人工智能在Web安全中的应用

冯景辉@Baidu

从ModSecurity开始说起

SecRule REQUEST_COOKIES|!REQUEST_COOKIES:/__utm/|REQUEST_COOKIES_NAMES|ARGS_NAMES|ARGS|XML:/* "(?i:(?:(union(.*?)select(.*?)from)))"

- 多阶段
- 多字段
- 转换函数
- 多动作

绕过仍然无法避免

```
REVERSE(noinu)+REVERSE(tceles)
```

```
un?+un/**/ion+se/**/lect+
```

SQL Tokenizer Parser Analyzer

语法解析

- 关键词解析
- 语法规则
- 基本函数

语义分析

- SQL补全
- 环境感知
- 注入检测
- 语义行为

libinjection

兼容性

除了MySQL，其他SQL

误报

本质上，系统将尽量补全SQL，而SQL一旦通过语法分析，只要存在Token，误报就容易出现

机器学习初探

典型的机器学习场景

有监督学习 VS 无监督学习

图像识别

NLP

关联新闻

机器学习初探

特征选取

基于Payload的特征选择，需要结合安全特性，比如关键字、字符特征、甚至请求长度，同时避免过拟合

样本训练

选取大量黑白标注样本，同时要控制样本类型的分布

特征选取 01

算法选择 02

样本训练 03

日志审计 04

算法选择

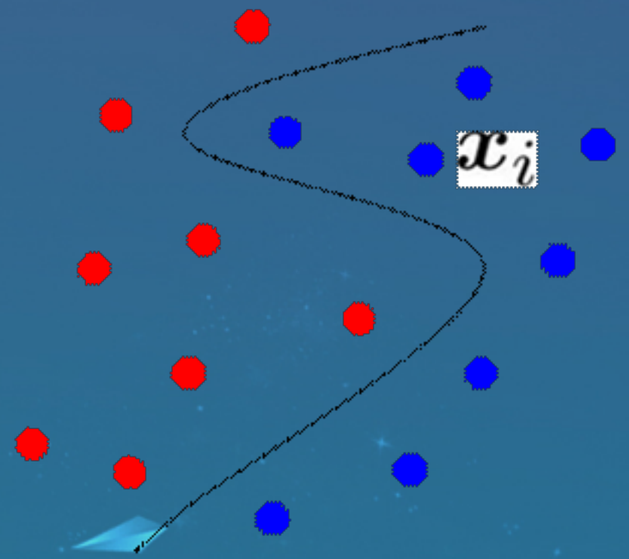
有监督学习有诸多常用算法、SVM、HMM、贝叶斯等等

日志审计与回归

当前有监督学习主要应用在离线日志分析中，快速发现未知攻击样本

支持向量机-XSS检测应用

SVM的典型问题



结构风险最小，
而非经验风险最小

特征选取

- URL长度
- 第三方域名个数
- 敏感字符
- JS关键字

召回率

93%

准确率

90%

支持向量机-不足

不适合大规模数据集训练

广泛采用的LibSVM，在最坏情况下复杂度为O(n^2)（训练样本数平方）

本质上与规则无异

可以对抗基本变形，只是对原有规则系统提供一定的宽容度

准确度无法满足需呀

对原有系统提供一个离线检查机制

是否能够结合更多的识别方法

隐马尔可夫

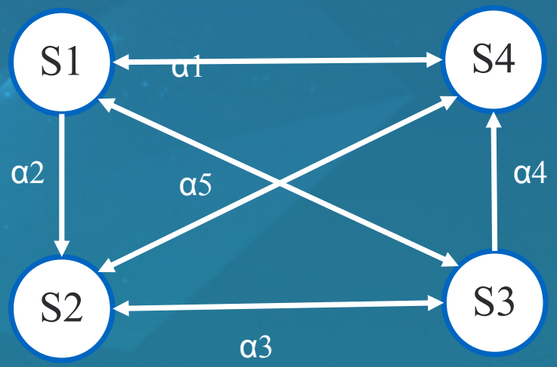
最大熵模型

<script>alert(0)</script>

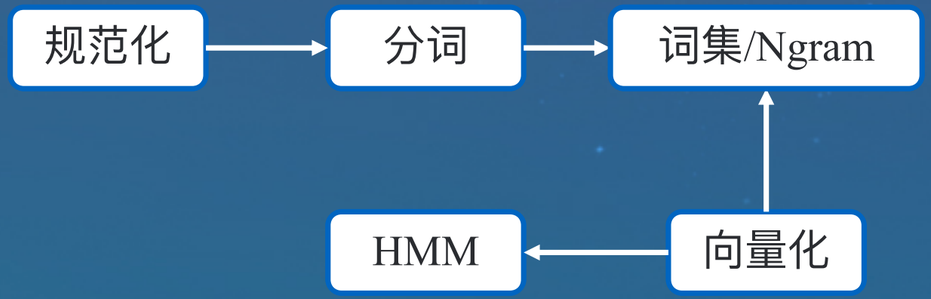
S1:符号
S2:字符
S3:数字
S4:分割符号

观察序列

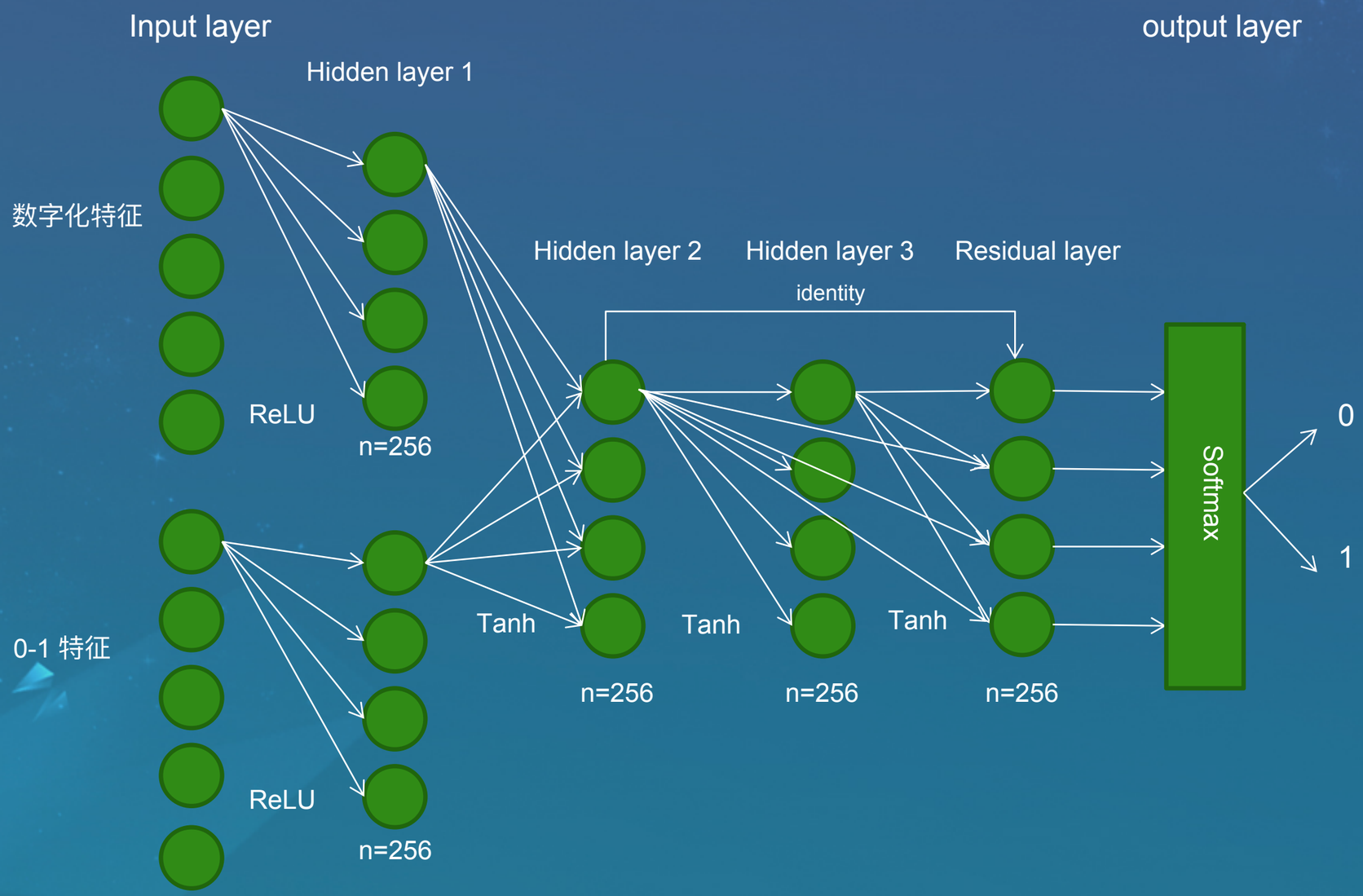
隐含序列



加入词法之后



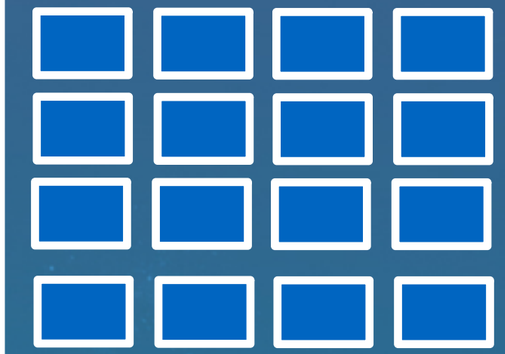
从浅层学习走向深度神经网络



从浅层学习走向深度神经网络

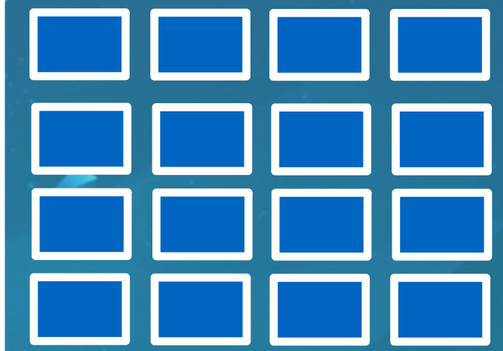
特征提取

请求



结构特征

- URL长度
- 特殊字符数量
- JS关键字数量
- SQL关键字数量
- UA



经验特征

- “(“数量
- Union
- 参数个数
- 单参数section

数字化特征

205	3	34.5	143234
285	68	296	7
13850	157	11218	847
1.23e+9	422	1004	177
0	398	13.333	125

数值型特征

0	0	0	0
0	0	0	1
0	1	0	0
0	1	1	0
0	1	0	0

布尔特征

见证奇迹的时刻

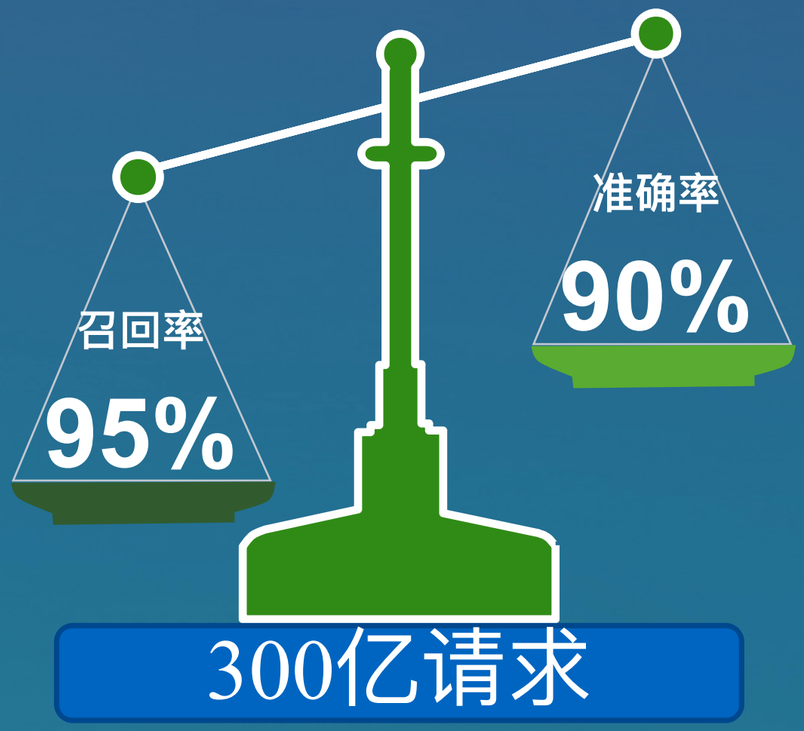
一些奇怪的发现

```
P                                   O                                   S                                   T
/index/index.php?_c=zip://d://KAS/WebSource/ueditor/php/upload/file/20170531/1496216087803962.zip#xxx&_m=captcha
cmd=echo "\n\n\n", system("dir C:");exit;

%2527!=(hex(user())>0x23)%2523
```

通过不断调整特征，对于变形与绕过有了神奇的抵抗能力，但是准确率却无法提升

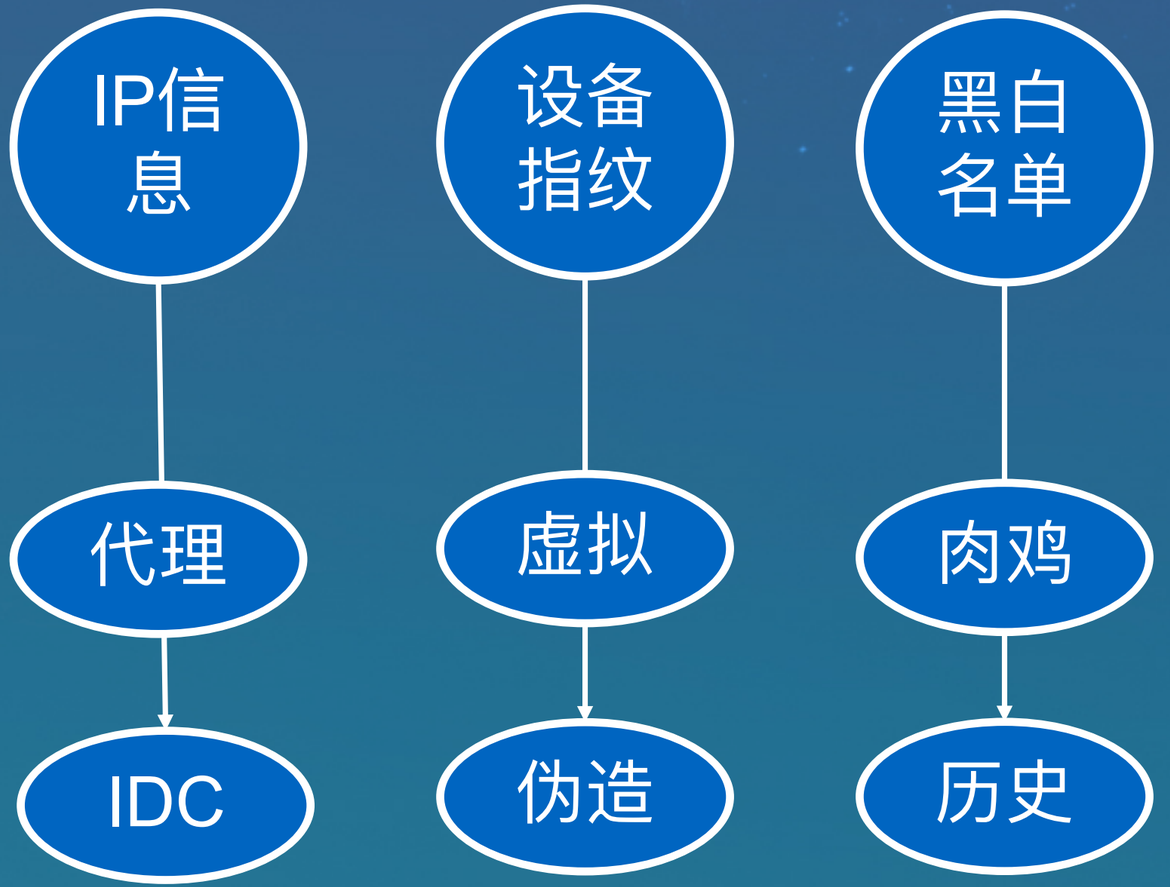
如果我们在结合Response呢?



威力不止如此

如果机器学习只做文本特征检测，
不能称之为人工智能

威胁特征全貌



威力不止如此

如果机器学习只做文本特征检测，
不能称之为人工智能

威胁特征全貌

威力不止如此

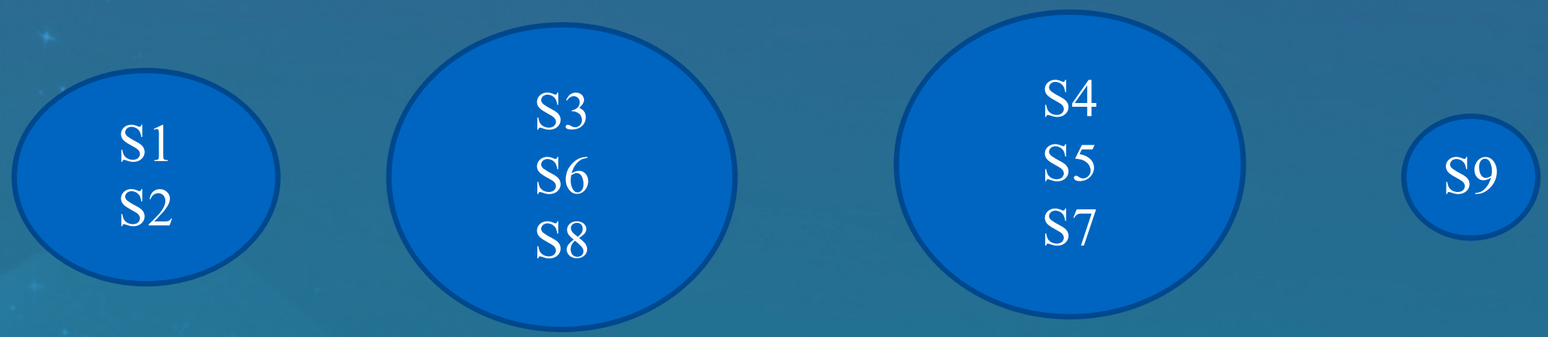
如果机器学习只做文本特征检测，
不能称之为人工智能

威胁特征全貌

用户行为分析-电商案例

	A0	A1	A2	A3	A4	An
API	登录	浏览	搜索	购物车	抽奖	结算
Session	{A2,A1,A4,A8}	{A2,A1,A4,A4}	{A1,A1,A4,A8}	{A0,A1,A4,A8}		

无监督学习
K-means



产出	盗号	僵尸	恶意爬虫	未知威胁

用户行为分析-难点

业务抽象

通过n-gram算法，产生业务pattern,分析URL，将请求归类，实现业务抽象

不判别好坏，只寻找少数派，相信大多数用户都是正常业务

去噪

去除网络、浏览器等干扰，将Session中所有业务向量化

因为无法识别异常类型，还需要人工介入和辅助模型识别

关系向量化

每Session的API集合，交集

异常识别的准确率高达95%

算法

如何选择K值，还要考虑到的向量集合的方差

总结

有监督学习，有效降低规则维护工作量，但对于准召相比语法引擎没有突破

在样本空间扩大之后，DNN相比SVM能有效提高召回率，但更多的应用在离线场景

UBA可以解决当前技术在高维空间上的不足，是安全的对抗的下一个风口

无监督学习是未来，能突破样本空间限制

THANKS

Thanks for watching