异常检测算法--Isolation Forest

南大周志华老师在2010年提出一个异常检测算法Isolation Forest，在工业界很实用，算法效果好，时间效率高，能有效处理高维数据和海量数据，这里对这个算法进行简要总结。

iTree

提到森林，自然少不了树，毕竟森林都是由树构成的，看Isolation Forest（简称iForest）前，我们先来看看Isolation Tree（简称iTree）是怎么构成的，iTree是一种随机二叉树，每个节点要么有两个女儿，要么就是叶子节点，一个孩子都没有。给定一堆数据集D，这里D的所有属性都是连续型的变量，iTree的构成过程如下：

- 随机选择一个属性Attr；
- 随机选择该属性的一个值Value；
- 根据Attr对每条记录进行分类，把Attr小于Value的记录放在左女儿，把大于等于Value的记录放在右孩子；
- 然后递归的构造左女儿和右女儿，直到满足以下条件：
- 传入的数据集只有一条记录或者多条一样的记录；
- 树的高度达到了限定高度；

Algorithm 2 : $iTree(X, e, l)$

Inputs: X - input data, e - current tree height, l - height limit

Output: an iTree

if $e \geq l$ or $|X| \leq 1$ **then**
  return $exNode\{Size \leftarrow |X|\}$
else
  let Q be a list of attributes in X
  randomly select an attribute $q \in Q$
  randomly select a split point p from max and min values of attribute q in X
  $X_l \leftarrow filter(X, q < p)$
  $X_r \leftarrow filter(X, q \geq p)$
  return $inNode\{Left \leftarrow iTree(X_l, e+1, l),$
    $Right \leftarrow iTree(X_r, e+1, l),$
    $SplitAtt \leftarrow q,$
    $SplitValue \leftarrow p\}$
end if

iTree构建好了后，就可以对数据进行预测啦，预测的过程就是把测试记录在iTree上走一下，看测试记录落在哪个叶子节点。iTree能有效检测异常的假设是：异常点一般都是非常稀有的，在iTree中会很快被划分到叶子节点，因此可以用叶子节点到根节点的路径h(x)长度来判断一条记录x是否是异常点；对于一个包含n条记录的数据集，其构造的树的高度最小值为log(n)，最大值为n-1，论文提到说用log(n)和n-1归一化不能保证有界和不方便比较，用一个稍微复杂一点的归一化公式：

$$s(x,n) = 2^{\left(-\frac{h(x)}{c(n)}\right)}$$

，

$$c(n) = 2H(n-1) - (2(n-1)/n), \text{其中} H(k) = ln(k) + \xi,\ \xi\text{为欧拉常数}$$

$s(x,n)$就是记录x在由n个样本的训练数据构成的iTree的异常指数，$s(x,n)$取值范围为[0,1]，越接近1表示是异常点的可能性高，越接近0表示是正常点的可能性比较高，如果大部分的训练样本的s(x,n)都接近于0.5，说明整个数据集都没有明显的异常值。

昵称：CodeMeals
园龄：5年5个月
粉丝：268
关注：19
+加关注

< 2018年1月 >

日	一	二	三	四	五	六
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3
4	5	6	7	8	9	10

常用链接

我的随笔
我的评论
我的参与
最新评论
我的标签

我的标签

数据挖掘(20)
python(13)
算法(10)
java(9)

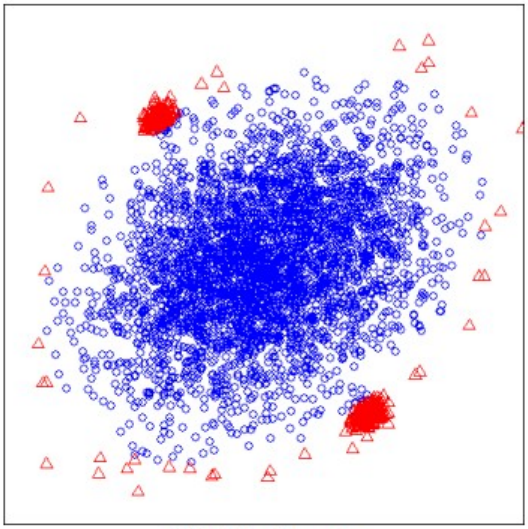
LeetCode(6)
实践中的bug(4)
数据库(3)
c++(3)
android(2)
深度学习(2)
更多

随笔档案
2017年6月 (1)
2015年11月 (1)
2015年9月 (1)
2014年11月 (1)
2014年9月 (1)
2014年8月 (4)
2014年7月 (2)
2014年6月 (5)
2014年5月 (8)
2014年4月 (1)
2014年3月 (1)
2014年1月 (1)
2013年12月 (1)
2013年11月 (2)
2013年8月 (5)
2013年7月 (6)
2013年6月 (7)
2013年5月 (5)
2013年4月 (2)
2013年3月 (1)
2013年2月 (2)
2013年1月 (3)
2012年11月 (1)
2012年10月 (1)

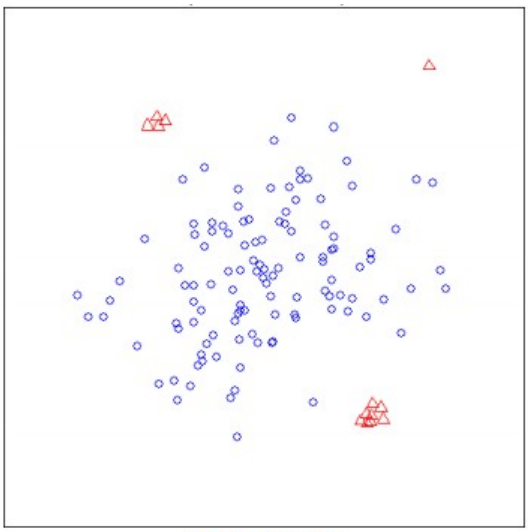
随机选属性，随机选属性值，一棵树这么随便搞肯定是不靠谱，但是把多棵树结合起来就变强大了；

iForest

iTree搞明白了，我们现在来看看iForest是怎么构造的，给定一个包含n条记录的数据集D，如何构造一个iForest。iForest和Random Forest的方法有些类似，都是随机采样——部分数据集去构造每一棵树，保证不同树之间的差异性，不过iForest与RF不同，采样的数据量Psi不需要等于n，可以远远小于n，论文中提到采样大小超过256效果就提升不大了，明确越大还会造成计算时间的上的浪费，为什么不像其他算法一样，数据越多效果越好呢，可以看看下面这两个个图，



(a) Original sample
(4096 instances)



(b) Sub-sample
(128 instances)

左边是元素数据，右边是采样了数据，蓝色是正常样本，红色是异常样本。可以看到，在采样之前，正常样本和异常样本出现重叠，因此很难分开，但我们采样之和，异常样本和正常样本可以明显的分开。

除了限制采样大小以外，还要给每棵iTree设置最大高度$l = ceiling(log_2^\Psi)$，这是因为异常数据记录都比较少，其路径长度也比较低，而我们也只需要把正常记录和异常记录区分开来，因此只需要关心低于平均高度的部分就好，这样算法效率更高，不过这样调整了后，后面可以看到计算$h(x)$需要一点点改进，先看iForest的伪代码：

2012年9月 (1)

2012年7月 (1)

友情链接

老高

新浪微博(@也爱数据挖掘)

积分与排名

积分 - 106847

排名 - 2747

最新评论

1. Re:数据挖掘系列（6）决策树分类算法

博主，有没有具体的实现呢，想观摩拜读下

--开心就好硕

2. Re:TensorFlow上实践基于自编码的One Class Learning

博主你好！请问这个训练结束之后是不是也要保存训练好的模型？

--我是南山码农

3. Re:数据挖掘之KNN分类

@PacosonSWJTU谢谢...

--CodeMeals

4. Re:异常检测算法--Isolation Forest

你好，看了你的文章，自己思考后有些地方不太明白，麻烦博主指点，假设我有一堆数据想判别是否有异常值，但是iforest需要训练集和测试集，我不知道这要如何划分，不知道如果我利用sklearn里面的方法，.....

--xt_judy

5. Re:模块度与Louvain社区发现算法

hi,我想问下最后一句话，"用节点id和和社区id构成的边组成新图，再用联通图来调整节点的社区"，我记得联通图是更新顶点的属性值为联通图内最小的id，这样子，会有个问题，假设社区id原先是图里面最小顶.....

Algorithm 1 : $iForest(X, t, \psi)$

Inputs: X - input data, t - number of trees, ψ - sub-sampling size

Output: a set of t *iTrees*

Initialize $Forest$
set height limit $l = ceiling(\log_2 \psi)$
for $i = 1$ to t **do**
  $X' \leftarrow sample(X, \psi)$
  $Forest \leftarrow Forest \cup iTree(X', 0, l)$
end for
return $Forest$

IForest构造好后，对测试进行预测时，需要进行综合每棵树的结果，于是

$$s(x,n) = 2^{(-\frac{E(h(x))}{c(n)})}$$

$E(h(x))$表示记录x在每棵树的高度均值，另外h(x)计算需要改进，在生成叶节点时，算法记录了叶节点包含的记录数量，这时候要用这个数量$Size$估计一下平均高度，h(x)的计算方法如下：

Algorithm 3 : $PathLength(x, T, e)$

Inputs : x - an instance, T - an iTree, e - current path length; to be initialized to zero when first called

Output: path length of x

if T is an external node **then**
  return $e + c(T.size)$ {$c(.)$ is defined in Equation 1}
end if
$a \leftarrow T.splitAtt$
if $x_a < T.splitValue$ **then**
  return $PathLength(x, T.left, e + 1)$
else $\{x_a \geq T.splitValue\}$
  return $PathLength(x, T.right, e + 1)$
end if

处理高维数据

在处理高维数据时，可以对算法进行改进，采样之后并不是把所有的属性都用上，而是用峰度系数Kurtosis挑选一些有价值的属性，再进行iTree的构造，这跟随机森林就更像了，随机选记录，再随机选属性。

只使用正常样本

这个算法本质上是一个无监督学习，不需要数据的类标，有时候异常数据太少了，少到我们只舍得拿这几个异常样本进行测试，不能进行训练，论文提到只用正常样本构建IForest也是可行的，效果有降低，但也还不错，并可以通过适当调整采样大小来提高效果。

全文完，转载请注明出处：http://www.cnblogs.com/fengfenggirl/p/iForest.html

标签： 数据挖掘

CodeMeals
关注 - 19
粉丝 - 268
+加关注

« 上一篇：分享我的书签
» 下一篇：模块度与Louvain社区发现算法

posted @ 2015-09-05 14:19 CodeMeals 阅读(25524) 评论(14) 编辑 收藏

评论列表

#1楼 2016-01-24 14:36 learn_coding

--明日菜心

阅读排行榜

1. 异常检测算法--Isolation Forest(25518)
2. 数据挖掘系列（6）决策树分类算法(23689)
3. 数据挖掘系列（10）——卷积神经网络算法的一个实现(21190)
4. 如何找出知乎的所有神回复(14500)
5. 数据挖掘系列（1）关联规则挖掘基本概念与Aprior算法(11842)

评论排行榜

1. 如何找出知乎的所有神回复(28)
2. 一个简单的多线程爬虫(26)
3. 数据挖掘系列（2）--关联规则FpGrowth算法(22)
4. 像Hacker News一样排序博客园首页文章(18)
5. 开发一个简单实用的android紧急求助软件(17)

推荐排行榜

1. 像Hacker News一样排序博客园首页文章(18)
2. PageRank实践-博客园用户PageRank排名(17)
3. 如何找出知乎的所有神回复(13)
4. 数据挖掘系列（1）关联规则挖掘基本概念与Aprior算法(11)
5. 在茫茫人海中发现相似的你——局部敏感哈希（LSH）(8)

您好，看了您的这篇文章感觉受益匪浅。我想问一下您，iforest算法您认为是否可以写成分布式的算法呢？

支持(0) 反对(0)

#2楼[楼主] 2016-01-25 19:32 CodeMeals

@ learn_coding
可以的，我们组有同学实现了MR的版本

支持(0) 反对(0)

#3楼 2016-01-25 19:35 learn_coding

@ CodeMeals
那您方便把代码发给我学习一下么？我做个学习方面的参考。非常感谢呢。
我的邮箱是1123941131@qq.com

支持(0) 反对(0)

#4楼 2016-01-28 14:27 大闹天空的程序猴

您好，我想问下你们用来做iforest的异常样本在哪里得到的？

支持(1) 反对(0)

#5楼 2016-03-07 12:42 yongmou-

"南大周志华老师在2010年提出一个异常检测算法Isolation Forest"，明明是周老师的一个学生提出的算法，这能一样么？

支持(1) 反对(0)

#6楼 2016-03-17 09:04 linpeikun16

@ 大闹天空的程序猴
同求测试数据集

支持(0) 反对(0)

#7楼 2016-10-01 10:01 摇光在望

写的真的太好了，请问可以发一份源代码学习一下么？我做个学习方面的参考。非常谢谢你！我的邮箱是390240733@qq.com

支持(0) 反对(0)

#8楼 2016-10-29 09:57 fionaplanet

楼主，看了你的博客非常感兴趣，想问下isolation Forest的代码能够发我学习下，现在正在做非法交易的识别。

支持(0) 反对(0)

#9楼 2016-12-29 16:59 临冬

时隔一年看到博文，受益匪浅，博主，目前我想做一下iforest异常检测方面的内容，对于算法改进，博主觉得除了在维度随机选择之外还有什么可改进的地方吗？ 同时我也准备做成分布式的spark版本，不博主觉得可行吗？

支持(1) 反对(0)

#10楼 2017-01-29 06:33 YaleZhu

这个iForest算法是刘博士(Fei Tony Liu)在莫纳什大学就读期间由陈开明(Kai-Ming Ting)教授和周志华(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。

R语言实现的版本下载 sourceforge.net/projects/iforest/
Python语言实现的版本 scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

支持(0) 反对(0)

#11楼 2017-01-29 06:33 YaleZhu

@ 摇光在望
R语言实现的版本下载 sourceforge.net/projects/iforest/
Python语言实现的版本 scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

支持(0) 反对(0)

#12楼 2017-05-03 22:14 一个看上算法AI的小伙

好厉害好厉害

支持(0) 反对(0)

#13楼 2017-05-03 22:15 一个看上算法AI的小伙

论文在哪里呢，楼主方便给链接吧 想拜读一下

支持(0) 反对(0)

#14楼 2017-07-17 22:46 xt_judy

你好，看了你的文章，自己思考后有些地方不太明白，麻烦博主指点，假设我有一堆数据想判别是否有异常值，但是iforest需要训练集和测试集，我不知道这要如何划分，不知道如果我利用sklearn里面的方法，直接把所有数据带入，当训练集，这样输出异常值，是否准确，还有就是sklearn方法里面有个参数contamination，用来设置异常数据存在的比率，那这样做会使结果不准确吗？麻烦博主指点，感激不尽

支持(0) 反对(0)

Visitors