

硕 士 研 究 生 读 书 报 告



题目  异常检测算法之孤立森林

作者姓名 黄闯

作者学号 21860451

指导教师  才振功

学科专业 计算机技术

所在学院 工程师学院

提交日期 2018年12月6日

Anomaly detection algorithm for isolated forest

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Cai zhen gong

By

Huang Chuang

Zhejiang University, P.R. China

2018

**摘要**

对一个数据集检测样本和对一个一般的二分类或者多分类数据集进行分类存在很大的不同，一般的数据集分类，无论是二分类还是多分类，一般来说，各类样本相对来说会样本量会比较均衡，但是异常点却会比较少，而且和正常点存在很大的不同。孤立森林基于异常点这一特性，利用这两个方面构造分类树，能够检测出异常值。孤立森林的时间复杂度能够达到线性，而且只需要很小的内存需求，而且其原作者表示，其表现超过随机森林，LOF，oneclass-svm，ORCA等异常检测方法，本文将对孤立森林进行原理剖析，适用领域以及数据集需要具备哪些特征，优劣势分析以及本质上的总结，去探索这种新方法真实面貌和随机森林之间的区别。

**关键词**：异常检测，孤立森林，异常数据集，决策树，随机森林

**Abstract**

There is a big difference between the detection of a dataset and the classification of a general binary or multi-categorical dataset. The general dataset classification, whether it is a two-category or a multi-classification, generally speaking, the relatives of the various types of samples The sample size will be more balanced, but the abnormal point will be less, and it is very different from the normal point. Isolated forests are based on the characteristics of anomalous points. Using these two aspects to construct a classification tree can detect outliers. The time complexity of isolated forests can be linear, and only requires a small memory requirement, and the original author said that its performance exceeds the random forest, LOF, oneclass-svm, ORCA and other anomaly detection methods. Anatomy, applicable areas and datasets need to have characteristics, advantages and disadvantages analysis and essential summaries to explore the difference between the true face of this new method and the random forest.

**Keywords**: anomaly detector, isolation forest，Abnormal data set, decision tree, random forest

简介

在诸多领域需要，都需要异常检测来保证系统的稳定性。除了大型服务集群需要异常检测，还有网络安全，对于安全流量的监控，也需要异常检测方面的支持，来保证网络的安全；还有机械系统，传感器数据往往非常多，而根据传感器数据来判断局部和整体安全，也是非常有必要的一个方向，异常检测就派上用场；还有化工行业，天文地理，水文监测，大气指数，可以说异常检测无处不在，相信不久的未来，异常检测将会应用的更加的成熟，来保证我们的生活安全和工业安全。

异常检测的需要，自然造就了今天在异常检测领域出现很多的检测方法，而这些方法，其发明者都会给他们命名，比如STL时序数据分解技术，其论文发表于1990年；比如EWMA方法，其论文发表于1987年；比如oneclass—svm，其论文发表于2001年，其起初的研究领域用于文本分析；还有LOF，其论文发表于2000年，其最初就是用来对异常点进行检测，不过只是基于密度的，也就是说，时序数据肯定是不行的了，那么孤立森林了，其论文最早发表于2008年，于2012年又发表了一个扩展版的论文，可以说是一个很年轻的算法了，孤立森林的发表，从一开始就是用来进行异常检测。

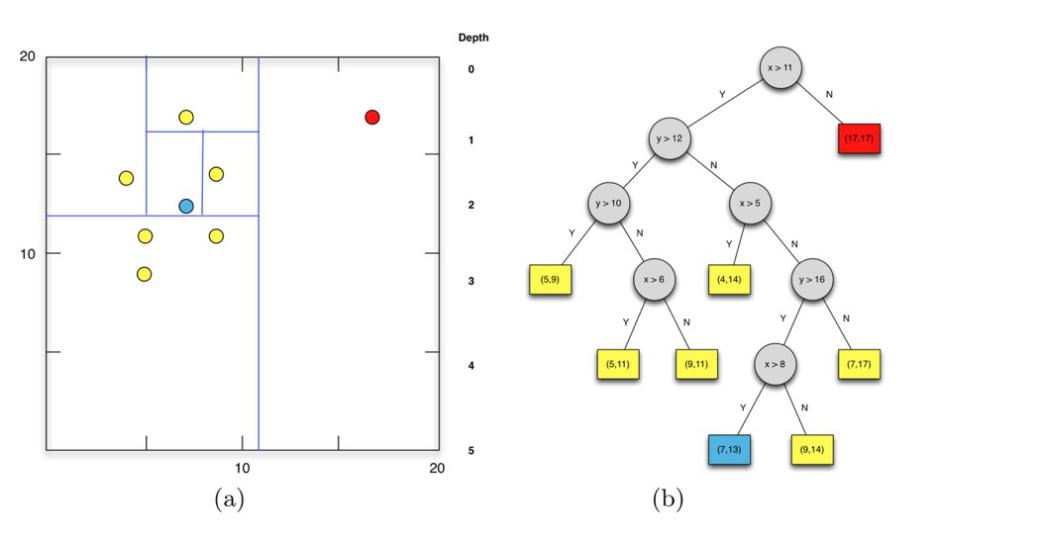
一、孤立森林简介

iForest 是[刘飞](https://link.jianshu.com?t=https%3A%2F%2Ffeitonyliu.wordpress.com%2Fabout%2F" \t "_blank)博士(Fei Tony Liu)在莫纳什大学就读期间由[陈开明](https://link.jianshu.com?t=https%3A%2F%2Ffederation.edu.au%2Ffaculties-and-schools%2Ffaculty-of-science-and-technology%2Fstaff-profiles%2Finformation-technology%2Fkai-ming-ting" \t "_blank)(Kai-Ming Ting)教授和[周志华](https://link.jianshu.com?t=http%3A%2F%2Fcs.nju.edu.cn%2Fzhouzh%2F" \t "_blank)(Zhi-Hua Zhou)教授指导发表的。第一个版本是在2008年ICDM上，获得年度最佳论文，扩充版本发表于TKDD。

传统的异常检测算法都是通过数据拟合出来一个模型，然后用现有数据测试，然后看看是不是异常点。但是孤立森林却不是这样，而是通过一个简单的超平面划分数据集，来检测异常点的存在。

其思想大致是：用一个随机超平面来分割数据空间，切一次可以生成两个子空间，然后继续用一个随机超平面来切割每个子空间，循环下去，直到每个子空间只有一个数据点为止。

如图所示：



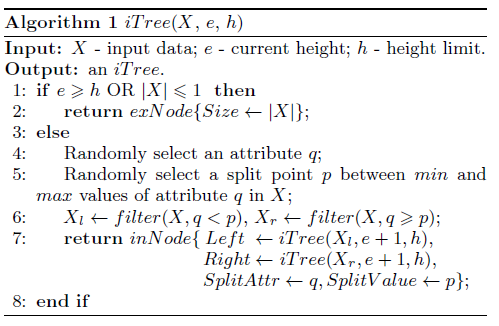
1. 孤立森林核心算法

孤立森林是如何创造出来的，和以前的方法有什么不同；以前的异常检测方法，比如统计方法，以分类为基础的方法，以聚类为基础的方法，但是他们有两方面的缺点：⑴训练出来的异常检测器对正常数据拟合的特别好，但是对出现次数比较少的数据拟合的就不是很好，预测出来的异常很多都是假的，⑵ 另外大多数已经存在的方法因为计算复杂度太高受限于低纬度数据，无法对对高纬度数据进行有效检测。

孤立森林利用过去方法存在的问题以及抓住异常点的特有特征——出现次数少而且和正常数据很不一样，利用这两点，提出了孤立森林的方法。

孤立森林这种方法，基本的思想非常简单，就是不断的分割数据集，当整棵孤立树分割的高度超过限制高度的时候就停止分割。

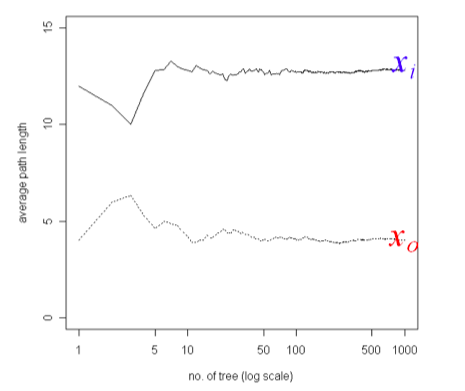
孤立树构造过程伪代码如图所示：



但是，用这种方法，如果你不对它进行限制的话，如果原始数据有数十万个数据点，那么分割将会是一件非常麻烦的事情，而且对于数据密集的区域，不断进行迭代分割意义也并不是很大。

所以，在伪代码中，对树的高度是有一定限制的。

那么利用原始数据，最终得到一棵树，如何能够判断数据点是正常的还是不正常的，其实，我们在图一中就可以看到，对于 需要很多次才能够孤立出来，但是对于仅仅需要几次就能够孤立出来了，如下图所示，二者平均路径长度是不一样的。

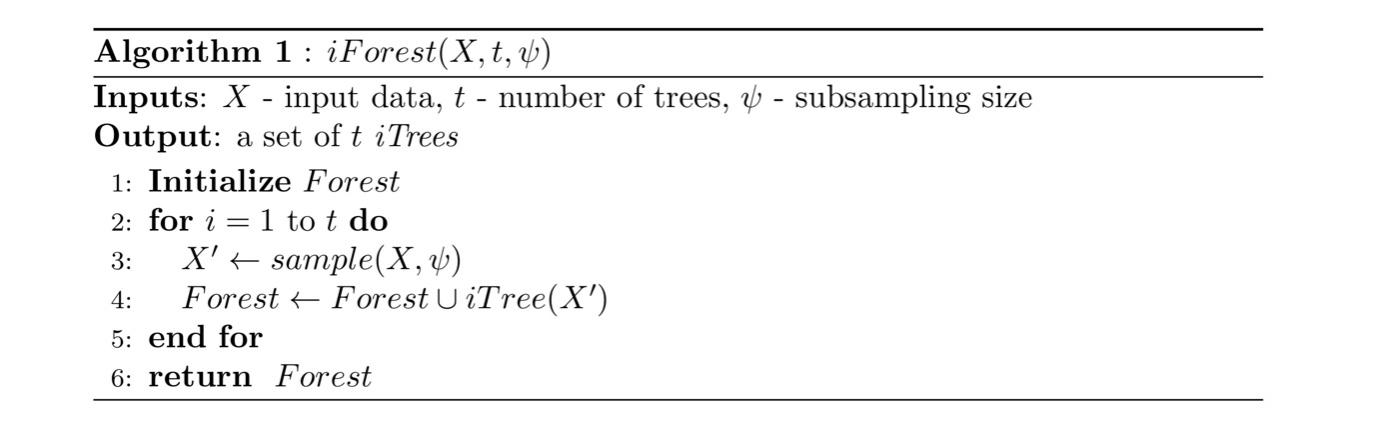


异常点往往很容易就能够孤立出来，但是正常点由于数据量比较大，需要好多次才能够孤立出来。

那么，是不是根据这种孤立的次数就能够判断一个点是不是异常点了？

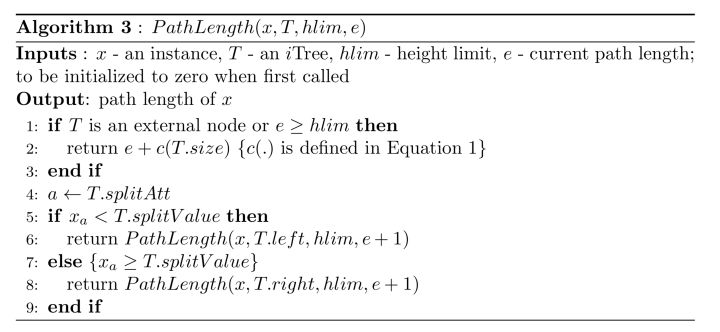
实际上，对于孤立森林来说，一个数据集不仅仅构造一棵孤立树，而是很多棵孤立树，要不然也不会叫做孤立森林了。

上面我们已经给出构造孤立树的过程，那么现在是如何构造孤立森林了，如图所示，我们能够看到构造孤立森林的伪代码：



那么这个切割出来又是如何进行异常点检测的？

虽然树和森林都构造出来了，但是结果是如何得到的，实际上孤立森林算是集成方法的一种，但是对于单个的孤立树来说，如何判断一个样本是不是异常值，我们需要根据该样本被划分的难易，也就是说，该样本在孤立树中的深度如何，查询某个样本的树深度，其伪代码如图所示：



1. 孤立森林应用场景

通过对孤立森林的思想以及算法原理的了解，我们知道，孤立森林对基于密度的样本可以很好的展现它优良的特性，但是在异常检测领域，并不是所有的样本都是以向量的形式呈现，在工业界，有很多场景下产生的数据都是以时间为基准，不断的产生数据，对于这种时序数据，孤立森林处理起来比较困难，也就是说，孤立森林的应用领域具有一定的局限性。

1. 孤立森林优劣势分析

孤立森林算法有着线性时间复杂度以及较低的内存需求，在大量非相关属性的高纬问题上表现很好，而且它的训练集不需要包含任何异常。

对于其时间复杂度，我们很容易理解，为什么孤立森林具有很低的时间复杂度，能够达到线性，原因就在于，孤立森林每次对样本集进行切分，都只是随机选取一个特征，然后选取该特征的一个值进行切分，这样把整个样本集切分完，实际上，其时间复杂度，也就相当于其切分的次数；往往为了不让切割之后的图形过于复杂，在使用孤立森林之前，会预先设置孤立森林的树的深度，这样以至于，其切分次数计算如下；

n——树的深度；

我们可以看到孤立森林的深度如果设置为n的话，切割完成需要大概次，也就是说，如果按照操作次数，其实这个时间复杂度是不会到达线性的，但是为什么原作者告诉我们，孤立森林的时间复杂度会成为线性的了，这是因为，在构造孤立森林的过程中，还用到了剪枝策略，也就是说并不是对每一个子样本都需要切割，因为并不是每个子样本领域都有多个样本点，对于密度比较低的子样本，如果其样本个数不超过一个，就不用再进行切割了，所以这个剪枝方法大大减少了迭代执行的次数。

孤立森林还有一个缺点就是无法对多属性进行有效的划分，一般来说，孤立森林表现最好的就是在二维平面上的划分。

1. 孤立森林和决策树，随机森林有什么区别

对于孤立森林来说，我们前期不需要对它设置参数，也不需要进行监督训练，只需要直接对样本进行一次孤立森林构造，就能够探测到异常点在哪里，这对其他机器学习方法来说是高效的，因为不需要考虑参数值的设置，以及训练样本的比例，还有就是测试集如何选择，需要不需要用到交叉验证等方法。还有就是对于监督学习来说，过拟合，欠拟合还有评估方法，都是需要考虑的重点，样本的采集等等。

但是对于孤立森林了，孤立森林对于很多人来说，以为其是多个树组成的算法框架，但是实际上，我们通过上面的样本算法可以看到，其实孤立森林本质就是由一个特殊的决策树构建而成，只不过这个决策树的决策准则和ID3，C4.5以及CART等常用的决策树不同，因为对于孤立森林来说，训练结束之后，将会形成一棵决策树，这颗决策树的根节点，并不是直接用来进行分类，对异常点进行分类，而是，需要根据这棵树的深度来对异常进行分类。

那么随机森林了，随机森林和孤立森林又有什么区别，随机森林是对某一个训练集，随机的抽取包含多个属性的子集，然后针对这些子集训练基学习器，孤立森林也是如此，从样本中抽出一个子子集，然后随机生成一棵孤立树，可以说二者是非常一样的。

总结

孤立森林具有线性时间复杂，可以用在含有海量数据的数据集上面。通常树的数量越多，算法越稳定。由于每棵树都是互相独立生成的，因此可以部署在大规模分布式系统上来加速运算。

孤立森林不适用于特别高维的数据。由于每次切数据空间都是随机选取一个维度，建完树后仍然有大量的维度信息没有被使用，导致算法可靠性降低。高维空间还可能存在大量噪音维度或无关维度，影响树的构建。对这类数据，使用子空间异常检测技术比较好。此外，切割平面默认是axis-parallel的，也可以随机生成各种角度的切割平面。

孤立森林对全局异常敏感，不擅长处理局部的相对稀疏点。

参考文献

[1]Liu F T, Ting K M, Zhou Z H. Isolation-Based Anomaly Detection[J]. Acm Transactions on Knowledge Discovery from Data, 2012, 6(1):1-39.

[2]Liu F T, Kai M T, Zhou Z H. Isolation Forest[C]Eighth IEEE International Conference on Data Mining. IEEE, 2009:413-422.

[3]The Identification of Outliers in Two-Way Contingency Tables Using 2 × 2 Subtables

[4] T. J. v W. Kotze and Douglas M. Hawkins. The Identification of Outliers in Two-Way Contingency Tables Using 2 × 2 Subtables[J]. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1984, 33(2):215-223.

[5]. Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.