ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.13, No.5, February 2017
E-mail: jslt@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于随机森林的HTTP异常检测

唐宇迪

（同济大学，上海 200092）

摘要：在互联网日益强大的今天，网络安全问题已经尤为重要，如何能够精准找到网络中的攻击行为具有重要的价值。基于该目标，该文提出了基于HTTP流量数据的异常检测模型，以随机森林为核心算法，围绕该算法提出了一种HTTP流量数据生成策略以及检测方法。

关键词：随机森林；HTTP异常检测；数据生成

中图分类号：TP393 文献标识码：A 文章编号：1009-3044(2017)05-0031-03

DOI:10.14004/j.cnki.ckt.2017.0575

1 HTTP异常数据生成策略

对于HTTP流量数据的异常行为检测，一个难点就在于如何定义正负样本也就是正常的HTTP行为和异常的HTTP行为。现阶段，普遍的做法是通过网络异常检测软件来对每个有风险的IP点进行检测，这种做法一方面并不能得出准确的结果另一方面也很难发现新的异常IP点。基于这点，本文提出了一种是用数据生成策略并基于聚类结果的随机森林检测模型[1]。

1.1 问题提出

对于HTTP流量数据可以通过聚类算法得出一些离群点，对于这些离群点使用集成的方式可以得出不同类型的IP点，例如将离群点当做具有潜在异常行为的IP点，将非离群点当做正常的IP点，将部分聚类算法认为是离群点的当做疑似点。通过聚类算法虽然可以出来部分具有异常行为的IP点，但是从整体的量上来说，离群点只占了整个HTTP流量数据中很少的一部分。原始数据中可能还存在着大量具有异常行为的IP点[2]。

对于原始的HTTP日志数据来说，进行异常行为检测的难点在于问题的本身是一个无监督的问题，没有给定的标签来指定什么样的IP点是正常的，什么是异常的。这使得很难对提取的特征数据进行分类的算法，基于这点不得不选择无监督的聚类算法。通过聚类算法得出了一些离群点，然后通过集成的方法将离群点分成3类，即异常点，疑似点，正常点。有了这些的基础，对与异常行为检测这个正负样本分类问题，将原本无监督的样本集分成了3个类别，即有较大可能是负样本的异常点集合，有可能是负样本的疑似点集合，正样本则对应着正常点集合。基于这种划分规则，便可将原本的无监督的问题转换成一个半监督问题[3]。

对于分类算法来说，一个重点就在于正负样本的选择，通过分析得知正样本的数量很丰富，但是相对来说负样本的数量却远远不够，由于聚类算法得出的异常点和疑似点数量都比较少，所以要进行分类算法首先需要对数据样本就行增强。如何选择一种合适的数据增强策略改善这种不均衡的数据分布成为分类准确性的关键所在。

分类算法第二个重点在于如何选择合适的分类器，目前已经有很多种经典分类算法，由于数据本身的无监督性，以及由聚类分析的不同特征的重要程度具有明显差别，这里需要综合考虑这些因素选择最合适的分类器[2]。

1.2 数据均衡问题解决

数据样本是否均衡对最终的分类结果可能会产生很大的影响，对于服务器端IP点来说得到的异常点一共有20个，疑似点有724个，正常点有279025个。这是一个极不均衡的样本分布。假设把异常点和疑似点都算作负样本，那也仅仅只有760个样本点，这远远不够的。为了解决数据均衡问题，基于异常点和疑似点我对负样本进行了随机的生成，策略如下：

对于异常点，通过聚类集成方法的分析已经证明了它们的异常行为的可靠性，所以要充分利用这些异常IP点的特征数据进行更多负样本的生成。由于每一个异常IP点的特征数据都具有7个特征，它们具有的异常行为，可能是这7个特征中一个或者某几个发生了明显的数值上的变换，基于这点，在生成新的异常特征数据时，对每一个异常点的7个特征都需要进行不同的变换，由异常点的特征数据发现，不同的异常IP点之间在7个特征上数据变化幅度较大，尤其表现在和连接数量相关的特征上，而与URI和COOKIE相关的特征的变化虽然趋于平缓，但整体仍具有浮动现象。基于这点以及异常行为潜在的多变性。我选择对特征数据进行随机的变换，将每一个异常IP点的每一个特征的取值随机映射到原始值的0.8-1.2倍之间，选择这个区间是因为，为了保证生成的数据尽可能地具有异常行为所以只选择了较小的变换范围，目的是保留住这些异常行为的数据特征，对于不同的映射区间下节会有详细的分析。并且保证7个特征都是随机进行特征数据的生成在这个区间上。例如异常点106.39.178.1的原始数据如下：

106.39.178.1	11092781	139.33521	34.952099	1.080522176	11084709	7919	466441

进行随机生成后的部分数据如下：

收稿日期：2017-01-08

作者简介：唐宇迪（1991—），男，广州人，硕士研究生，硕士，研究方向为信息安全，异常检测。

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

```
Generate from ip 106.39.178.1 sample0 [[  1.32059589e+07   1.64259827e+02   3.98584013e+01   1.11131375e+00
   1.30124790e+07   7.41376815e+03   3.79022633e+05]]
Generate from ip 106.39.178.1 sample1 [[  1.29644316e+07   1.28555036e+02   4.10421446e+01   1.11579979e+00
   9.72892896e+06   8.61678844e+03   4.24985400e+05]]
Generate from ip 106.39.178.1 sample2 [[  1.29868048e+07   1.46990140e+02   2.92473517e+01   1.06330348e+00
   1.17001952e+07   7.63497178e+03   4.22859877e+05]]
Generate from ip 106.39.178.1 sample3 [[  1.13187281e+07   1.61336381e+02   3.59924462e+01   1.22979785e+00
   1.20529224e+07   7.25667640e+03   5.19449695e+05]]
Generate from ip 106.39.178.1 sample4 [[  1.32227314e+07   1.36756588e+02   3.00354705e+01   1.25881562e+00
   8.92221327e+06   6.33677226e+03   4.08028414e+05]]
Generate from ip 106.39.178.1 sample5 [[  1.03207554e+07   1.17610690e+02   3.34912626e+01   9.68446629e-01
   1.10543754e+07   6.77457878e+03   4.72804639e+05]]
Generate from ip 106.39.178.1 sample6 [[  1.14326040e+07   1.17570024e+02   4.09615864e+01   1.03858118e+00
   1.14030646e+07   8.68378383e+03   3.76985413e+05]]
Generate from ip 106.39.178.1 sample7 [[  9.85704285e+06   1.12354885e+02   3.47546203e+01   1.19346468e+00
   9.84509299e+06   7.25803664e+03   5.30439515e+05]]
Generate from ip 106.39.178.1 sample8 [[  9.34705284e+06   1.25612244e+02   3.76728563e+01   1.04379243e+00
   9.43863507e+06   7.56901634e+03   4.67923308e+05]]
Generate from ip 106.39.178.1 sample9 [[  1.31980997e+07   1.52565695e+02   3.45477315e+01   1.16799912e+00
   1.24093262e+07   9.49180280e+03   4.69534596e+05]]
```

图1 负例生成数据样本

由于这些异常点有着极大的可能伴随着异常行为,而可利用的异常点的数量又非常少,所以这里我选择对每一个异常点都按照这样的规则随机生成了100个负样本。

对于疑似点,虽然这些点伴随着的异常行为的可能性没有异常点那么高,但是它们都是由聚类算法得出的离群点组成的。由离群点的特性可以得知,这些疑似点相比于正常点仍在某些特征上具有潜在的异常行为,并且疑似点的数量相比于异常点要多得多,这一方面可以很大程度丰富负样本的多样性不至于像异常点生成的负样本的数据特征行为都很相似另一方面可以生成更多的负样本。但是由于通过在聚类的分析得出的结果可知,这些疑似点存在异常行为的可能性要低于异常点。所以综合考虑这些因素,在这里我对724个疑似点中的每一个样本点按照同异常点生成的相同策略都随机生成10个负样本。

对于正常点,由于正样本数量已经足够,不需要对正样本进行生成,通过上述生成策略已经生成了10840个负样本,为了使正负样本更均衡,对正样本进行了随机选取,取10840个正样本作为分类算法的输入。

通过这样的生成策略,使得正负样本的个数更均衡,而且保证了数据的量,变换后正负样本分别有10840个特征数据。

2 基于HTTP异常检测的随机森林模型

对于分类算法来说有很多的分类器可供选择,在这里我选择随机森林模型的原因在于随机森林是一个用随机方式建立的,包含多个决策树的分类器。由于HTTP数据本身的无监督性以及在进行特征选择时无法对特征进行准确的评估只能从聚类分析其对异常行为影响的重要性。这也存在着潜在的问题就是有的特征可能对其异常行为产生负面的影响即不利于分类算法,但是由于网络异常行为的多样性和数据本身的无监督性,很难去准确分辨哪些特征的价值更高哪些可能具有负面影响。由于使用了数据增强策略,很多特征数据可能表现出较大的相似性,尤其是由异常点生成的数据样本。基于以上存在的问题选择随机森林模型的原因如下[4]:

1)随机森林模型在构造时在每个节点上,随机选取所有特征的一个子集,用来计算最佳分割方式。基于这点可以更全面的利用特征数据,使得即便某个特征可能存在负面的影响也不至于对分类结果产生较大的负面影响。

2)训练每棵树时,从全部训练样本(样本数为N)中选取一个可能有重复的大小同样为N的数据集进行训练(即bootstrap取样)。通过这样选择样本的方式可以有效避免生成数据具有较大相似性的问题。

首先对生成的特征数据同样进行归一化处理,为了验证随机森林模型的可靠性,我对几种经典的分类器如支持向量机,K近邻,决策树,Adaboosting,随机森林5种分类算法在生成的数据集上选择了同样的训练集和测试集,分别进行了默认参数的测试,即默认的参数都是基于样本数量的大小给予的没有进行任何的优化调节。从图中可以看出来随机森林的模型的效果要优于其他分类器的结果。这里的准确率的定义为:在生成的数据上进行的5倍交叉验证的准确率。

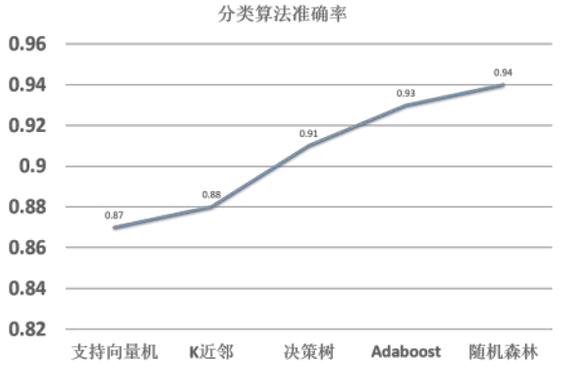

图2 分类算法对比

对于随机森林模型来说,对最终结果影响最大的就是建立树的个数,当把树的个数逐渐增多的时候自验证的准确率也会发生小范围提升,如上图所示当树的个数为10个时准确率为0.94,当树的个数为100个的时候自验证的准确率能平均得到0.98,再增加树的个数,准确率基本保持不变。

另一个重要的影响因素就是输入的特征数据,由于在进行负样本生成的时候选择了随机生成的区间值,下面综合分析一下该方案的优缺点。进行数据生成的原因在于对于本是无监督的HTTP日志数据很难使用分类算法进行快速的异常行为检测,但是根据聚类算法论证可以找到极小一部分的负样本,对于这部分负样本在生成更多的数据的时候可能有不同的策略,选择不同的变化幅度区间。对于特征数据来说,不同的特征变换的范围有着很大的差别,比如连接数量可能出现上千倍的变换即便在同是异常的IP点之间,而对于URI和COOKIE参数来说变换范围相对就小得多,这也符合实际的现象,考虑到这点,在下面的分析中,始终保持和URI,COOKIE相关的特征的变换范围区间在0.8到1.2之间不变,而只改变和连接数量相关的特征。这样做的意义在于如果选择的生成区间在比较小的范围内就会使得大量数据具有相似性,这样做虽然在训练集中可以得到较高的准确率,但是很有可能出现过拟合的现象使得在实际应用的效果欠佳。如果对生成的样本区间进行放大,一方面可以使得数据之间的相似性大量降低也可以找出更多的潜在的异常行为,但是这样就需要以牺牲一些准确率为代价。下图为对负样本生成区间进行放大后的准确率结果图:

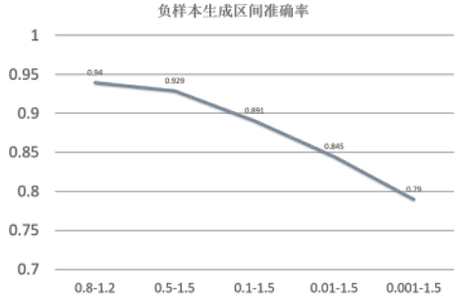

图3 负样本生成区间对结果的影响

从上图可以看出不同数据生成区间对最终随机森林模型的自验证准确率有着很大的影响,当数据生成区间在一个比较

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

小的区间的时候，比如0.8到1.2时自验证的准确率偏高，因为在这个区间上生成的样本都和异常点具有很强的相似性，由于异常点和正常点之间的数值差异本身就比较大，所以此时分类的准确率偏高。但是这样带来的问题是，只有异常情况很明显下才能被分类成具有异常行为的IP点。为了能找到更多的异常行为的IP点，可以稍微放大一下数据生成的区间，比如从0.8到0.5再到0.1，可以看出对特征数据的上限只增大到了1.5倍就不再继续增大了，这是因为，对于负样本来说，它们的特征数值的上限已经足够大了相比正常的IP点来说，但是它们的下限却要比正常点的上限还要高很多，所以在保证上限不变的情况下，适当增大下限的取值，可以让随机森林模型找到更多的潜在的具有异常行为的IP点。

3 特征重要性衡量

基于已经生成的随机森林模型，一方面可以对新的特征数据进行是否具有异常行为的分类任务，另一方面也可以通过该模型度量每一个特征的重要程度。计算某一特征X流程如下：

1）对每一颗决策树，选择相应的袋外数据，因为我们在建立每一颗决策树时都是随机的进行有放回的选取也就是重复抽样，所以最终大概仍有三分之一的样本点没有被抽取到对于每一颗决策树来说。用这部分数据计算模型的错误率，记作为errOOB1.

2）随机对袋外数据所有样本的X特征加入噪声干扰，一般来说简单的做法就是随机改变样本数据在X特征处的值，经过这样的变换后，再次计算模型的错误率，记作为errOOB2.

3）假设随机森林中有N颗数，则特征X的重要性的计算公式为：

$$\text{importance} = \sum (errOOB1 - errOOB2)/N$$

之所以用这样的式子来表示特征的重要性，原因在于如果对样本数据的X特征进行随机改变后，袋外数据的准确率大幅下降，也就是意味着errOOB2会出现大幅上升的现象，那么就说明这个特征对最终的分类结果产生了比较大的影响，所以该特征的重要性也就比较大。对生成的数据进行特征重要性的衡量，结果如下图所示：

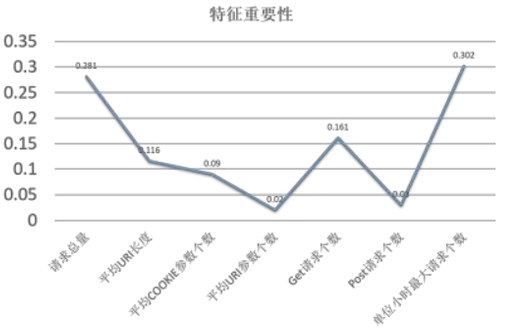

图4　基于随机森林模型的特征重要性衡量

从图中可以看出，不同的特征具有的重要程度具有很大幅度的变化在随机森林模型中，其中单位小时最大请求个数这个特征具有最大的重要程度，而和uri,cookie相关的特征重要程度相对较小。这些特征重要程度上也可以得知，一般的网络异常行为主要集中体现在请求的量上，与该指标相关的量会对最终一个IP点是否具有异常行为有着更大的权重。

参考文献：

[1] Shi, Lin.Abnormal organization of white matter network inpatients with no dementia after ischemic stroke[C]. PloS one, 2013: 8, 12.

[2] Soltani, Somayehl. A survey on real world botnets and detection mechanisms[J]. International Journal of Information & Network Security, 2014, 3(2).

[3] Narudin, Fairuz Amalina. Evaluation of machine learning classifiers for mobile malware detection[J]. Soft Computing,2016: 1, 20.

[4] Qian Quan, TianhongWang, Rui Zhang.Relative Network Entropy based Clustering Algorithm for Intrusion Detection[J]. Network Security, 2013, 15(1):16-22.

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net