


iForest（Isolation Forest）孤立森林异常检测 入门篇



YeZhu (/u/5d354b505d16)

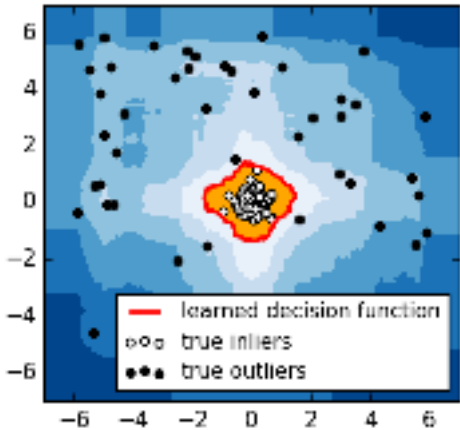
+ 关注

2017.01.29 09:11* 字数 1924 阅读 13445 评论 41 喜欢 32 赞赏 1

(/u/5d354b505d16)

iForest（Isolation Forest）孤立森林 是一个基于Ensemble的快速异常检测方法，具有线性时间复杂度和高精度度，是符合大数据处理要求的state-of-the-art算法（详见新版教材“Outlier Analysis (<https://link.jianshu.com?t=http%3A%2F%2Flink.springer.com%2Fbook%2F10.1007%2F978-3-319-47578-3>)”第5和第6章 PDF (<https://link.jianshu.com?t=http%3A%2F%2Fcharuaggarwal.net%2Foutlierbook.pdf>)）。其可以用于网络安全中的攻击检测，金融交易欺诈检测，疾病侦测，和噪声数据过滤等。本文将通俗解释实现方法和日常运用，即无需深厚的数学功底。

首先，我们先了解下该算法的动机。目前学术界对异常（anomaly detection (<https://link.jianshu.com?t=http%3A%2F%2Fcucis.ece.northwestern.edu%2Fprojects%2FDMS%2Fpublications%2FAnomalyDetection.pdf>））的定义有很多种，iForest 适用与连续数据（Continuous numerical data）的异常检测，将异常定义为“容易被孤立的离群点（more likely to be separated）”——可以理解为分布稀疏且离密度高的群体较远的点。用统计学来解释，在数据空间里面，分布稀疏的区域表示数据发生在此区域的概率很低，因而可以认为落在这些区域里的数据是异常的。一个例子如下（来源 (https://link.jianshu.com?t=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fauto_examples%2Fcovariance%2Fplot_outlier_detection.html%23sphx-glr-auto-examples-covariance-plot-outlier-detection-py)））：



黑色的点为异常点，白色点为正常的点（在一个簇中）。iForest检测到的异常边界为红色，它可以正确地检测到所有黑点异常点。

iForest属于Non-parametric和unsupervised的方法，即不用定义数学模型也不需要标记的训练。对于如何查找哪些点是否容易被孤立（isolated），iForest使用了一套非常高效的策略。假设我们用一个随机超平面来切割（split）数据空间（data space），切一次可以生成两个子空间（想象拿刀切蛋糕一分为二）。之后我们再继续用一个随机超平面来切割每个子空间，循环下去，直到每子空间里面只有一个数据点为止。直观上来讲，我们可以发现那些密度很高的簇是可以被切很多次才会停止切割，但是那些密度很低的点很容易很早的就停到一个子空间了。上图里面黑色的点就很容易被切几次就停到一个子空间，而白色点聚集的地方可以切很多次才停止。



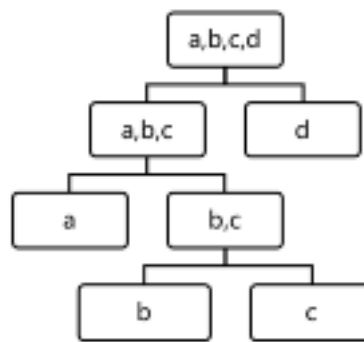
怎么来切这个数据空间是iForest的设计核心思想，本文仅介绍最基本的方法。由于切割是随机的，所以需要ensemble的方法来得到一个收敛值（蒙特卡洛方法），即反复从头开始切，然后平均每次切的结果。iForest由t个iTree (Isolation Tree) 孤立树组成，每个iTree是一个二叉树结构，其实现步骤如下：

1. 从训练数据中随机选择 Ψ 个点样本作为subsample，放入树的根节点。
2. 随机指定一个维度（attribute），在当前节点数据中随机产生一个切割点 p ——切割点产生于当前节点数据中指定维度的最大值和最小值之间。
3. 以此切割点生成了一个超平面，然后将当前节点数据空间划分为2个子空间：把指定维度里小于 p 的数据放在当前节点的左孩子，把大于等于 p 的数据放在当前节点的右孩子。
4. 在孩子节点中递归步骤2和3，不断构造新的孩子节点，直到孩子节点中只有一个数据（无法再继续切割）或孩子节点已到达限定高度。

获得t个iTree之后，iForest训练就结束，然后我们可以用生成的iForest来评估测试数据了。对于一个训练数据 x ，我们令其遍历每一棵iTree，然后计算 x 最终落在每个树第几层（ x 在树的高度）。然后我们可以得出 x 在每棵树的高度平均值，即 the average path length over t iTrees。*值得注意的是，如果 x 落在一个节点中含多个训练数据，可以使用一个公式来修正 x 的高度计算，详细公式推导见原论文 (<https://link.jianshu.com?t=http%3A%2F%2Fcs.nju.edu.cn%2Fzhouzh%2Fzhouzh.files%2Fpublication%2Ficdm08b.pdf>)。

获得每个测试数据的average path length后，我们可以设置一个阈值（边界值），average path length 低于此阈值的测试数据即为异常。也就是说“iForest identifies anomalies as instances having the shortest average path lengths in a dataset”(异常在这些树中只有很短的平均高度)。*值得注意的是，论文中对树的高度做了归一化，并得出一个0到1的数值，即越短的高度越接近1（异常的可能性越高）。

4个测试样本遍历一棵iTree的例子如下：



b和c的高度为3，a的高度是2，d的高度是1。

可以看到d最有可能是异常，因为其最早就被孤立（isolated）了。

生成一棵iTree的详细算法（来源 <https://link.jianshu.com?t=http%3A%2F%2Fdl.acm.org%2Fcitation.cfm%3Fid%3D2939779>）：

t=<http%3A%2F%2Fdl.acm.org%2Fcitation.cfm%3Fid%3D2939779>) :



Algorithm 1 *iTree*(X, e, h)

Input: X - input data; e - current height; h - height limit.
Output: an *iTree*.

```

1: if  $e \geq h$  OR  $|X| \leq 1$  then
2:   return exNode{ $Size \leftarrow |X|$ };
3: else
4:   Randomly select an attribute  $q$ ;
5:   Randomly select a split point  $p$  between  $\min$  and

```

X 为独立抽取的训练样本。参数 e 的初始值为0。 h 是树可以生成的最大高度。

iForest算法默认参数设置如下：

subsample size: 256

Tree height : 8

Number of trees: 100

通俗解释就是——建100棵iTree，每棵iTree最高8层，且每棵iTree都是独立随机选择256个数据样本建成。

个人见解：

1. iForest具有线性时间复杂度。因为是ensemble的方法，所以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。

2. iForest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（irrelevant attributes），影响树的构建。对这类数据，建议使用子空间异常检测（Subspace Anomaly Detection）技术。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，详见“On Detecting Clustered Anomalies Using SCiForest (https://link.jianshu.com?t=http%3A%2F%2Flink.springer.com%2Fchapter%2F10.1007%2F978-3-642-15883-4_18)”。

3. iForest仅对Global Anomaly 敏感，即全局稀疏点敏感，不擅长处理局部的相对稀疏点（Local Anomaly）。目前已有改进方法发表于PAKDD，详见“Improving iForest with Relative Mass (https://link.jianshu.com?t=http%3A%2F%2Flink.springer.com%2Fchapter%2F10.1007%2F978-3-319-06605-9_42)”。

4. iForest推动了重心估计（Mass Estimation）理论发展，目前在分类聚类和异常检测中都取得显著效果，发表于各大顶级数据挖掘会议和期刊（如SIGKDD，ICDM，ECML）。

参考文献：

iForest 是刘飞 (<https://link.jianshu.com?t=https%3A%2F%2Ffeitonyliu.wordpress.com%2Fabout%2F>)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%2Federation.edu.au%2Ffaculties-and-schools%2Ffaculty-of-science-and-technology%2Fstaff-profiles%2Finformation-technology%2Fkai-ming-ting>)(Kai-Ming Ting)教授和周志华 (<https://link.jianshu.com?t=http%3A%2F%2Fcs.nju.edu.cn%2Fzhouzh%2F>)(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，扩充版本发表与TKDD。



Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008.

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation-based anomaly detection." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6.1 (2012): 3.

论文下载：

<http://cs.nju.edu.cn/zhoush/zhoush.files/publication/icdm08b.pdf>
(<https://link.jianshu.com?t=http%3A%2F%2Fcs.nju.edu.cn%2Fzhoush%2Fzhoush.files%2Fpublication%2Ficdm08b.pdf>)

<http://cs.nju.edu.cn/zhoush/zhoush.files/publication/tkdd11.pdf> (<https://link.jianshu.com?t=http%3A%2F%2Fcs.nju.edu.cn%2Fzhoush%2Fzhoush.files%2Fpublication%2Ftkdd11.pdf>)

源码下载：

R语言 <https://sourceforge.net/projects/iforest/> (<https://link.jianshu.com?t=https%3A%2F%2Fsourceforge.net%2Fprojects%2Fiforest%2F>)

Python语言 <http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html>
([https://link.jianshu.com?t=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fgenerated%2Fsklearn.ensemble.IsolationForest.h](https://link.jianshu.com?t=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fgenerated%2Fsklearn.ensemble.IsolationForest.html)tml)

Java语言 <http://weka.sourceforge.net/packageMetaData/isolationForest/index.html>
(<https://link.jianshu.com?t=http%3A%2F%2Fweka.sourceforge.net%2FpackageMetaData%2FisolationForest%2FIndex.html>)

Matlab语言 <https://github.com/zhuye88/iForest> (<https://link.jianshu.com?t=https%3A%2F%2Fgithub.com%2Fzhuye88%2FiForest>)

全文完，转载必须注明出处：© Ye Zhu (<https://link.jianshu.com?t=http%3A%2F%2Fwww.yezhu.com.au%2F>) 2017


小礼物走一走，来简书关注我

赞赏支持



(/u/bde97fa2387c)

算法研究 (/nb/9372510) 举报文章 © 著作权归作者所有



YeZhu (/u/5d354b505d16) ♂

写了 4755 字，被 65 人关注，获得了 38 个喜欢

(/u/5d354b505d16)

[+ 关注](#)

2012年毕业于英国伦敦帝国理工学院，2017年在澳大利亚莫纳什大学获得机器学习和数据挖掘博士学位。 ...

喜欢


32

[更多分享](#)



下载简书 App ▶

随时随地发现和创作内容




(/apps/download?utm_source=nbc)


被以下专题收入，发现更多相似内容




程序员 (/c/NEt52a?utm_source=desktop&utm_medium=notes-included-collection)




架构算法设计模... (/c/c568ddab391a?utm_source=desktop&utm_medium=notes-included-collection)




大数据 (/c/b641f7c33fd2?utm_source=desktop&utm_medium=notes-included-collection)



语言 (/c/4636009e33b8?utm_source=desktop&utm_medium=notes-included-collection)



数据乐园 (/c/a3017f6e996e?utm_source=desktop&utm_medium=notes-included-collection)




机器学习与数据挖掘 (/c/9ca077f0fae8?utm_source=desktop&utm_medium=notes-included-collection)

(/p/cdbed82a34bc?



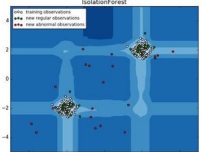
utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
[3/4]我所经历的大数据平台发展史（三）：互联网时代·上篇 (/p/cdbed82a...

//我所经历的大数据平台发展史（三）：互联网时代·上篇http://www.infoq.com/cn/articles/the-development-history-of-big-data-platform-paet02 编者按：本文是松子（李博源）的大数据平台发展史...




葡萄喃喃呓语 (/u/2c67926c48ce?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/1b020e2605e2?



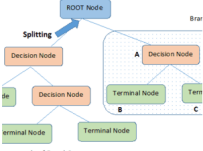
utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
0x14 异常挖掘，Isolation Forest (/p/1b020e2605e2?utm_campaign=mal...

摘要：iForest用于挖掘异常数据，如网络安全中的攻击检测和流量异常分析，金融机构则用于挖掘出欺诈行为。算法对内存要求很低，且处理速度很快，其时间复杂度也是线性的。可以很好的处理高维数据和大量...



云戒 (/u/809656718e88?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/ff9b7b031fed?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
关于基于树的建模的完整教程（从R&Python） (/p/ff9b7b031fed?utm_ca...

翻译自analyticsvidhya 基于树的学习算法被认为是最好的和最常用的监督学习(supervised learning)方法之一。基于树的方法赋予预测模型高精度,稳定性和易于解释的能力。与线性模型不同,它们非常好地映射...



珞珈村下山 (/u/cc62f0e70e83?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

机器学习(Machine Learning)&深度学习(Deep Learning)资料(Chapter 1) (...)

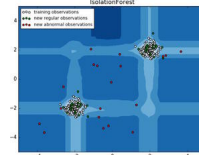
机器学习(Machine Learning)&深度学习(Deep Learning)资料(Chapter 1) 注:机器学习资料篇目一共500条,篇目二开始更新 希望转载的朋友,你可以不用联系我,但是一定要保留原文链接,因为这个项目还在继续也...



Albert陈凯 (/u/185a3c553fc6?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/ab7713dc884f?)



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

Isolation Forest (/p/ab7713dc884f?utm_campaign=maleskine&utm_co...

摘要: iForest用于挖掘异常数据,如网络安全中的攻击检测和流量异常分析,金融机构则用于挖掘出欺诈行为。算法对内存要求很低,且处理速度很快,其时间复杂度也是线性的。可以很好的处理高维数据和大量...



xulao3 (/u/68317f7e5163?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/eff774d2470d?)



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

原来我们都期待遇到一个人 (/p/eff774d2470d?utm_campaign=maleskine...

01 有多久你没有期待了?期待一件好事的发生。不知道从什么时候开始,我们把自己隐藏在钢筋水泥的城市里,淹没在人来人往的人潮里,按着那既定的脚步,走,哪怕不走,也会有固定的力量把你推向某个既定...



瑞和她的浅岛繁花 (/u/aafc4d608cad?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

2017-07-14 (/p/5b31d114bc77?utm_campaign=maleskine&utm_conten...

如果我是一本小说,书名是谁心所欲!为什么叫这个名字呢!因为我的终极目标我的渴望我的追求就是可以跟着自己的心去生活去选择!我可以无拘无束的生活!我可以不以外的观点价值观所束缚!我可以自由...



真与真 (/u/b6f477cb4553?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

自律的大顺20170711 (/p/3d9a2a16fe19?utm_campaign=maleskine...

儿今天跟爸爸去了公司,做了一天的作业。回来爸爸说,晚上不要再做作业了,玩会。儿略带愧疚地对我说,今天英语那本暑假作业没做完,我说,没关系。那本作业好厚呀!不着急,明天再做。爸爸让儿去玩...



大爱无疆杨青 (/u/0e40f1ba346b?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

夜游园 (/p/05d97147b207?utm_campaign=maleskine&utm_content=no...

忙碌了一个下午,夜晚趁着还未困倦,在校园内漫步着。吹着徐徐清风,沿着河流出发,望着眼前那些模糊不清的柳树,在清风下摆动着。在文人的心中,此刻它们定式风情万种,撩起心底那最后的宁静,轻轻的...



七月浅 (/u/2d8244b4b285?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)



