空间处理方法（High Contrast Subspaces，HICS），该方法可以实现高对比度子空间的离群点检测，并将多个子空间结果进行合并得到最终结果。由于高维空间数据存在稀疏性，基于距离的度量方法检测效果较差，因此 Kriegel 等提出了基于角度的 ABOD 算法[63]。该算法认为在高维空间中，基于角度的度量方法比距离更加稳定，可以通过计算余弦相似度反映高维数据点的离群程度。如图 10 所示，在高维空间中，如果其他点位于样本点的同一方向且相距较远，则该样本点属于离群点。Chen 等[66]在 ABOD 算法的基础上提出了 HODA 算法，提出使用数据降维和网格划分两种方式对数据集进行处理，根据剩余网格的数据计算异常因子 ABOF，提高了算法的检测精度。上述两种方法虽然在一定程度上提高了高维数据中离群点的检测准确性，但是存在时间复杂度高的缺陷，Pham 等[67]提出使用 L1-depth 作为离群因子，以解决基于角度算法时间复杂度高的问题。

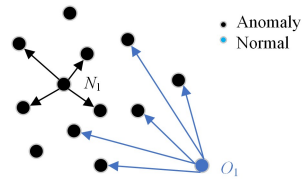


图 10 二维空间基于角度的离群点检测

Fig. 10 Angle based outlier detection in two-dimensional space

3.2 时序数据离群点检测

在传统的离群点检测过程中，数据大多以完整的形式存储于数据库，算法可以对数据进行多次、无规则提取。但是，随着数据采集规模与速度的增加，数据存储成本变得十分高昂，有些数据（如传感器数据）无需长期储存，只需要对实时产生的数据进行处理，这些按照时间顺序无限更新的数据序列就是时序数据。与传统的静态数据相比，时序数据的离群点检测过程具有以下特点：1）数据实时更新，只要不中断数据源，数据会一直更新；2）单次检测性，由于数据是源源不断的，对全部数据进行存储不切实际，大多只需要对数据进行实时检测。时序数据描述了检测对象在不同时间段的变化规律和发展趋势，可以为研究人员下一步的决策提供依据，因此，对时序数据的离群点检测是当前数据挖掘领域的一个研究重点。Varun 等[68]总结了 t-STIDE、FSA 等 7 种时序异常检测算法的优势与不足，并分析了不同算法应用于具体领域数据集的表现情况。Hawkins 等[69]提出了分层时间记忆算法（Hierarchical Temporal Memory，HTM），该算法模拟人脑皮层的工作原理，可以对复杂的数据集进行模式识别。Ahmad 等[70]提出将 HTM 算法用于时序数据异常检测中，并通过实验证明了该方法的有效性。清华大学的 Xu 等[71]提出了基于关联差异的异常检测模型 Anomaly Transformer，如图 11 所示，该模型使用了一种全新的注意力机制 Anomaly-Attention 来统一建模先验关联和序列关联，并且引入极小极大（Mini max）关联学习策略增大正常点和异常点的差异，在服务器监测、地空探索、水流观测等时序数据应用中，该模型均展现出了优秀的异常检测结果，具有很强的应用落地价值。

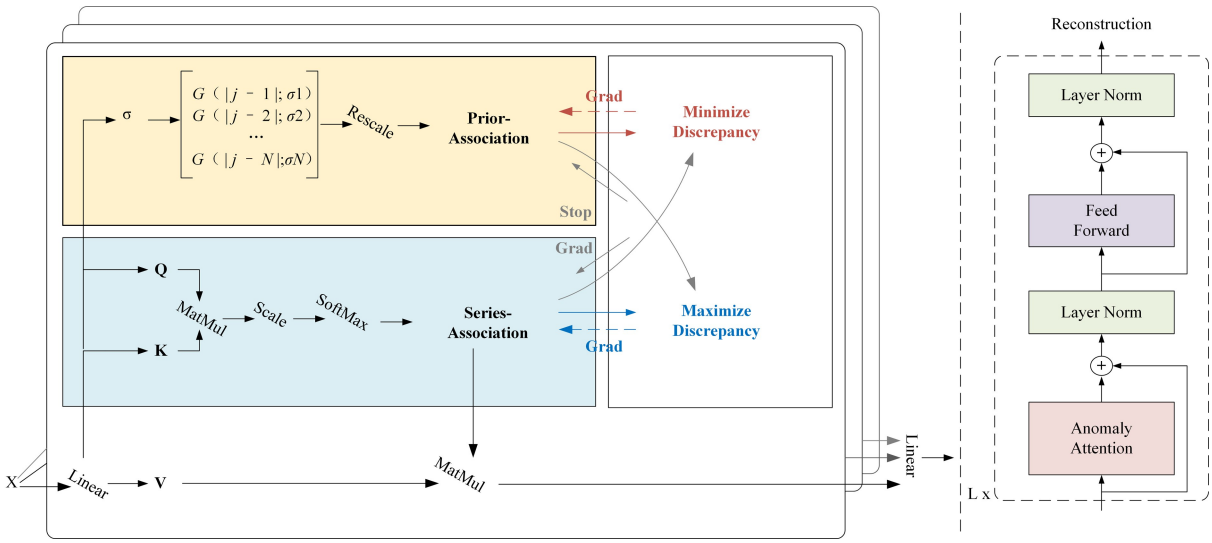


图 11 Anomaly Transformer 模型的结构

Fig. 11 Structure of Anomaly Transformer model

3.3 基于分布式的离群点检测

随着数据规模与数据维度的不断增加，传统依靠单台服务器进行串行计算的方法难以应对大数据背景下的数据挖掘挑战，这促进了分布式计算和云计算在数据挖掘领域的快速发展。分布式技术将一个业务拆分成多个子业务，部署在多个服务器上进行计算，可以极大提高数据的处理效率，解决由计算中心化导致的数据处理效率低以及计算资源浪费的问题。正是基于以上优势，分布式计算也被广泛应用于离群点检测领域，目前常用的分布式计算框架有 MapReduce[72] 和 Spark[73] 等。MapReduce 由谷歌团队 Dean 等于 2004 年提出，如图 12 所示[74]，它有两个关键步骤，即 Map 阶段和 Reduce 阶段，Map 阶段将待处理数据划分成多个 split 块，选取集群中的任意空闲节点对每个 split 块进行处理，Reduce 阶段将 Map 阶段切片处理后的各节点运行结果进行归并计算。MapReduce[75] 具有可靠性、高扩展性、低成本等优点，但是存在以下缺陷：1）所有计算都需要转换成 Map 和 Reduce 两个操作，不能适用所有应用场景，对于复杂的数据处理过程难以描述，表达能力有限；2）待处理数据存储在磁盘中，I/O 开销大；3）计算延迟高，不能胜任复杂、多阶段的计算服务。Spark[76] 最早来源于 UC Berkeley AMP lab 发表的一篇论文，该论文提到了一种弹性分布式数据集（RDD）的概念。RDD 是一种分布式的内存抽象，能够在大规模数据集群中进行内存运算，并且有一定的容错能力，Spark 的核心数据结构就是 RDD。

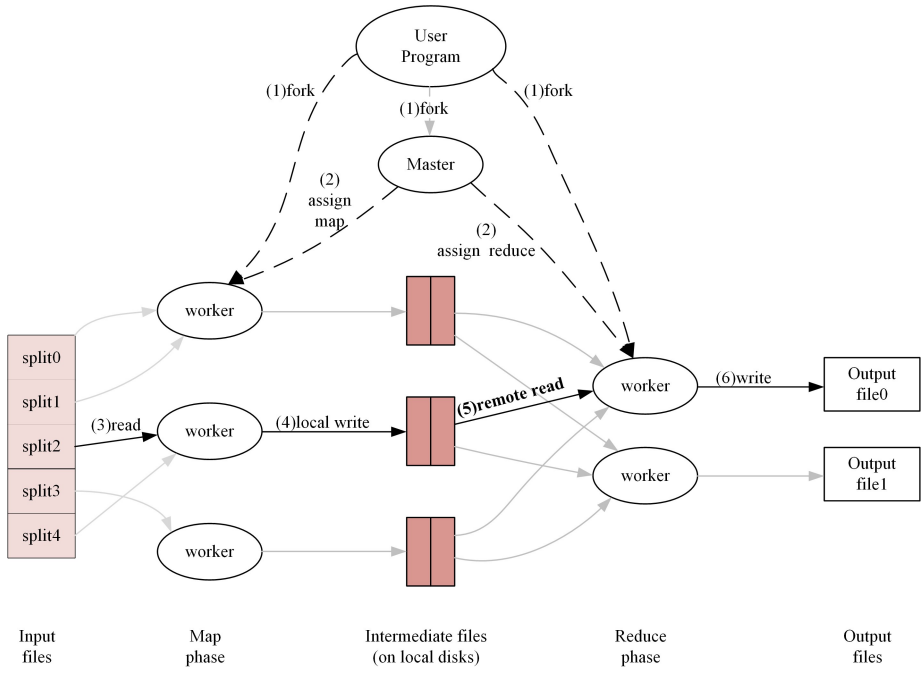


图 12 MapReduce 执行概述

Fig. 12 MapReduce execution overview

图 13 给出了 Spark 的任务调度过程。Spark 继承了 MapReduce 的优势并改进了其明显缺陷：1）基于内存进行数据

处理，将需要反复利用的中间结果保存到内存中，提高了数据的处理速度；2）Spark 通过引入有向无环图（DAG）进行分布式并行计算，减少了迭代过程中的磁盘 I/O 操作；3）Task 以线程的方式进行维护，任务启动快，可以批量创建，提高并行能力。尽管 Spark 存在以上优势，但是它并不能完全取代 MapReduce 的地位，因为 Spark 基于内存进行数据处理，对内存和 CPU 的性能要求很高，大多适用于数据量小、对实时性要求高的场景，而 MapReduce 可以使用廉价的通用服务器来搭建分布式运算集群。因此，对于数据量特别大、实时性要求不高的应用场景，使用 MapReduce 更合适。

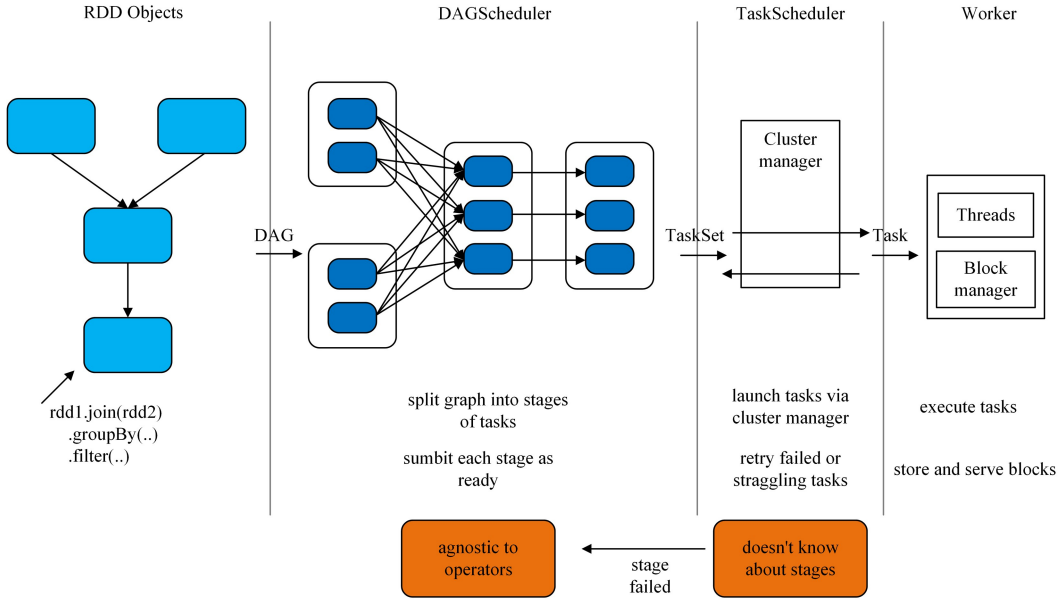


图 13 Spark 任务调度流程

Fig. 13 Spark task scheduling process

3.4 基于图模型的离群点检测

图作为一种主流的数据结构，被广泛用于表达结构化数据中包含的复杂信息。基于图的离群点检测方法通过建立数学模型，将数据转化为图的表达形式，图中的每个节点表示样本点，边表示样本点之间的抽象关系。该方法不仅可以发现数据之间的相似度信息，还能发现数据之间的关系信息，有效捕捉数据对象之间的长期依赖属性，直观表现实体与实体之间的联系。如图 14 所示，根据可用的数据标签、输入网络的性质以及检测的异常类型，可将图结构方法分为基于社区方法、基于概率方法等[77]。Ma 等[78]对基于图的离群点检测技术和方法进行了综述，强调了采用图方法进行离群点检测技术的重要性，它可以显示数据之间的相互依赖状态，是一个表现直观并且十分稳定的异常检测工具。Chen 等[79]从静态图和动态图的角度出发，梳理了基于深度学习的图异常检测研究现状，重点分析了基于深度神经网络的图表示方法。Moonesinghe 等[80]提出了基于图的离群点检测框架 Outrank，该框架根据原始数据集拓展完全链接的无向图，并在预定义的图上应用马尔可夫随机游走方法，使用随机游走的平稳分布值作为异常值得分。Bandyopadhyay 等[81]提出了 DONE 算法和 AdONE 算法，如图 15 所示，采用深度自编码器来捕捉图模型中结构和属性的重构损失。Su 等[82]分析了基于复杂网络的离群点检测研究进展。基于图的离群点检测方法虽然有一定优势，但在当前的研究过程中也存在诸多挑战，例如使用图方法需要首先构建数据的网络表示，但通常很难预测最适合使用图结构的数据种类。因此，图方法是未来离群点检测领域的重要研究方向之一。

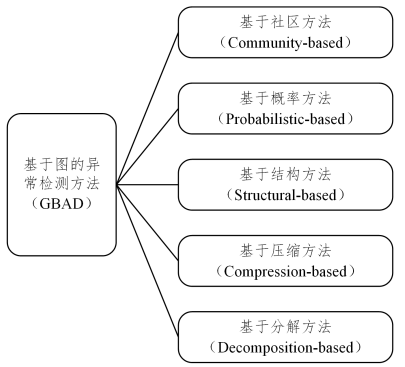


图 14 5 种 GBAD 方法

Fig. 14 Five GBAD methods

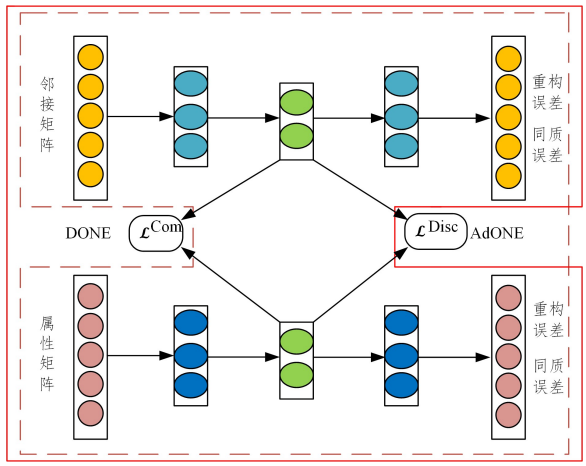


图 15 DONE 和 AdONE 模型的结构

Fig. 15 Structure of DONE and AdONE models

3.5 模糊聚类算法在离群点检测的应用

传统的离群点检测方法通常采用“硬划分”的方式，将每个样本归属于一个单独的类别。这种方法仅考虑样本所属类簇内的样本，缺乏与其他类簇内样本的对比。然而，在实际情况中，离群实体的行为既具有“伪装性”又具有“复杂性”，因此仅在一个类簇内进行分析是不够的。

为了解决这个问题，可以采用模糊聚类方法[83]对离群实体行为数据进行处理。模糊聚类利用模糊数学方法对数据进行分析，并生成样本类别的隶属度矩阵，使得一个样本可以同时属于多个类别。通过应用模糊聚类算法处理离群实体行为数据，可以更好地保留样本与各个类簇之间的关联信息。

相比传统的方法，基于模糊聚类的离群检测方法能够综合考虑不同类别的样本特征，识别出更复杂和隐蔽的异常行为。通过对离群实体行为数据进行模糊聚类处理，可以更全面地评估样本的离群程度，并得出更准确的离群检测结果[84]。

4 离群点检测评价指标与工具集

4.1 评价指标

对离群点检测算法的有效性评估是检验算法优劣的重要方法，可以把离群点检测结果当作二分类问题来评价，如表5所列，将混淆矩阵作为基础构建模型的评价指标，包括准确率、召回率、真正率等。

表5 混淆矩阵

Table 5 Confusion matrix

		标签类别	
		离群点	正常点
预测类别	离群点	TP	FP
	正常点	FN	TN

如式(1)所示，精确率(Precision)指算法将离群点样本预测正确的数量占所有预测为离群点数量的比例。如式(2)所示，召回率(Recall)指算法正确预测为离群点的样本数量占所有真实为离群点数量的比例。以上两种指标的值越高，算法的效果越好。如式(3)所示，假正率(False Positive Rate，FPR)指将正常点预测为离群点的数量占正常点数量的比例，值越低表示算法的效果越好。在真实情况下，不同类别样本的分类代价是不相同的，例如在垃圾邮件检测过程中，为防止重要邮件的误删，我们一般选用Precision较高的模型，但这容易将一些垃圾文件保留，导致Recall降低。而在癌症诊断中，我们宁愿误判也不希望真实癌症患者被漏诊，因此会选择Recall高的模型，但这容易将健康的人误诊为癌症患者，导致Precision降低。精确率与召回率是相反的，在异常检测中如何构建一个使两者均衡的分类器是目前面临的挑战。

$$Precision=TP/(TP+FP) \tag{1}$$

$$Recall=TPR=TP/(TP+FN) \tag{2}$$

$$FPR=FP/(TN+FP) \tag{3}$$

如图16所示，为权衡Precision与Recall的精度，本文使用ROC曲线。通过使用不同的分类阈值得到多组混淆矩阵，将混淆矩阵计算得到的TPR作为纵坐标，将FPR作为横坐标绘制坐标轴，曲线越靠近坐标轴的左上角，表明算法效果越好(图16中曲线a优于曲线b)。不同的ROC曲线相交时存在难以明显区分算法优劣的问题，可以计算ROC曲线下方的面积AUC(Area Under the Curve)，对算法性能进行量化，以更好地反映算法效果。

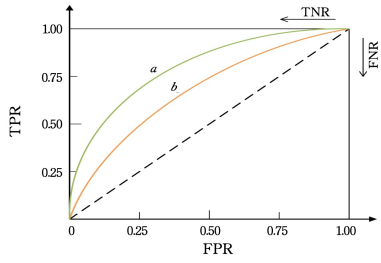


图16 ROC曲线

Fig. 16 ROC curve

4.2 代价因子在离群点检测评价中的作用

代价因子[85]是一种用于量化离群点检测算法性能的指标，它反映了将一个样本误分类为离群点或正常点所带来的代价或成本。例如，在某些离群点检测任务中，将正常点错误地标记为离群点(FP)可能带来较大的成本，而将离群点错误地标记为正常点(FN)的成本可能较小。因此，可以使用代价因子来调整混淆矩阵中各个元素的权重，以更准确地评估离群点检测算法的性能。除此之外，代价因子还具有以下作用[86]。

1)性能度量：代价因子可以用来度量离群点检测算法的性能。通过比较算法的代价因子，可以评估算法的准确性和效果。代价因子可以是真实离群点的漏检率、误检率、精确度、召回率等指标，这些指标反映了算法在检测离群点时的表现。

2)决策依据：代价因子可以作为决策依据，帮助确定阈值或决策边界。通过分析代价因子，可以设定一个适当的阈值或决策边界来判断数据点是否为离群点。代价因子有助于权衡误检率和漏检率，根据应用需求选择合适的阈值，以平衡检测的准确性和召回率。

3)算法比较：代价因子可以用于比较不同离群点检测算法的性能。通过计算和比较代价因子，可以评估不同算法在不同数据集上的表现。代价因子可以帮助选择最适合特定任务和数据集的离群点检测算法，并提供有关算法性能的量化指标。

4)优化算法：代价因子可以用于优化离群点检测算法的参数选择和模型调整。通过优化代价因子，可以改进算法的准确性和效率，提高离群点检测的性能。代价因子可以用作目标函数或约束条件，通过调整算法参数来最小化代价因子，从而优化算法的性能。

4.3 离群点检测工具

在实际应用领域，已经产生诸多离群点检测的科学工具包，具体如下。

1)Scikit-learn Outlier Detection

Scikit-learn[87]是Bisong于2007年基于SciPy构建的专门用于机器学习的开源框架，它提供了诸多机器学习的离群点检测算法，例如LOF和Isolation Forest等。该工具包安装和使用十分容易，并且有详细的教程和使用文档。

2)Python Outlier Detection(PyOD)

PyOD[88]是一个全面且可扩展的 Python 工具包,常用于检测多源数据中的离群点,在科学研究和商业项目中应用广泛,该工具包含 30 多种离群点检测算法,包括经典的 LOF 算法以及最新的 SUOD 算法等。

3)ELKI

ELKI[89]是一种基于 Java 编写的开源(AGPLv3)数据挖掘工具,它的重点是算法研究,包括聚类分析和离群点检测中的无监督算法。

4)MATLAB

MATLAB 是一个由美国 Math Works 公司开发并进行维护的数学分析软件,在机器学习和深度学习领域应用广泛。它包含离群点检测的诸多经典算法如 LOF,IForest 等,并且具有良好的用户交互性,算法可以直接在 MATLAB 中运行。

4.4 数据集

标准的离群点检测数据集对于评价算法的性能至关重要[90],基于离群点检测的应用场景,需要时序、高维、不规则等多种类型数据,目前主流的离群点检测数据集如下。

1)The UCI Repository

UCI 数据库提供了数百个公开数据集,由加州大学欧文分校提出并进行维护,离群点检测常使用该数据库来评估算法的性能。但是,这些数据集大多基于分类算法设计,在离群点检测场景中,一般需要按照分类思想对数据集进行检测。

2)Outlier Detection Datasets(ODDS)

ODDS 与 UCI 数据库不同,它只提供对离群点检测的数据集,拥有不同应用领域的多种类型数据,包括医学和工业等领域的多维、时间序列,以及单变量和多变量的数据集。

3)ELKI Outlier Datasets

ELKI 工具包中有一套专门用于异常检测和算法评估的数据集,常被用于评估离群点检测算法和参数的性能。

5 总结与展望

本文系统总结了离群点检测算法的发展历程,详细介绍了经典离群点检测算法的优点与不足。下文展望了离群点检测算法的前沿问题和研究方向,以促进该领域进一步完善与发展。

1)时序、高维数据离群点挖掘

随着互联网技术迅猛发展,数据规模呈现井喷式增长,部分数据的维度可能达到上百维,同时时序数据也广泛存在于传感器网络、视频监控、通信转换、金融服务等领域[91]。因此,对高维数据离群点实现精准检测以及对时序数据离群点实现实时性检测是目前一个巨大的挑战。

2)对离群点的解释性问题

离群点检测方法大多是一个“黑盒”模型,对于该点为什么属于离群点的解释性工作开展相对较少,这违背了离群点检测的初衷。因此,下一步对离群点产生的原因进行合理分析和解释尤为重要。

3)可扩展的检测模型

纵向来看,随着时间的推移和研究的深入,对离群点的评判指标可能发生变化,数据也会增加新的特征或者剔除旧的特征,这可能导致过去优秀的模型精确度降低,重新训练新的检测模型成本非常高,并且将失去对原有数据信息的理解,因此要求模型具有持续学习能力;横向来看,针对某个特定领域开发的检测模型一般只适用于该领域,如果能提高模型在不同领域的兼容性,实现模型复用,将极大降低成本。因此,模型的可扩展性是一个值得研究的重要方向。

4)数据可视化技术

经过分析和处理的数据结果应及时交由分析人员审阅,并使其能更加直观地获取其中信息。因此,优秀的数据可视化技术十分重要。

5)深度学习在离群点检测中的应用

深度学习在其他应用领域已经取得相当大的进展,从目前的研究趋势来看,基于深度学习的离群点检测方法同样具有一定优势,但是研究相对较少。因此,深度学习技术在离群点检测中的应用仍有待进一步发展。

结束语 离群点挖掘现已被广泛应用于各个领域,各种检测技术飞速发展并且日趋成熟,但仍存在完善与发展的空间,尤其是结合深度学习方法以解决传统经典算法在新环境下面临的问题与挑战,实现性能和精度上的提升。本文首先介绍了离群点的成因以及分类,对主流离群点检测算法的优势与不足进行了总结,分析了新环境下离群点检测的应用,最后针对前沿领域对未来的发展方向进行了总结和展望。

参 考 文 献

[1] HAWKINS D M. Identification of outliers [M]. Vol. 11. London: Chapman and Hall, 1980.

[2] WANG H, BAH M J, HAMMAD M. Progress in outlier detection techniques: A survey[J]. IEEE Access 7(2019): 107964-108000.

[3] JIANG F, WANG K L, YU X, et al. Summary of Intrusion Detection Models Based on Deep Learning[J]. Control and Decision, 2020, 35(5): 1199-1204.

[4] ZHANG W A, HONG Z, ZHU J W, et al. A survey of network intrusion detection methods for industrial control systems[J]. Control and Decision, 2019, 34(11): 2277-2288.

[5] CHENG Z, CHAI S. A cyber intrusion detection method based on focal loss neural network[C]// 2020 39th Chinese Control Conference(CCC). IEEE, 2020.

[6] ZHOU Y J, HE P F, QIU R F, et al. Research on Intrusion Detection Based on Random Forest and Gradient Boosting Tree [J]. Journal of Software, 2021, 32(10): 3254-3265.

[7] LIU Y, YANG K. Credit Fraud Detection for Extremely Imbalanced Data Based on Ensembled Deep Learning[J]. Journal of Computer Research and Development, 2021, 58(3): 539-547.

[8] POURHABIBI T, ONG K L, KAM B H, et al. Fraud detection: A systematic literature review of graph-based anomaly detection approaches[J]. Decision Support Systems, 2020, 133: 113303.

[9] AL-HASHEDI K G, MAGALINGAM P. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019 [J]. Computer Science Review2021, 40: 100402.

[10] FIORE U, AD S, PERLA F, et al. Using generative adversarial

networks for improving classification effectiveness in credit card fraud detection[J]. Information Sciences,2019,479:448-455.

[11] FERNANDO T,GAMMULLE H,DENMAN S,et al. Deep learning for medical anomaly detection-a survey[J]. ACM Computing Surveys(CSUR),2021,54(7):1-37.

[12] HAN C,RUNDO L,MURAO K,et al. MADGAN:Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction[J]. BMC bioinformatics,2021,22(2):1-20.

[13] SHVETSOVA N,BAKKER B,FEDULOVA I,et al. Anomaly detection in medical imaging with deep perceptual autoencoders [J]. IEEE Access,2021,9:118571-118583.

[14] POORNIMA I,PARAMASIVAN B. Anomaly detection in wireless sensor network using machine learning algorithm[J]. Computer communications,2020,151:331-337.

[15] FRANCESCO C,GIANCARLO F,ANTONIO G,et al. Short-long term anomaly detection in wireless sensor networks based on machine learning and multi-parameterized edit distance[J]. Information Fusion,2019,52:13-30.

[16] ZHOU J T,DU J,ZHU H,et al. Anomalynet:An anomaly detection network for video surveillance[J]. IEEE Transactions on Information Forensics and Security,2019,14(10):2537-2550.

[17] SULTANI W,CHEN C,SHAH M. Real-world anomaly detection in surveillance videos[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[18] CHANDOLA V,BANERJEE A,KUMAR V. Anomaly detection:A survey[J]. ACM computing surveys(CSUR),2009,41(3):1-58.

[19] XU X,LIU J W,LUO X L. Research on outlier mining[J]. Application Research of Computers,2009,26(1):34-40.

[20] XUE A R,YAO L,JU S G,et al. Survey of Outlier Mining[J]. Computer Science,2008(11):13-18,27.

[21] MEI L,ZHANG F L,GAO Q. Overview of outlier detection technology[J]. Application Research of Computers,2020,37(12):3521-3527.

[22] WU J F,JIN W D,TANG P. Survey on Monitoring Techniques for Data Abnormalities[J]. Computer Science,2017,44(S2):24-28.

[23] LEI H L,TUERHONG G,WUSHOUER M,et al. Review of Novelty Detection[J]. Computer Engineering and Applications,2021,57(5):47-55.

[24] JOHNSON T,KWOK I,NG R T. Fast Computation of 2-Dimensional Depth Contours[C]// KDD. 1998:224-228.

[25] KNOX E M,NG R T. Algorithms for mining distancebased outliers in large datasets[C]// Proceedings of the International Conference on Very Large Data Bases. 1998:392-403.

[26] RAMASWAMY S,RASTOGI R,SHIM K. Efficient algorithms for mining outliers from large data sets[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. 2000.

[27] ESTER M,KRIEGEL H P,SANDER J,et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// KDD. 1996:226-231.

[28] ERTÖZ L,STEINBACH M,KUMAR V. Finding topics in collections of documents:A shared nearest neighbor approach[J]. Clustering and information retrieval. Springer, Boston, MA, 2004:83-103.

[29] GUHA S,RASTOGI R,SHIM K. ROCK:A robust clustering algorithm for categorical attributes[J]. Information systems, 2000,25(5):345-66.

[30] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967: 281-297.

[31] KOHONEN T. Self-organization and associative memory[M]. Springer Science & Business Media. 2012.

[32] HE Z,XU X,DENG S. Discovering cluster-based local outliers [J]. Pattern recognition letters,2003,24(9/10):1641-1650.

[33] AMER M,GOLDSTEIN M. Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer[C]// Proceedings of the 3rd RapidMiner Community Meeting and Conference(RCOMM 2012). 2012:1-12.

[34] MUHAMMAD M,DANIEL ANI U,ABDULLAHI A A,et al. Device-Type Profiling for Network Access Control Systems using Clustering-Based Multivariate Gaussian Outlier Score [C]// The 5th International Conference on Future Networks & Distributed Systems. 2021.

[35] ALHUSSEIN I,ALI A H. Application of DBSCAN to Anomaly Detection in Airport Terminals[C]// 2020 3rd International Conference on Engineering Technology and its Applications (IICETA). IEEE,2020.

[36] ANKERST M,BREUNIG M M,KRIEGEL H P,et al. OPTICS:Ordering points to identify the clustering structure[J]. ACM Sigmod Record,1999,28(2):49-60.

[37] BREUNIG M M,KRIEGEL H P,NG R T,et al. LOF:identifying density-based local outliers[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. 2000.

[38] XU X,LEI Y,ZHOU X. A lof-based method for abnormal segment detection in machinery condition monitoring[C]// 2018 Prognostics and System Health Management Conference(PHM-Chongqing). IEEE,2018.

[39] TANG J,CHEN Z,FU A W C,et al. Enhancing effectiveness of outlier detections for low density patterns[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin,Heidelberg,2002.

[40] JIN W,TUNG A K H,HAN J,et al. Ranking outliers using symmetric neighborhood relationship[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer,Berlin,Heidelberg,2006.

[41] KRIEGEL H P,KRÖGER P,SCHUBERT E,et al. LoOP:local outlier probabilities[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009.

[42] PAPADIMITRIOU S,KITAGAWA H,GIBBONS P B,et al. Loci:Fast outlier detection using the local correlation integral [C]// Proceedings 19th International Conference on Data Engineering(Cat. No. 03CH37405). IEEE,2003.

[43] TANG B,HE H. A localdensity-based approach for outlier de-

tection[J]. Neurocomputing, 2017, 241: 171-180.
[44] KIRAN B R, THOMAS D M, PARAKKAL R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos[J]. Journal of Imaging, 2018, 4(2): 36.
[45] CHEN Z, YEO C K, LEE B S, et al. Autoencoder-based network anomaly detection[C]// 2018 Wireless Telecommunications Symposium(WTS). IEEE, 2018.
[46] WU Y K, LI W, NI M Y, et al. Anomaly Detection Model Based on One-class Support Vector Machine Fused Deep Auto-encoder [J]. Computer Science, 2022, 49(3): 144-151.
[47] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010, 11(12): 3371-3408.
[48] DOERSCH C. Tutorial on variational autoencoders[J]. arXiv: 1606.05908, 2016.
[49] ZHANG C H, ZHOU X T, ZHANG Y A, et al. Application Research of Deep Auto Encoder in Data Anomaly Detection[J]. Computer Engineering and Applications, 2020, 56(17): 93-99.
[50] DI MATTIA F, GALEONE P, DE SIMONI M, et al. A survey on gans for anomaly detection[J]. arXiv: 1906.11632, 2019.
[51] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery[C]// International Conference on Information Processing in Medical Imaging. Cham: Springer, 2017: 145-157.
[52] ZENATI H, FOO C S, LECOUAT B, et al. Efficient gan-based anomaly detection[J]. arXiv: 1802.06222, 2018.
[53] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks[J]. Medical Image Analysis, 2019, 54: 30-44.
[54] DONAHUE J, KRÄHENBÜHL P, DARRELL T. Adversarial feature learning[J]. arXiv: 1605.09782, 2016.
[55] AKCAY S, ATAPOUR-ABARGHOUEI A, BRECKON T P. Ganomaly: Semi-supervised anomaly detection via adversarial training[C]// Asian Conference on Computer Vision. Cham: Springer, 2018.
[56] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]// International Conference on Machine Learning. PMLR, 2017.
[57] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017.
[58] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[J]. arXiv: 1409.2329, 2014.
[59] LIU F T, TING K M, ZHOU Z H. Isolation forest[C]// 2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008: 413-422.
[60] LIU F T, TING K M, ZHOU Z H. On detecting clustered anomalies using sciforest[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2010.
[61] ZHONG Y Y, CHEN S C. High-order Multi-view Outlier Detection[J]. Computer Science, 2020, 47(9): 99-104.
[62] AGGARWAL C C, YU P S. Outlier detection for high dimensional data[C]// Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data. 2001.
[63] KRIEGEL H P, SCHUBERT M, ZIMEK A. Angle-based outlier detection inhigh-dimensional data[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008.
[64] KRIEGEL H P, KRÖGER P, SCHUBERT E, et al. Outlier detection in axis-parallel subspaces of high dimensional data[C]// Pacific-asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2009.
[65] KELLER F, MULLER E, BOHM K. HiCS: High contrast subspaces for density-based outlierranking[C]// 2012 IEEE 28th International Conference on Data Engineering. IEEE, 2012.
[66] CHEN S N, QIAN H Y, LI W. Hybrid outlier detection algorithm based on angle variance for high-dimensional data[J]. Application Research of Computers, 2016, 33(11): 3383-3386.
[67] PHAM N. L1-depth revisited: A robust angle-based outlier factor in high-dimensional space[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer, 2018.
[68] CHANDOLA V, MITHAL V, KUMAR V. Comparative evaluation of anomaly detection techniques for sequence data[C]// 2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008.
[69] HAWKINS J, AHMAD S. Why neurons have thousands of synapses, a theory of sequence memory in neocortex[J]. Frontiers in neural circuits, 2016, 10: 174222.
[70] AHMAD S, LAVIN A, PURDY S, et al. Unsupervised real-time anomaly detection for streaming data[J]. Neurocomputing, 2017, 262: 134-147.
[71] XU J, WU H, WANG J, et al. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy[J]. arXiv: 2110.02642, 2021.
[72] DEAN J, GHEMAWAT S. MapReduce: Simplified data processing on large clusters[J]. Communications of ACM, 2008, 51(1): 107-113.
[73] ZAHARIA M, CHOWDHURY M, DAS T, et al. Resilient Distributed Datasets: A {Fault-Tolerant} Abstraction for {In-Memory} Cluster Computing[C]// 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12). 2012.
[74] KANNA P R, SANTHI P. Hybrid intrusion detection using mapreduce based black widow optimized convolutional long short-term memory neural networks[J]. Expert Systems with Applications, 2022, 194: 116545.
[75] FATHNIA F, BARAZESH M R, BAYAZ M H J D. Runtime Optimization of a New Anomaly Detection Method for Smart Metering Data Using Hadoop Map-Reduce[C]// 2019 International Power System Conference(PSC). IEEE, 2019.

[76] ALNAFESSAH A,CASALE G. Artificial neural networks based techniques for anomaly detection in Apache Spark[J]. Cluster Computing,2020,23(2):1345-1360.

[77] POURHABIBI T,ONG K L,KAM B H,et al. Fraud detection: A systematic literature review of graph-based anomaly detection approaches[J]. Decision Support Systems,2020,133:113303.

[78] MA X,WU J,XUE S,et al. A comprehensive survey on graph anomaly detection with deep learning[J]. IEEE Transactions on Knowledge and Data Engineering,2021,35(12):12012-12038.

[79] CHEN B F,LI J D,LU X J,et al. Survey of Deep Learning Based Graph Anomaly Detection Methods[J]. Journal of Computer Research and Development,2021,58(7):1436-1455.

[80] MOONESINGHE H D K,TAN P N. Outrank:a graph-based outlier detection framework using random walk[J]. International Journal on Artificial Intelligence Tools, 2008,17(1): 19-36.

[81] BANDYOPADHYAY S,VIVEK S V,MURTY M N. Outlier resistant unsupervised deep architectures for attributed network embedding[C]// Proceedings of the 13th International Conference on Web Search and Data Mining. 2020.

[82] SU J,DONG Y H,YAN M J,et al. Research progress of anomaly detectionfor complex networks[J]. Control and Decision, 2021,36(6):1293-1310.

[83] MOJARAD M,NEJATIAN S,PARVIN H,et al. A fuzzy clustering ensemble based on cluster clustering and iterative Fusion of base clusters[J]. Applied Intelligence,2019,49:2567-2581.

[84] GUO Y L,ZUO X J,CUI J Y. An abnormal behavior detection algorithm based on fuzzy clusteringfor multi-categories affiliation of power entities[J]. Journal of Hebei University of Science and Technology,2022,43(5):528-537.

[85] CHEN Z,SHENG V,EDWARDS A,et al. An effective cost-sensitive sparse online learning framework for imbalanced streaming data classification and its application to online anomaly detection[J]. Knowledge and Information Systems, 2023, 65(1):59-87.

[86] CHEN X,LIU H,XU X,et al. Identification of Suitable Technologies for Drinking Water Quality Prediction:A Comparative Study of Traditional, Ensemble, Cost-Sensitive, Outlier Detection Learning Models and Sampling Algorithms[J]. ACS ES&T Water,2021,1(8):1676-1685.

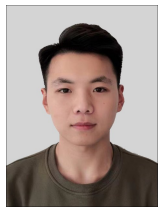
[87] BISONG E. Introduction to Scikit-learn[C]// Building machine learning and deep learning models on Google cloud platform. Apress,Berkeley,CA,2019:215-229.

[88] ZHAO Y,NASRULLAH Z,LI Z. Pyod:A python toolbox for scalable outlier detection[J]. arXiv:1901.01588,2019.

[89] SCHUBERT E,ZIMEK A. ELKI:A large open-source library for data analysis-ELKI Release 0.7.5 "Heidelberg"[J]. arXiv: 1902.03616,2019.

[90] FU L F,CHEN Z,AO C L. Dynamic outlier detection algorithm for network large data set based on classification and regression trees decision tree[J]. Journal of Jilin University(Engineering and Technology Edition),2023,53(9):2620-2625.

[91] HUANG J R,WANG Q,CAI X J,et al. Multi-objective Adaptive DBSCAN Outlier Detection Algorithm[J]. Journal of Chinese Computer Systems,2022,43(4):702-706.



KONG Lingchao, born in 1998, postgraduate, is a student member of CCF (No. G8696G). His main research interests include data mining and fault detection.

LIU Guozhu, born in 1965, Ph.D, professor, master supervisor. His main research interests include network security and fault detection.

(责任编辑:喻藜)