基于机器学习的web异常检测

机器学习 阿里聚安全 检测 异常 web

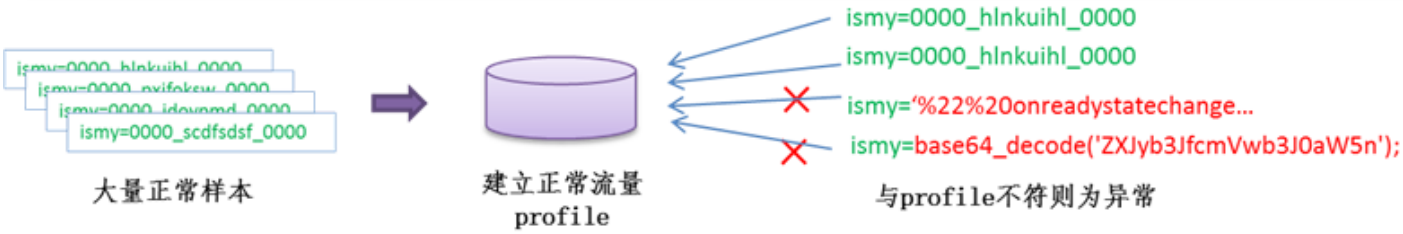
阿里聚安全 2017年02月08日发布

基于机器学习的web异常检测

Web防火墙是信息安全的第一道防线。随着网络技术的快速更新，新的黑客技术也层出不穷，为传统规则防火墙带来了挑战。传统web入侵检测技术通过维护规则集对入侵访问进行拦截。一方面，硬规则在灵活的黑客面前，很容易被绕过，且基于以往知识的规则集难以应对0day攻击；另一方面，攻防对抗水涨船高，防守方规则的构造和维护门槛高、成本大。

基于机器学习技术的新一代web入侵检测技术有望弥补传统规则集方法的不足，为web对抗的防守端带来新的发展和突破。机器学习方法能够基于大量数据进行自动化学习和训练，已经在图像、语音、自然语言处理等方面广泛应用。然而，机器学习应用于web入侵检测也存在挑战，其中最大的困难就是标签数据的缺乏。尽管有大量的正常访问流量数据，但web入侵样本稀少，且变化多样，对模型的学习和训练造成困难。因此，目前大多数web入侵检测都是基于无监督的方法，针对大量正常日志建立模型(Profile)，而与正常流量不符的则被识别为异常。这个思路与拦截规则的构造恰恰相反。拦截规则意在识别入侵行为，因而需要在对抗中“随机应变”；而基于profile的方法旨在建模正常流量，在对抗中“以不变应万变”，且更难被绕过。

基于异常检测的web入侵识别，训练阶段通常需要针对每个url，基于大量正常样本，抽象出能够描述样本集的统计学或机器学习模型(Profile)。检测阶段，通过判断web访问是否与Profile相符，来识别异常。



对于Profile的建立，主要有以下几种思路：

1. 基于统计学习模型

基于统计学习的web异常检测，通常需要对正常流量进行数值化的特征提取和分析。特征例如，URL参数个数、参数值长度的均值和方差、参数字符分布、URL的访问频率等等。接着，通过对大量样本进行特征分布统计，建立数学模型，进而通过统计学方法进行异常检测。

2. 基于贝叶斯分析的机器学习模型

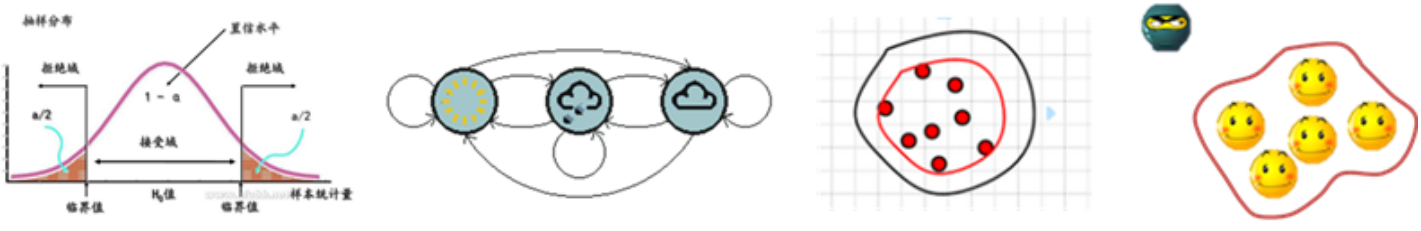
Web异常检测归根结底还是基于日志文本的分析，因而可以借鉴NLP中的一些方法思路，进行文本分析建模。这其中，比较成功的是基于隐马尔科夫模型(HMM)的参数值异常检测。

3. 基于单分类模型

由于web入侵黑样本稀少，传统监督学习方法难以训练。基于白样本的异常检测，可以通过非监督或单分类模型进行样本学习，构造能够充分表达白样本的最小模型作为Profile，实现异常检测。

4. 基于聚类模型

通常正常流量是大量重复性存在的，而入侵行为则极为稀少。因此，通过web访问的聚类分析，可以识别大量正常行为之外，小搓的异常行为，进行入侵发现。



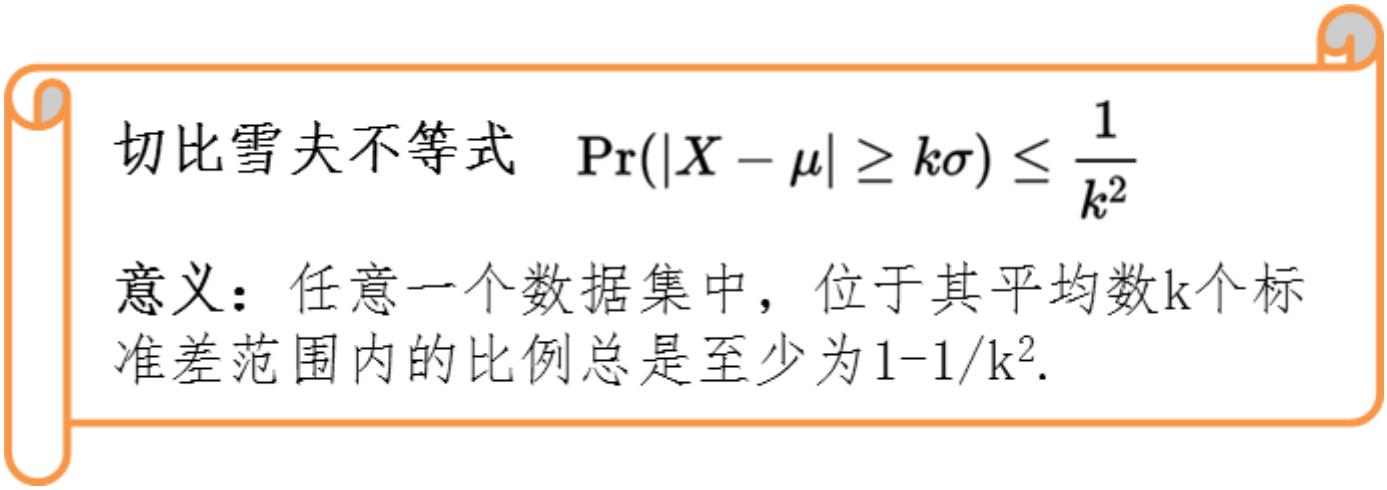
基于统计学习模型

据。

这里以斯坦福大学CS259D: Data Mining for CyberSecurity课程[1]为例，介绍一些行之有效的特征和异常检测方法。

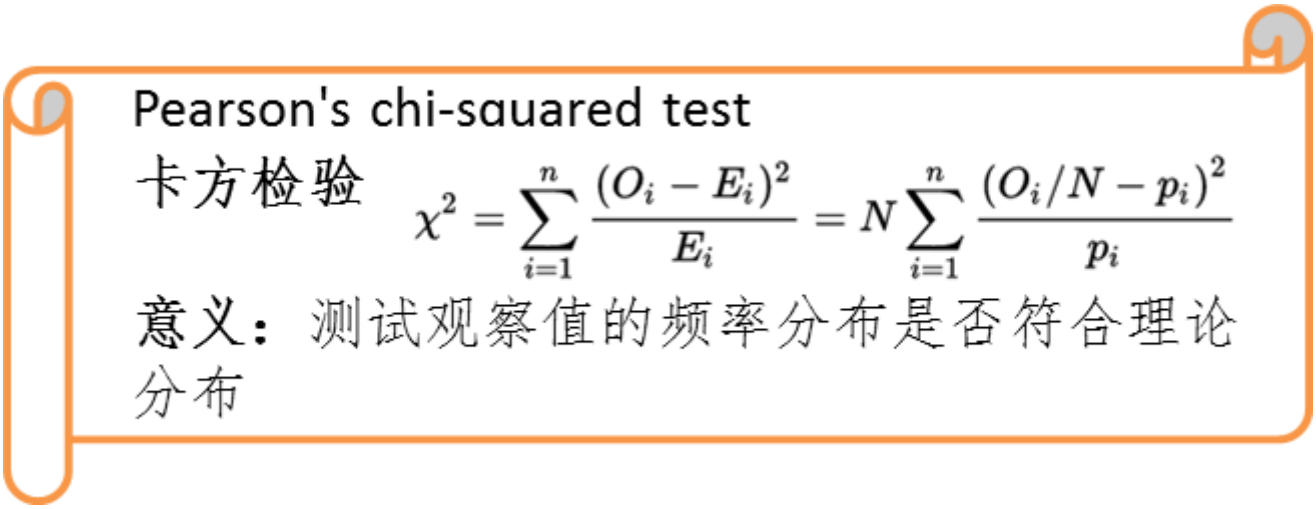
特征1：参数值value长度

模型：长度值分布，均值μ，方差σ2，利用切比雪夫不等式计算异常值p



特征2：字符分布

模型：对字符分布建立模型，通过卡方检验计算异常值p



特征3：参数缺失

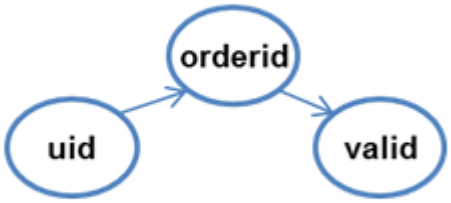
模型：建立参数表，通过查表检测参数错误或缺失

特征4：参数顺序

模型：参数顺序有向图，判断是否有违规顺序关系

1. 通过有向图表示参数顺序

uid=123&orderid=12345&valid=true



2. 求取强连图子图（SCC，Tarjan algorithm）

3. 形成顺序约束表

特征5：访问频率（单ip的访问频率，总访问频率）

模型：时段内访问频率分布，均值μ，方差σ2，利用切比雪夫不等式计算异常值p

特征6：访问时间间隔

模型：间隔时间分布，通过卡方检验计算异常值p

最终，通过异常打分模型将多个特征异常值融合，得到最终异常打分：

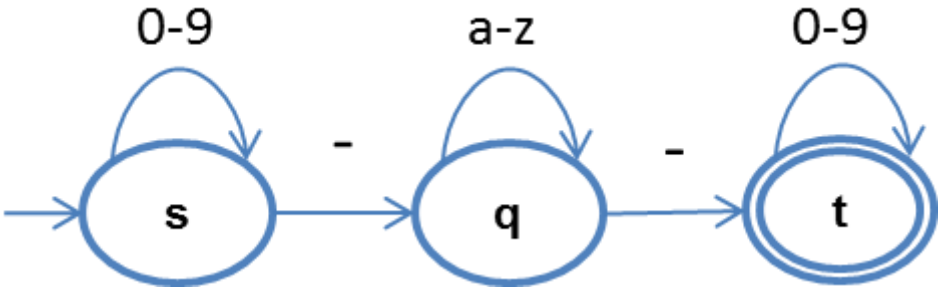
$$\sum_m w_m \times (1 - p_m)$$

基于文本分析的机器学习模型

URL参数输入的背后，是后台代码的解析，通常来说，每个参数的取值都有一个范围，其允许的输入也具有一定模式。比如下面这个例子：

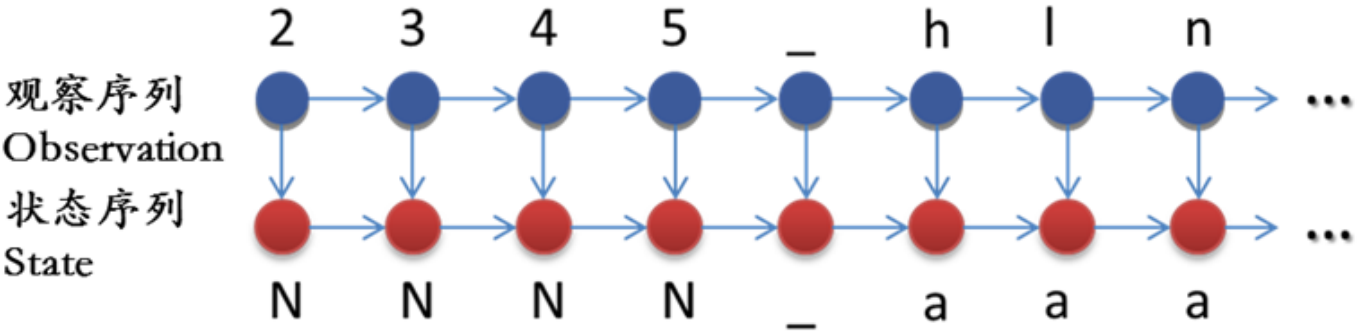
```
https://somedomain.com/alibaba/report?mid=6492_abc_7756
https://somedomain.com/alibaba/report?mid=1234_feagada_7680
https://somedomain.com/alibaba/report?mid=2345_hlnkl_9000
https://somedomain.com/alibaba/report?mid=base64_decode
```

例子中，绿色的代表正常流量，红色的代表异常流量。由于异常流量和正常流量在参数、取值长度、字符分布上都很相似，基于上述特征统计的方式难以识别。进一步看，正常流量尽管每个都不相同，但有共同的模式，而异常流量并不符合。在这个例子中，符合取值的样本模式为：**数字 *字母* 数字**，我们可以用一个状态机来表达合法的取值范围：

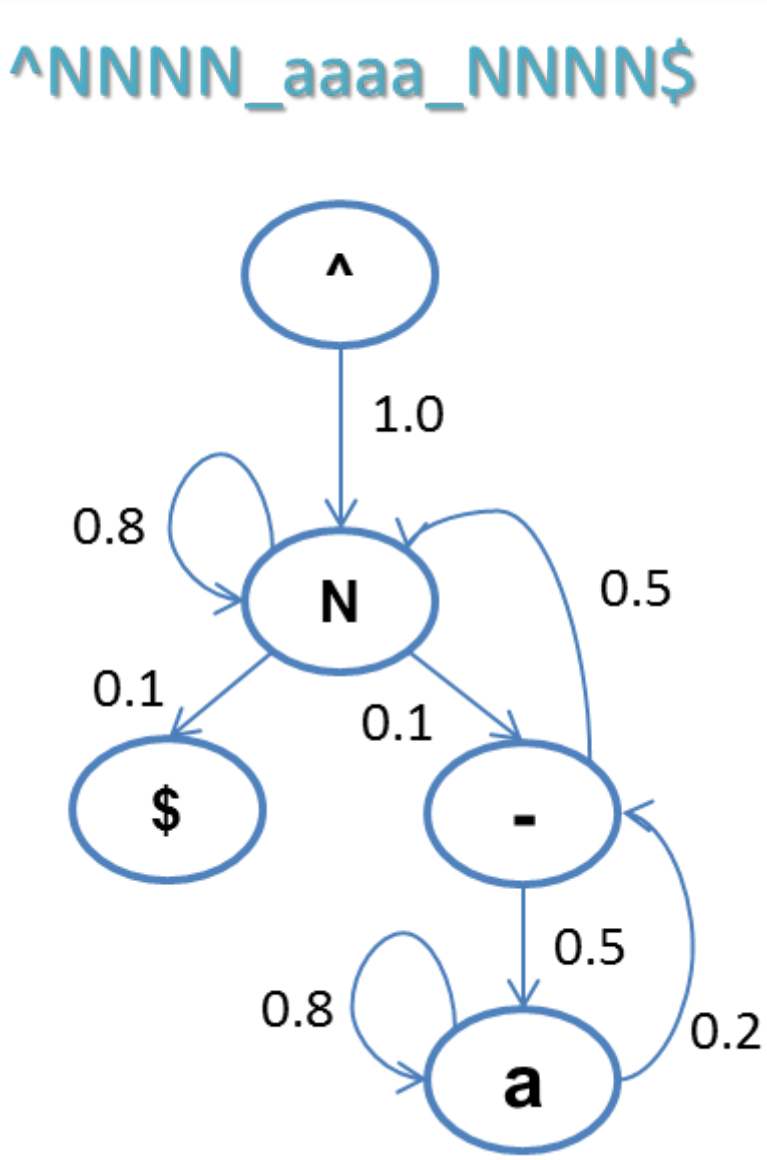


对文本序列模式的建模，相比较数值特征而言，更加准确可靠。其中，比较成功的应用是基于隐马尔科夫模型(HMM)的序列建模，这里仅做简单的介绍，具体请参考推荐文章[2]。

基于HMM的状态序列建模，首先将原始数据转化为状态表示，比如数字用N表示状态，字母用a表示状态，其他字符保持不变。这一步也可以看做是原始数据的归一化(Normalization)，其结果使得原始数据的状态空间被有效压缩，正常样本间的差距也进一步减小。



紧接着，对于每个状态，统计之后一个状态的概率分布。例如，下图就是一个可能得到的结果。“^”代表开始符号，由于白样本中都是数字开头，起始符号(状态^)转移到数字(状态N)的概率是1；接下来，数字(状态N)的下一个状态，有0.8的概率还是数字(状态N)，有0.1的概率转移到下划线，有0.1的概率转移到结束符(状态$)，以此类推。



利用这个状态转移模型，我们就可以判断一个输入序列是否符合白样本的模式：

Observation ismy=2345_hlnkl_9000

State ismy=NNNN_aaaa_NNNN

$P(w) = 1.0*(0.8)^3*0.1*0.5*(0.8)^3*0.2*(0.8)^3*0.1$

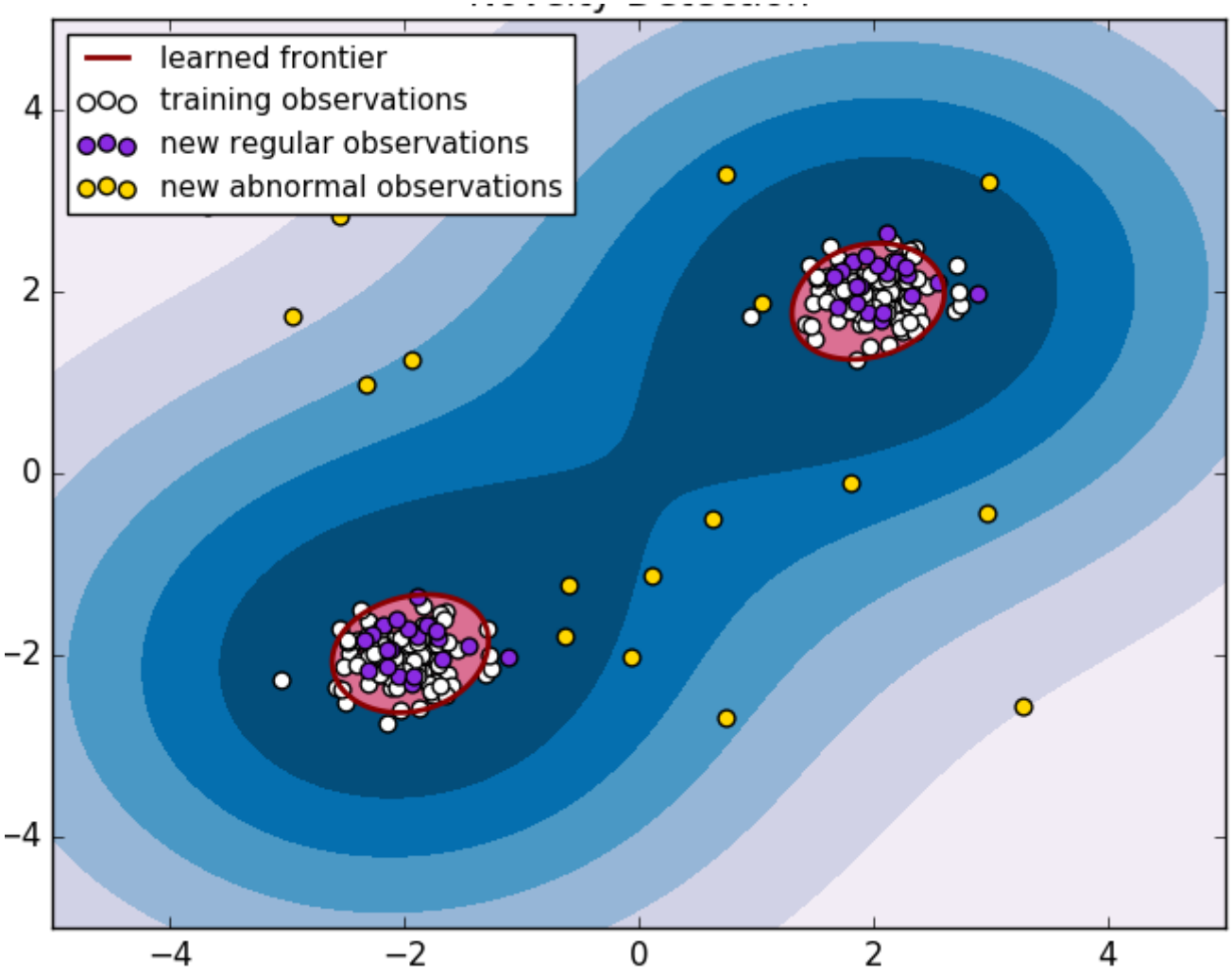
Observation ismy=base64_decode('ZXJyb3JfcmVwb3J0aW5n');

State ismy=aaaaNN_aaaaaa('AAAaaNaaaaAaaNANaANa');

$P(w) = 0.0*(0.8)^3*0.0*0.1*(0.8)^5*0.0...$

正常样本的状态序列出现概率要高于异常样本，通过合适的阈值可以进行异常识别。

在二分类问题中，由于我们只有大量白样本，可以考虑通过单分类模型，学习单类样本的最小边界，边界之外的则识别为异常。



这类方法中，比较成功的应用是单类支持向量机(one-class SVM)。这里简单介绍该类方法的一个成功案例McPAD的思路，具体方法关注文章[3]。

McPAD系统首先通过N-Gram将文本数据向量化，对于下面的例子，

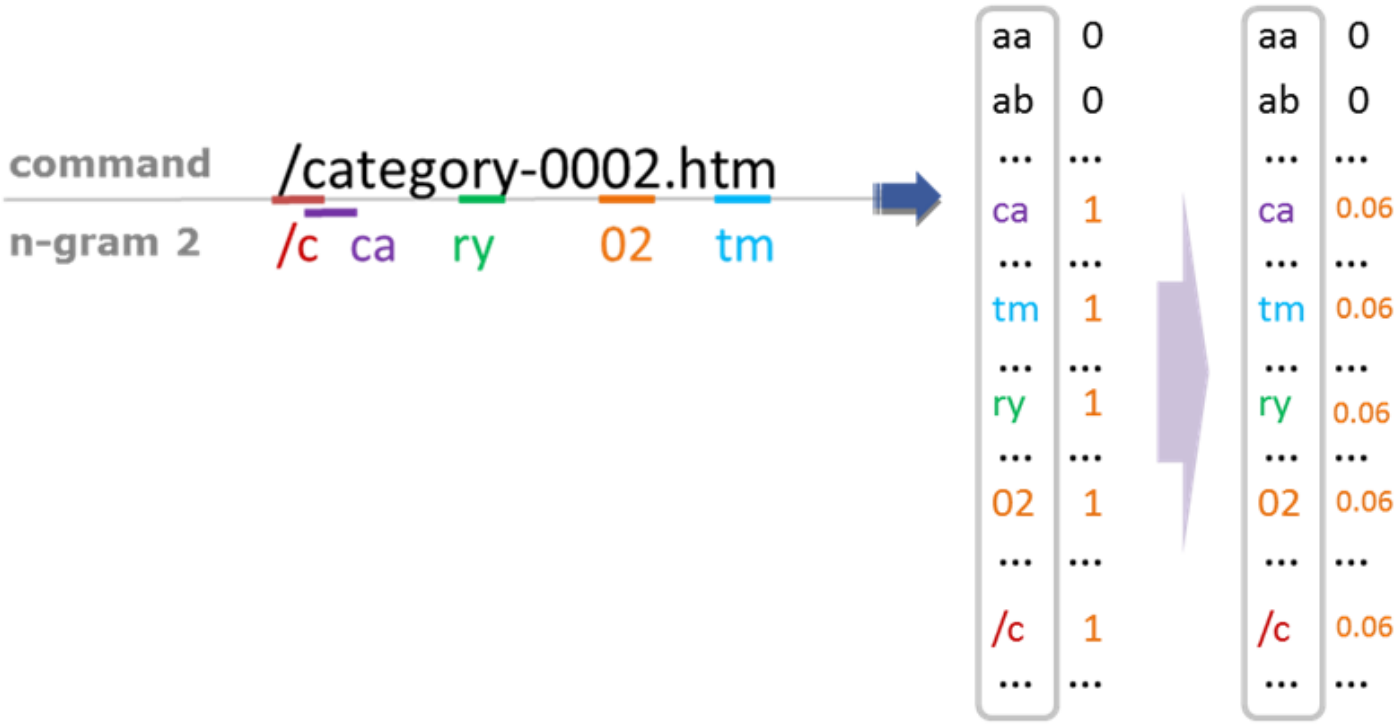
http://abc.com/test?path=/category-0001.htm
http://abc.com/test?path=/category-0002.htm

首先通过长度为N的滑动窗口将文本分割为N-Gram序列，例子中，N取2，窗口滑动步长为1，可以得到如下N-Gram序列。

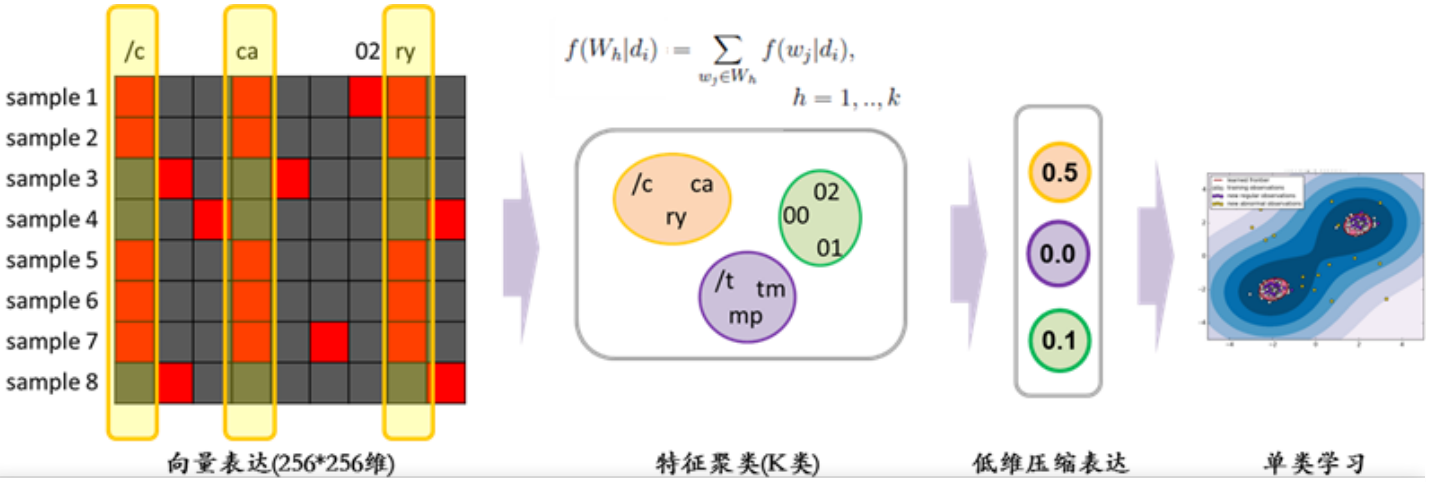
下一步要把N-Gram序列转化成向量。假设共有256种不同的字符，那么会得到256*256种2-GRAM的组合(如aa, ab, ac …)。我们可以用一个256256长的向量，每一位one-hot的表示(有则置1，没有则置0)文本中是否出现了该2-GRAM。由此得到一个256*256长的0/1向量。进一步，对于每个出现的2-Gram，我们用这个2-Gram在文本中出现的频率来替代单调的“1”，以表示更多的信息：

$$f(\beta|B) = \frac{\#\text{ of occurrences of } \beta \text{ in } B}{l - n + 1}$$

至此，每个文本都可以通过一个256*256长的向量表示。

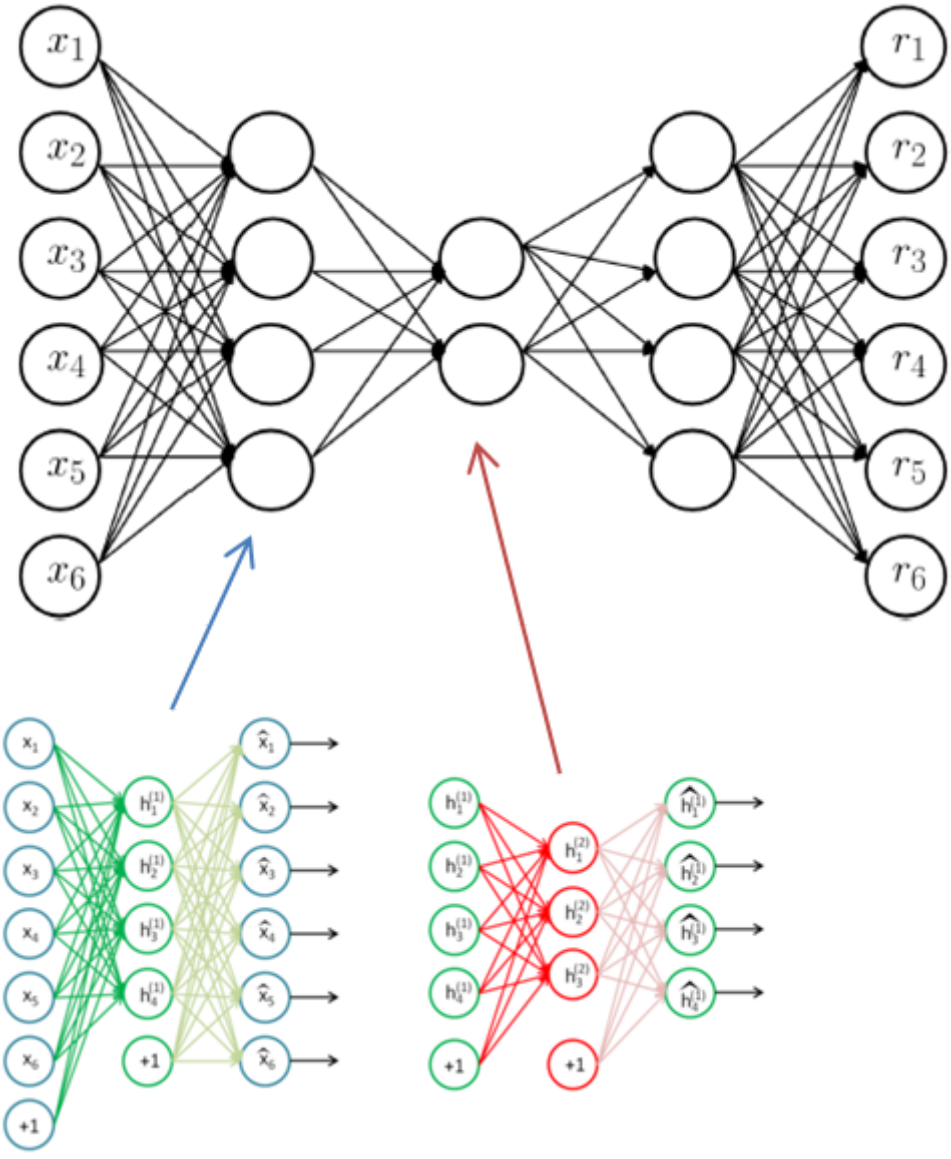


现在我们得到了训练样本的256*256向量集，现在需要通过单分类SVM去找到最小边界。然而问题在于，样本的维度太高，会对训练造成困难。我们还需要再解决一个问题：如何缩减特征维度。特征维度约减有很多成熟的方法，McPAD系统中对特征进行了聚类达到降维目的。

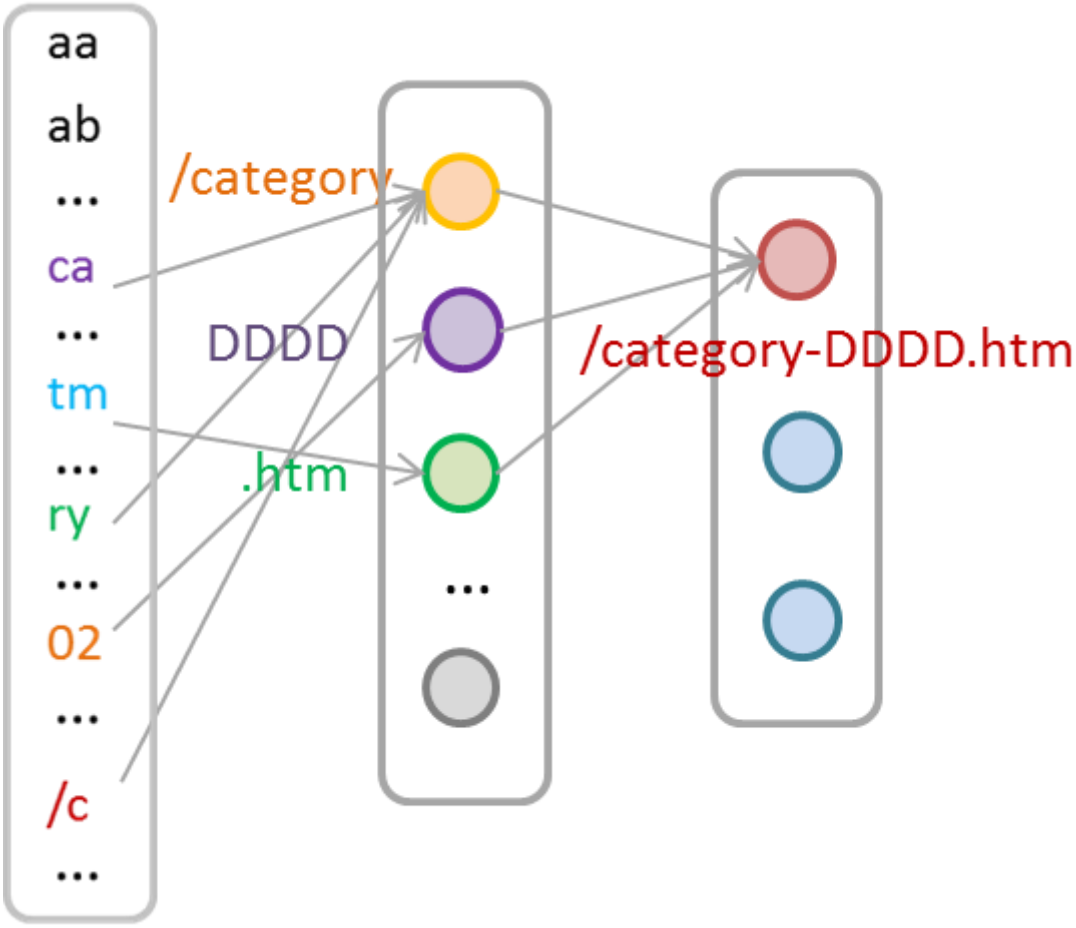


上左矩阵中黑色表示0，红色表示非零。矩阵的每一行，代表一个输入文本(sample)中具有哪些2-Gram。如果换一个角度来看这个矩阵，则每一列代表一个2-Gram有哪些sample中存在，由此，每个2-Gram也能通过sample的向量表达。从这个角度我们可以获得2-Gram的相关性。对于2-Gram的向量进行聚类，指定的类别数K即为约减后的特征维数。约减后的特征向量，再投入单类SVM进行进一步模型训练。

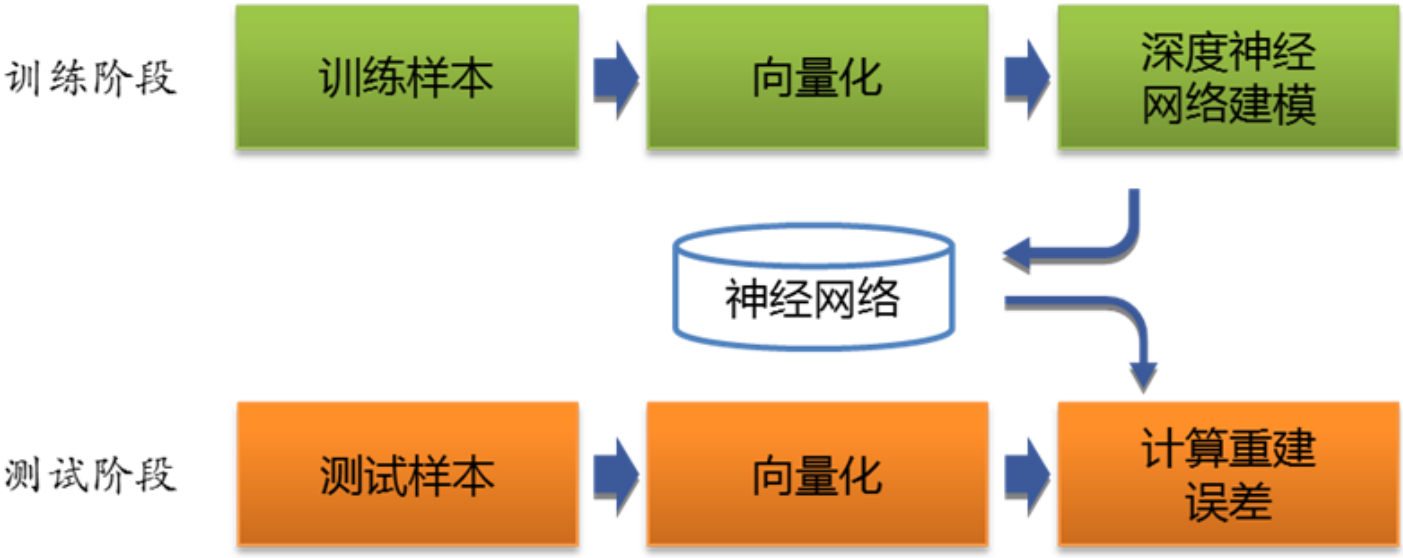
再进一步，McPAD采用线性特征约减加单分类SVM的方法解决白模型训练的过程，其实也可以被深度学习中的深度自编码模型替代，进行非线性特征约减。同时，自编码模型的训练过程本身就是学习训练样本的压缩表达，通过给定输入的重建误差，就可以判断输入样本是否与模型相符。



我们还是沿用McPAD通过2-Gram实现文本向量化的方法，直接将向量输入到深度自编码模型，进行训练。测试阶段，通过计算重建误差作为异常检测的标准。



基于这样的框架，异常检测的基本流程如下，一个更加完善的框架可以参见文献[4]。



本文管中窥豹式的介绍了机器学习用于web异常检测的几个思路。web流量异常检测只是web入侵检测中的一环，用于从海量日志中捞出少量的“可疑”行为，但是这个“少量”还是存在大量误报，只能用于检测，还远远不能直接用于WAF直接拦截。一个完备的web入侵检测系统，还需要在此基础上进行入侵行为识别，以及告警降误报等环节。

2017阿里聚安全算法挑战赛将收集从网上真实访问流量中提取的URL，经过脱敏和混淆处理，让选手利用机器学习算法提高检测精度，真实体验这一过程。并有机会获得30万元奖金，奔赴加拿大参加KDD----国际最负盛名的数据挖掘会议！

报名地址：https://tianchi.shuju.aliyun....

推荐阅读

1. CS259D: Data Mining for CyberSecurity, 课程网址：http://web.stanford.edu/class...
2. 楚安，数据科学在Web威胁感知中的应用，http://www.jianshu.com/p/942d...
3. McPAD : A Multiple Classifier System for Accurate Payload-based Anomaly Detection, Roberto Perdisci
4. AI2 : Training a big data machine to defend, Kalyan Veeramachaneni

作者：七雨@阿里聚安全 ，更多阿里安全类技术文章，请访问阿里聚安全博客

2017年02月08日发布 ···

赞 | 0 收藏 | 4

算法开启的人工智能时代！阿里聚安全算法挑战赛公开报名！ 887 浏览

谷歌发布基于机器学习的Android APP安全检测系统：Google Play Protect 602 浏览

30万奖金！还带你奔赴加拿大相约KDD！？阿里聚安全算法挑战赛带你飞起！ 675 浏览

评论

默认排序 | 时间排序

文明社会，理性评论

发布评论

阿里聚安全

479 声望

关注作者

发布于专栏

阿里聚安全

阿里聚安全（http://jaq.alibaba.com）由阿里巴巴移动安全部出品，面向企业和开发者提供企业安全解决方案，全面覆盖移动安全、数据风控、内容安全、实人认证等维度，并在业界率先提出“以业务为中心的安…

45 人关注

关注专栏