基于学习的分类器的实际逃避：

案例研究

                                                                          Nedim Srndi c和Pavel Laskov'

认知系统系

蒂宾根大学

德国图宾根

{ nedim.srndic，pavel.laskov } @ uni-tuebingen.de

*摘要* - 基于学习的分类器越来越多地用于检测各种形式的恶意数据。但是，如果在线部署，攻击者可能会试图通过操纵数据来逃避攻击。这种攻击的例子之前已经在假设攻击者完全了解部署的分类器的情况下进行了研究。实际上，这样的假设很少成立，尤其是对于在线部署的系统。关于部署的分类器系统的大量信息可以从各种来源获得。在本文中，我们通过一个真实的部署系统PDF RATE ，实验性地研究了分类器规避的有效性作为一个测试案例。我们为实际逃避策略开发了一个分类法，并调整已知的逃避算法来实现分类中的特定场景。我们的实验结果显示，即使是简单的攻击，PDF RATE 的分类分数和检测精度也大幅下降。我们进一步研究防范分类器逃避的潜在防御机制。我们的实验显示，针对PDF RATE 提出的原始技术只有在执行的攻击与预期攻击完全匹配时才有效。在讨论我们的研究结果时，我们分析了一些潜在的技术，用于提高学习系统对数据对抗操纵的稳健性。

一，引言

数据分析方法（如机器学习）越来越多地用于安全应用。对于像恶意软件分析这样的任务来说，部署学习方法已经变得非常重要。数据驱动的分析能够将看似不同种类的恶意软件样本自动归属到适量的正版恶意软件系列[1]，[2]。最近的工作也见证了机器学习的多种创新应用，用于检测各种安全违规行为，例如驱动下载[3]，[4]，恶意网页[5]，[6]，被盗用账户和伪造社交网络中的身份[7]，[8]，不需要的P2P流量[9]等等。

显然，在任何安全关键的环境中部署学习方法都需要能够抵御潜在的攻击。机器学习方法的安全性先前已经从概念[10]，方法[11]，[12]，[13]和[14]以及实际的[15]，[16]，[17]观点进行了讨论。尽管越来越多的证据表明基于学习的对抗数据操纵方法的敏感性，但这似乎阻碍了它们被接受为数据密集型安全任务的多功能工具。通常，所提出的基于学习的技术的安全性分析是非正式进行的，有时由实验评估支持。

基于学习的方法的安全评估面临着几个挑战。主要的理论障碍是在数据分析的背景下缺乏对安全的正式定义。与隐私相比，已经提出了几种形式，例如隐私保护数据挖掘[18]或差分隐私[19]，与机器学习没有建立安全目标的正式联系。从实际角度来看，针对学习算法的攻击的成功与否关键取决于攻击者可用的知识数量。以前报告的大多数成功的攻击都假定攻击者对学习模型有完全的了解[20]，[15]，[16]，[21]，[17]，[22]。因此，可以认为，减少模型泄露的知识量，

尽管如此，攻击者仍然可能学习的是“野外”部署的以学习为基础的方法，以及如何利用这些信息。为了调查这个问题，我们提供了一个案例研究的结果，我们在一个真正的基于学习的系统上进行了PDF RATE[[1]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn1" \o ")，一个用于检测PDF恶意软件的在线服务[23]。对于任何提交的PDF文件，PDF RATE 提供了对其恶意的概率估计。我们的研究解决了攻击者试图通过修改提交的PDF文件来逃避检测的情况，以便其恶意功能保持完好，但PDF RATE 返回的概率得分降低。

我们继续介绍两种逃避策略，适用于攻击者可用的知识数量不同的几种攻击场景。由于PDF RATE 是一个研究系统，它的方法和技术细节在最初的研究论文[23]和随附的技术报告[24]中有相当详细的记载。基于这些信息，可以部分重构用于模型创建和评估的特征，重现训练过程，甚至独立地获得一些训练数据[[2]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn2" \o ")。为了系统地探讨攻击者的选择，我们定义了一套**正交的**回避策略，反映了不同程度的可用知识，详见第二节。我们的规避方法的基本思想是将虚拟内容插入PDF文件中，这被PDF渲染器忽略，但会影响PDF RATE中使用的特性的计算，如V-B所述。一旦我们可以影响PDF RATE 特性的一个子集，我们开发构建攻击实例的算法，在V-C节中介绍。在第六部分的实验中，我们评估了我们的策略对从PDF RATE 已知的数据集中随机抽取的100个恶意文件的有效性。

我们的研究结果显示，即使有最少量的可用信息，即可以自由修改六分之一并增加六分之一特征的能力，我们的攻击将PDF RATE 的分类分数从几乎100％降低到约33 ％。有关分类器的其他信息，如其类型（微不足道）的知识和拥有训练数据集（更难以获得），进一步将中值分数降低到约28％。

我们已经分析了PDF RATE 的作者提出和评估的防御策略，虽然我们不知道它是否部署在在线系统中。文献[23]的攻击方案假定攻击者利用一小部分信息特征。结果表明，通过将一小部分预期的攻击数据包含到训练集中，可以有效地阻止这种攻击。我们重建了这一攻击，并验证了原有防御战略的有效性。然而，这种主动防御最终只能有效抵抗攻击的“紧张”。每当执行的攻击与预期攻击不符时，主动防御的效果基本消失，检测精度降到10％以下。

我们的贡献可以概括如下：

•     我们提出了一个基于学习的检测技术安全性实践评估的一般模型。该模型可以对攻击者可利用的各种信息泄漏进行系统的探索，并且适用于具有可修改子集的特征的PDF RATE 之外的系统。

•     我们提出了两种在各种情况下针对部署分类模型进行的回避攻击。

•     我们演示了第一次针对在“野外”部署的基于学习的分类器的自动化实际攻击，这些分类器在没有学习模型知识的情况下完成，完全在问题空间中进行。

•     我们为在我们的研究中进行的所有实验提供了一个开源软件框架，用于独立验证和扩展我们的结果。

II。EVASION ATTACKS AGAINST LEARNING SYSTEMS 逃避攻击学习系统

任何在现实世界环境中部署的，基于学习的系统，其中存在着极为重要的经济，政治或军事利益，肯定会通过操纵系统来吸引个人或团体的注意力，从而影响其决定。这种活动有很多例子。除了

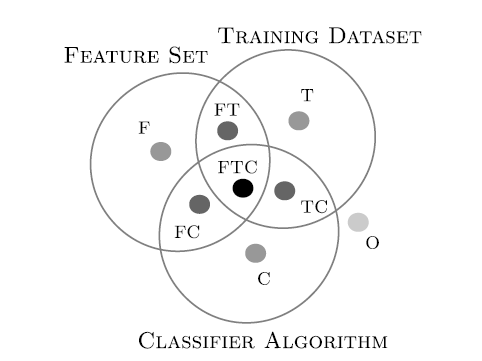


图1.分类系统的逃避场景的分类。在每个场景中，如果场景点在该组件的圈内，则关于给定的分类器组件的知识是高的，否则是低的。

引言中提到的计算机安全应用，这种操作的潜在场景包括敌对广告[25]，垃圾邮件检测[26]，[27]，写作风格的认可[28]，剽窃检测[29]等等。

在这项工作中，我们关注*分类器*，这是一种特殊类型的学习系统，它将新数据分为两个或多个预定义的类别。分类器通常通过计算一些数字或概率分数并将其与固定阈值进行比较来做出预测。旨在操纵分类器的对手的目标是*将*其*混淆*为提供错误的分类。对于二元分类问题，错误分类被称为false positives和*错误否定*false negatives。从对抗的角度来看，有关基于学习的系统的信息越多，系统成功的机会就越大。

构成每个基于学习的分类器系统的基本组件是：

•    分类器使用的一组特征

•    训练数据集用于分类器训练

•    分类算法及其参数。

因此，为了对手 的利益，最大化他们对目标分类器组件的知识。例如，对手一个谁知道某个分类的功能集和训练数据集有逃避它比对手的机会较高乙谁只知道功能集。从这个意义上说，两个对手A和B在不同的逃避场景下运作。逃避场景是从对手的角度来看逃避的问题。它以结构化的方式描述对手可用的分类器系统信息：概述对手对于特征集合，训练数据集合和分类算法的知识是低还是高。

为了系统地探索针对分类器系统的逃避攻击，我们基于知识分子拥有的关于分类器系统三个组成部分的知识量，提出了图1所示的逃避场景分类。

我们的分类包含8个逃避场景。他们的名字描述了对手可用的信息。如果对应于分类器组件特征集的字母F，T或C中的任何一个分别存在于场景的名称中，则训练数据集和分类器算法分别存在于对手所具有的给定分类器组件的知识水平这种情况很高，否则低

名为O的情景指的是敌手对所有三个分类器组件都知之甚少的情况。

在我们的分类中，高知识并不一定意味着完整的知识，反之亦然。目前还没有一个严格的标准来决定某个分类器组件的知识水平应该被分类为高还是低。我们认为知识水平高，如果它可以用于对手的实质性优势，否则低。

我们的研究仅限于4个特征集合的知识水平高的逃避场景。如果没有特征的知识，攻击者面临着一个重大的挑战，要么从分类结果的观察中推断它们，要么直接测量分类器对原始数据变化的敏感性。目前，我们还没有意识到解决这些问题的技术，因此将对未来工作的特征知识进行调查。

在下面的小节中，我们描述了在4个感兴趣的场景中分级躲避攻击的高级算法。

*A.情景F*

在情况F中，只有特征集可用于敌手，在不同程度上。攻击者可能会意识到某些或全部特征，错误地将过时的特征视为正在使用，能够读取子集或全部特征或者能够以不同程度修改某些或全部特征。然而，为了能够修改样本并进行规避，需要操纵足够的特征子集。

没有关于分类器和训练数据集的知识的对手可能仍然执行规避。如果他能够访问数据样本，并且被目标分类器证明是良性的，他可以尝试使他的恶意示例与已知的良好示例一致。这种策略被称为*模仿攻击*。这个PDF RATE 攻击的特定实现在第V-C1节中介绍。一般来说，如果攻击者可以在攻击过程中向目标系统提交探测，模仿攻击是最有效的，以确保对模仿攻击源例子的良性分类或者在多个良性源中进行选择。然而，目标系统的在线探测可能是可检测的，因此不如完全离线的攻击（其中只有最终结果被提交给目标系统）不可取。

收集足够数量的恶意样本（例如在黑市上发现的）的对手可以将它们与一系列良性样本相结合，从而构建*替代数据集*。这个数据集可以用来训练一个现成的，*代理分类器*，然后可以使用为这个特定分类器调整的特殊目的攻击来躲避。代理分类器攻击背后的基本原理是预测模型的推理是基于一般统计特性，这是在许多学习方法中共享的。因此，很可能用一个合适的代理分类器来近似一个未知的分类器，这个分类器的行为可以被攻击者控制。这个策略的有效性主要取决于攻击者可用的数据质量。如果代理数据是训练数据真实分布的真实样本，则可以预期所得到的攻击对于原始未知目标分类器是有效的。基于代理分类器的攻击可以离线执行，

*B.情景FT*

这种情况下，除了已知的特征之外，攻击者还可以利用目标分类器训练数据集的知识。数据集可能完全或部分泄漏，从而在生成成功的攻击样本的过程中实现更准确的决策。

对良性训练点的了解使攻击者能够使用模仿攻击来生成与其非常相似的逃逸样本，从而与场景F相比增加了成功攻击的可能性。在*原始*数据集上训练一个有效的分类器会创建一个替代分类器，更好地接近目标分类器，而不是在场景F中训练，再次开辟了使用定制方法逃避替代分类器的方式。对训练数据的了解使得攻击者可以在提交最终结果之前离线执行整个攻击。

*C.情景FC*

在场景FC中，攻击者知道特征集以及关于分类器的一些细节，例如类型，参数或具体实现。根本没有关于训练数据集的信息而没有替代数据集的对手几乎不知道分类器的优点。使用替代数据集，他们可以训练正确类型的替代分类器，但是这种近似的准确性取决于收集的数据的质量。这种攻击也可以离线执行，类似于基于替代分类器的其他攻击。

*D.情景FTC*

如果知道所有三个分类器组件的细节，那么对手就有最大的机会躲避目标分类器。在这种情况下，他可以在线下完全重现在线分类器，只有在找到足够好的规避样本时才提交攻击结果。离线模仿攻击或离线分类器特定攻击击败离线分类器也具有很强的击败在线分类器的可能性。

在介绍目标系统PDF RATE 和用于实现上述一般攻击场景的具体算法之前，我们将在下一节中简要介绍可移植文档格式（PDF）。

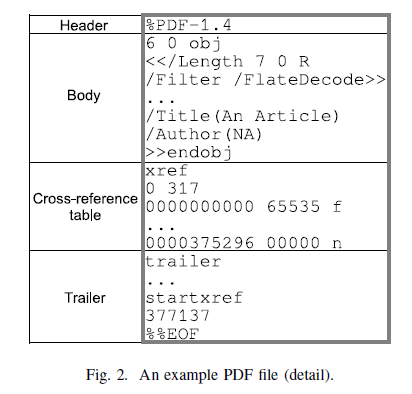


图2.示例PDF文件（详细信息）

III。Ť HE P ORTABLE d OCUMENT ˚F ORMAT

该*可移植文档格式（PDF）*是一种文件格式，能够使处于和打印一致，独立于底层环境的文件创建和发布为一个开放的标准，ISO 32000-1：2008 [30]。一个PDF文件，如图2所示，由一个带有PDF幻数和格式版本的*标题*，*正文*，一组构成文件结构和内容的PDF对象，*交叉引用表*（CRT）索引正文和*预告片中*的对象指着CRT。标题，CRT和预告片的开始分别由关键字％PDF，外部参照和预告片表示。文件主体中的对象与关键字obj和一对整数标识符一起引入。对象可以是不同的类型，例如*数字*， *字符串*，*名称*（标识符），*字典*（键值对的集合，其中键总是名称对象，值可以是任何对象类型，包括另一个字典），*间接引用*（“指针“到其他对象），*流*（带有附加编码和/或压缩内容的字典）等等.ISO 32000-1规定了一些字典具有特殊的含义，例如Type为JavaScript（包含可执行的JavaScript代码），Metadata（包含诸如其作者，标题，创建和修改日期等）或页面（描述单个页面）。被称为*对象流的*特殊类型的流可以包含其他对象作为编码和/或压缩的流内容的一部分。

标题，主体，CRT和尾标构成PDF文件的*文件结构*，即文件字节的内容，可以通过PDF格式的软件直接读取。文件体中的对象通过间接链接到其他对象或其直接嵌入，形成一个类似于图形的逻辑结构，称为PDF *文档结构*。*内容流*描述页面的外观。但是，我们限制对文件结构级别的兴趣，因为PDF RATE 不分析PDF文件，即只读取其原始字节。

IV . PDFRATE

PDF RATE 采用随机森林算法，根据其元数据和某些结构特征将PDF文件分类为良性或恶意[[3]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn3" \o ")。以下小节提供PDF RATE 的特性，分类算法，数据集和对抗性考虑的概述; 进一步的细节可以从原始文件[23]中获得。

*A.特点*

PDF RATE 总共使用了202个整数，浮点和布尔特性。在[24]中描述了其中的135个，其余的未知。在图4的VB部分中显示了一个特征子集的一个特定文件的值。这些功能反映了各种属性，如文件的大小和版本，PDF元数据项目的字符数（如作者姓名，创建和修改日期），结构属性（例如Acrobat表单的数量及其在文件中的相对位置等）。特征是由作者手动定义的，并分别针对对手的最佳分类性能和鲁棒性进行选择。通过对原始字节运行一组正则表达式来提取要素

PDF文件。通过不执行适当的PDF解析，作者

PDF RATE 有意识地偏好速度和简单性，而不是完整性和正确性，因为某些特征可能位于编码和/或压缩对象流中，超出了正则表达式的范围。

这些特征表现出显着的相互依赖性。当一个特征的值被修改时，许多其他特征可能会受到影响，因为它们直接或间接地依赖于目标特征。例如，通过修改作者元数据字段（author\_lc）的小写字母数量，相关特征author\_len将受到影响，但是与文件大小（大小）等直接相关的特征也会受到影响。大小的改变触发看似完全不相关的特征pos\_acroform\_ \*的进一步改变，表示一个或多个关键字AcroForm的相对文件偏移量。特征相互依赖性使特征值的对抗难以控制。

*B.数据集*

有三个数据集参与了PDF RATE 的创建和评估。已经对三种模型进行了培训，分别使用这些模型来评估用户提交的新数据。

三个数据集中的两个用于[23]中的PDF RATE 实验性评估：*Contagio*和*Operational*。该*Contagio*数据集贡献的恶意软件研究的恶意和良性PDF文件的集合，可供下载[[4]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn4" \o ")。培训PDF RATE 是在*Contagio*数据集的子样本上进行的，该样本包含5,000个良性文件和5,000个恶意文件[[5]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn5" \o ")。经过训练的分类器在*操作*数据库上进行了评估，包括在“大型大学校园”收集的100,000个PDF文件。据推测，相同的数据集被用来训练目前可用的模型*乔治梅森大学（GMU）*使用的PDF 率。

最后一个数据集“ *社区* ”是由PDF RATE 用户提交和评估的文件创建的，未在其原始评估中使用。

*C.分类算法*

PDF RATE 采用随机森林[31]，一种集合学习方法，包括独立训练的决策树的数量*t RF*。在训练步骤中，使用CART方法学习每棵树，但只使用可用训练样本的一个子集。通过从训练数据中随机抽样一定数量的树来为每棵树生成一个不同的子集，用替换 - 称为引导聚合或装袋的过程。当新的决策节点被添加到树中时，只考虑*f个RF*特征的随机选择的子集，其中*f RF*小于功能的总数。通过在给定的新数据点上的所有决策树中进行多数投票作出决定。随机森林以其卓越的泛化能力和对数据噪声的稳健性而闻名。PDF RATE 使用Leo Breiman's和Adele Cutler的原始随机森林实现的R端口，作为软件包RANDOM F OREST [[6]提供](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn6" \o ")。*t RF*和*f RF*分别是称为*ntree*和*mtry*的RANDOM F OREST的 参数。值*ntree* = 1000 和*mtry* = 43由PDF RATE使用。

所有这三个分类的PDF部署率，即接受了有关的那些*Contagio*， *GMU*和*社区*数据集，生产他们的结果是他们的决策函数在区间输出，即真正的价值[0 *，* 1] 表示的百分比将提交的文件标记为恶意的决定。在[23]中没有给出阈值来确定一个文件应该被认为是恶意的。请注意，通过提供此百分比值而不是二元决策PDF RATE揭示了更多关于分类引擎的信息，而不是决策所必需的信息，从而使对手能够在制定逃避方法时作出更明智的决定。

*D.敌对的考虑*

在描述我们的攻击之前，我们讨论PDF RATE 的特性对于我们研究的对抗性设置是至关重要的。

在我们的评估中，我们只关心*Contagio*分类器的逃避。我们不考虑*GMU*和*社区*分类器，因为他们的训练数据集对我们来说是不可用的，因此我们无法评估第II节中定义的全部范围的攻击场景。除了免费提供外，*Contagio*数据集似乎保持不变。定期再培训是一项重要的安全措施，会使我们的逃避方法的有效性的一致性评估变得复杂，因为每一个分级更新都会使以前的结果过时。此外，虽然PDF 率通过将恶意文件分类为“有针对性”和“机会主义”来提供第二级分析，但我们的研究仅限于回避最初的二进制分类器。

从对抗的角度来看，攻击者可获得的关于PDF RATE 的知识水平很高。其特征定义的可用性有助于创建操纵的样本。虽然对数据噪声有强大的抵御能力，但随机森林分类器并不是针对对抗噪音的弹性而设计的。在部署的系统中也不进行定期再训练。这些弱点使PDF RATE 成为我们案例研究的绝佳候选人。其他更突出的基于机器学习的恶意软件检测器具有未知或难以控制的特征。

尽管存在弱点，PDF RATE 确实存在对抗性的考虑初始设计。在文献[23]中考虑的攻击模型假设对于良性训练文件，攻击者知道6个最重要的特征（即，在随机森林模型的可变重要度量的列表之上）的手段和标准偏差。对手模型假定攻击者可以创建伪装的恶意样本，其中顶部特征的子集被设置为从正态分布中抽取的具有给定平均值和表征良性样本的标准偏差的随机值。这种攻击将被称为“良性随机噪声”（BRN）。文献[23]表明，BRN攻击会严重降低分类器的检测精度。为了对付这种攻击，提出了一个积极的防御战略：以完全相同的方式修改训练集中恶意数据点的子集。这个简单的防御策略被证明是非常有效的。

BRN攻击是综合实现的，即*在特征空间直接*修改前6个特征。因此，它没有解决是否可以用所需特征向量生成真实PDF文件的问题。由于强烈的相互依存关系，这种假设在实践中是不现实的。在我们对第六部分提出的防御机制的评估中，我们偏离了特征空间，并使用真实的PDF文件来评估这种攻击。此外，我们调查了对我们自己的模仿攻击提出的对策的稳健性。

五，METHODOLOGY 方法

由于我们对逃避场景的研究假设了一个隐身的攻击者，我们方法的关键要素包括重新实现PDF RATE 部署的方法。我们首先使用可用的公共知识重构PDF RATE 的一个子集。下一步是开发一种操作PDF文件的技术，这会影响选定的特征子集。我们方法的最后一步是设计攻击算法来实现第二部分介绍的通用攻击策略。

上述技术和方法在我们的实验性逃避框架M IMICUS中实施。该框架由支持特征提取，PDF文件修改，上传到PDF RATE 和分数检索，分类器训练以及对它们执行攻击的Python模块组成。M IMICUS 是免费的开源软件，适合与其他攻击和攻击目标进行扩展。它可以下载[[7]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn7" \o ")，与所有训练数据（作为特征向量），分类器模型以及完全重现我们的实验结果所需的代码捆绑在一起。我们实验中使用的所有攻击文件都可以从Contagio数据库中获得。

*A.重写*PDF RATE *的特性*

我们的四个逃避方案有一个共同的假设：攻击者知道被攻击的系统的特征。然而，关于特定特征的知识水平可能差异很大。攻击者可能根本不知道某些特征的存在。即使是具有已知描述的特征，攻击者也可能部分或不能控制其值。最后，功能之间的相互依赖可以防止攻击者对其值的任意操纵。

有关PDF RATE 特性的知识来自三个来源：原始研究论文[23]，技术报告[24]以及在线部署PDF RATE 的行为。

如[23]所述，共有202个特征被使用

PDF 率。然而，[24]中只描述了其中的135个，程度各不相同。这将潜在受攻击者控制的特征集合限制为大约报告数量的三分之二。此外，不能排除由于发展的自然进展，部署的系统与报告的系统相比没有不同的特征。

作为第一步，我们通过遵循[24]中的特征提取的通用指南，重新实现了135个已知特征的提取。随后，为每个特征开发了正则表达式，除了直接读取的大小。在此过程中，我们还检查了由PDF RATE 产生的元数据输出，其中包含PDF文件的固定测试套件，其中包含许多功能的各种值。可以从元数据输出中准确地推导出一些特征的值，例如，Page或obj关键字的计数。正则表达式进一步细化，直到所有测试文件达到一致的行为。尽管重新实施过程需要费时费力的专家工作，但对于激励对手而言，这将是一个小小的障碍。

感谢*Contagio*数据集的可用性，通过比较我们在该数据集上的分类结果与[23]中报告的结果，我们能够验证重新实现的正确性。此外，我们还证实，尽管实施特征的集合存在差异，但是我们的PDF RATE 本地克隆与第VI-A3节中提供的基准数据集中的在线系统产生类似的分类分数。

*B.修改*PDF 速率*特征值*

我们研究的PDF文件修改方法的发展受以下设计目标的指导：一旦修改，所述文件必须与原始文件看起来与任何PDF分析器不可区分，但可靠地影响PDF RATE 的特征提取。这样做的原因是，这样的语义保留方法可以安全地应用于我们的实验中的恶意PDF文件，而不管它们可能利用的各种漏洞，而没有破坏其可能的微妙*操作方式*的风险。

M IMICUS 的特征修改组件可以任意修改35的值和33的增量值

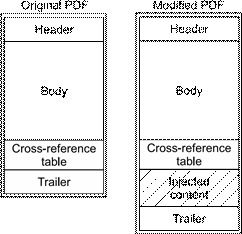


图3. PDF修改方法取原始PDF（左）并在交叉引用表（CRT）和预告片之间插入新内容。这样一个修改后的文件（右）将PDF RATE 混淆为接受新注入的内容作为文件的一部分，而PDF阅读器从拖车直接跳转到CRT，完全跳过注入的内容。

PDF RATE的特性，详见附录A.对其他特性的修改需要对PDF文件的结构进行微妙的修改，增加了实现工作量和破坏恶意功能的风险。

我们的文件修改方法是由PDF阅读器的操作和PDF RATE 之间的差异所驱动的。在[32]中将这种方法描述为解释各种文件格式时存在*语义鸿沟的*一个例子。PDF RATE 在PDF文件的原始字节上评估一组正则表达式，*从*文件*的开头到结尾*读取。相反，PDF阅读器按照ISO 32000-1规定的PDF语法解析PDF文件。一个符合标准的PDF阅读器从其结尾开始读取一个文件。它检查拖车找到交叉引用表（CRT）的位置，然后*跳转*直接找到它以便找到文件正文中的对象。这个差异如图3所示，显示了我们修改之前和之后的PDF文件的布局。

我们的解决方案利用此语义间隙：只要该文件标题，正文和CRT不被修改或移动时，拖车可以任意移动远离CRT 8，从而生成在该文件中的空的空间，其中任意的内容可以注射。这些内容将被PDF RATE 处理，但PDF阅读器将*始终忽略它*。

所描述的*内容注入方法*留下了文件修改，如果知道要查找什么，文件修改是微不足道的。我们认为可以重写PDF文件，而是在原地修改内容，从而将修改全部隐藏起来。但是，这种方法不会没有技术上的复杂性。它可能通过中断PDF文件的渲染而影响PDF阅读器，并可能对嵌入式漏洞的可靠性产生负面影响 - 这是内容注入完全避免的问题。我们的修改方法是通过将一组空格分隔的字符串模式注入到8之间的间隙中来实现的。只有拖尾保持在文件的末尾才是重要的。

CRT和目标PDF文件的预告片。这些模式旨在使特定的PDF RATE 正则表达式匹配它们，从而影响提取的特征值。例如，注入一个带有5个obj关键字的文件中，字符串“obj obj”将把它的count\_obj特征值从5改为7，因为PDF RATE 的正则表达式将全部匹配它们。作为另一个例子，作者元数据字段的长度可以通过注入一个新的“作者”字段，由3个字符“/ Author（abc）”作为PDF，RATE 倾向于只考虑文件中的最后一个元数据字段。通过在拖车之前注入我们的有效载荷，我们可以确保这个条件得到满足。

使用描述的修改方法，可以安全地假定PDF阅读器的行为不会被改变[[8]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn8" \o ")，但PDF RATE 会被诱骗从修改的文件中读取所需的特征值。我们的实验已经证实了两种PDF阅读器A DOBE R EADER 和E VINCE的这种行为。另外，我们已经将所有涉及到我们的逃避实验的恶意文件提交给W EPAWET [3]，并且验证了对任何样本的攻击有效性都没有受到影响。

如前所述，PDF RATE 的特点是相互依存的，即通常不可能将特征空间中的数据点完美地转换成问题空间中的文件。给定攻击前的恶意文件*F B*和攻击算法生成的数据点*P A*在*F B*上运行，攻击者想要生成攻击文件*F A*，通过修改*F B*的特征值来最优地击败分类器以*P 一*。但是，由于特征的相互依赖性，结果文件*FA*0 6 = *F A*，具有不同的特征，这可能会或可能不会击败分类器。图4示出了使用在部分V-C2中描述的G D -K DE 攻击的具体示例。特征pos\_acroform\_min表示关键字AcroForm的第一次出现的相对文件偏移量，不可由M IMICUS 修改，但是*间接地*受总文件大小增加的影响。另一方面，虽然特征author\_len表示作者元数据字段的长度是可以直接修改的，但是由于其他可修改的特征，即author\_lc，author\_num，author\_oth和author\_uc，表示不同的字符，所以其值为11而不是所需的0作者字段中的类，将总字符数驱动为11。

将数据点从特征空间转换到问题空间的另一个重要考虑是在特征空间中操作的算法可以构造在问题空间中不可行的数据点。示例是攻击算法试图分配负值的大小和版本功能。枚举所有的特征相互依赖关系并且*事先*说明它们的效果既*不可能*，也*不可能*在翻译成问题空间之前识别无效的数据点。

我们处理这两个限制的方法是机会主义的：我们从特征向量生成文件

                                       在文件        之后的特征

                            author\_lc：                  0                2                2

                          author\_len：                  0                0              11

                         author\_num：                 0                3                3

                          author\_oth：                 0                5                5

                            author\_uc：                 0                1                1

                  count\_acroform：                  1                0                1

                      count\_endobj：               11            918            465

                count\_endstream：                  1            169              85

                            count\_eof：                 0                2                2

                          count\_font：                  0              86              86

             count\_image\_large：                   0                1                1

             count\_image\_small：                   0                6                6

             count\_image\_total：                   0                0              11

           count\_image\_xsmall：                   0                4                4

                count\_javascript：                    3                0                3

                            count\_obj：               14            922            922

                      count\_objstm：                 0              28              28

                          count\_page：                 0              29              29

                      count\_stream：                  1            169              85

                      count\_trailer：                   1                0                1

                          count\_xref：                  1                0                1

                     createdate\_ts：                      -1 7.52e + 8 7.52e + 8

                     image\_totalpx ：                  0                  0 813898

                          moddate\_ts：                    -1 1.0e + 09 1.0e + 09

pos\_acroform\_avg：0.07043 0.07043 0.00716 pos\_acroform\_min：0.07043 0.07043 0.00716 pos\_acroform\_min：0.07043 0.07043 0.00716 size：        2726 -426760       26782 version：     0        -4       0

图4.           示例中特征子集的特征值变化

在方案F中的G D -K DE 攻击*。BEFORE*列显示从恶意候选文件*F B中*提取的具有SHA-1散列a39cf14b806db14a9e877b665324d203e5a5a666 的特征值。G D -K DE 将特征空间中的这些值转换成数据点*P A*（*AFTER*）。点*P 一个*用来修改文件*˚F 乙*的文件空间，并生成攻击文件*˚F 一个*。然而，功能相互依赖导致文件*FA*0而是用稍微不同的特征值（*FILE*）来生成。

通过一个接一个地转换特征，彼此独立并且不考虑局限性，希望得到的文件的特征与期望的值不会*太远*。虽然这种方法导致难以预测的结果，但是所得到的文件具有足够接近期望值的特征值并且适合于逃避。作为具体示例，比较图4的*AFTER*（期望结果）和*FILE*（实际结果）列。

如果所需值超出了特征和文件的有效范围，例如，如果尝试将大小修改为小于已有文件的正值，则M IMICUS中实现的其他安全机制会阻止特征修改。我们的方法实施的具体的下限和上限是通过枚举在VI-A节中描述的对手可用的PDF文件的数据集中的所有文件的特征并且提取每个特征的最小值和最大值来收集的。阻止特征修改的另一个原因是特征数据类型不匹配，例如，当数据类型不可知的算法想要将布尔特征设置为7.最后，结果是具有接近期望特征的有效PDF文件，适合逃避。

*C.攻击算法*

M IMICUS 的第二个主要组件是它的攻击算法。他们的主要目标是生成特征向量可能分类得分较低的PDF文件。为此，我们已经调整了两种先前已知的方法来处理PDF RATE 特性的特定上下文。

*1）模仿攻击：*模仿攻击在安全文献中是众所周知的。它的想法是转换一个恶意样本，以便尽可能模仿所选择的良性样本，使得最终的模拟样本更难以检测到。这种攻击很容易实现，可以应用于任何分类算法，并不一定依赖于特定的学习分类器模型。因此，适用于每个逃避场景的评估。我们的实现需要一个恶意文件，并且试图一次修改所有可修改的特性，以获取选择的模仿目标（一个良性文件）的特征值。为了提高模仿攻击的效果，我们用每个攻击文件的不同良性目标重复30次。生成的30个文件使用本地分类器进行评估，费率。

由于VB中讨论的特征和技术限制的存在，不可能产生一个完全对应于由模仿攻击产生的特征向量的文件。在特征空间中模仿完成之后，将特征向量转换为文件是很重要的。后者在技术上是直截了当的：我们只是将恶意特征向量合并到一个选定的良性向量中，同时保护现有的值。一次一个地修改功能，而将它们翻译成文件并不是一个好的策略，因为功能之间的相互依赖性支配着这个转换，并且产生了许多无法控制的变化。使用单步转换使得这种相互依赖性不那么突出。

模仿攻击的普遍性，即其具体学习算法和基础数据集的独立性，使其适用于其他基于学习的系统，如PDF RATE ，具有已知和可修改的特征子集。

反恶意攻击是通过将恶意内容注入到一个良性的PDF文件中进行的[33]，在小规模实验中被证明对PDF RATE 有效。

*2）梯度下降和核密度估计（* G D K DE *）攻击：*针对PDF RATE 评估的第二次攻击是基于使用梯度下降和核密度估计的方法（因此在本文中我们称之为G D -K DE）用已知的，可区分的决策函数来击败分类器[22]。它需要知道一个特定的学习模型和一套良性样本。此外，由于它是基于梯度下降的，因此它只适用于可微分类器，如SVM，**人工神经网络**等，不能应用于随机森林分类器。因此，G D -K DE 攻击仅适用于可区分替代分类器（F和FT）的情况。

G D -K DE算法通过遵循分类器的决策函数和良好示例的估计密度函数的加权和的梯度来进行。梯度下降的起点是恶意样本的特征向量。起始样本通常被正确分类为恶意; 目标是移动到分类算法将点分类为良性的区域。为了避免移向具有负分类特征空间的不可行区域，算法的目标函数具有第二项，即良性范例的密度。这确保了最终的结果靠近真正良性例子所在的区域。密度函数必须事先估计，使用标准的核密度估计技术[34]。与模仿攻击类似，我们运行GD K DE 在将特征空间转换为文件之前完成。

VI。Ë XPERIMENTAL Ë 估价

本节将介绍的实验评估迄今为止提出的规避技术的有效性。在我们的评估协议中，我们承担了攻击者的角色，并结合了所有可用的方法来部署PDF RATE 的最新版本。攻击者无法控制PDF RATE 的部署，因此不能保证系统在各个实验之间没有变化。由于我们的评估是针对在静态数据集上进行培训的模型进行的，发生在一周内，生产系统的任何变化都不太可能发生。

在另一组实验中，我们也调查了我们的攻击对原始文件中提出的防御措施的影响[23]。

*A.数据集*

在我们的实验中使用了三个数据集：两个数据集，*Contagio*和*Surrogate*，用于攻击实施所需的本地分类器的训练，而*攻击*数据集包含恶意文件，用作生成针对PDF RATE的攻击样本的起始点。

*1）*  *Contagio数据集：*这个数据集是原始PDFRATE 训练数据集的精确副本，在第IV-B节中描述。它包含5,000个良性和4,999个恶意PDF文件[[9]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn9" \o ")。假设对手知道这个数据集被用于训练并获得访问权是合理的。

*2）*  *代理数据*集（*Surrogate dataset）：*这个数据集被设计成一个数据集，不需要访问*Contagio*数据的对手就可以收集到这个数据集来近似它。数据集中的恶意文件是2012年3月5日至22日期间上传到在线病毒扫描服务VirusTotal[[10]](file:///E:\bdfirst\2018\2018Q1\srndic-laskov-sp2014.htm#_ftn10)的PDF文件的随机子样本。这些文件比*Contagio*数据新，但是在PDFRATE发布之前已知。数据集中的四个文件被发现存在于*Contagio*数据集中，并被删除以确保数据的严格互补。良性文件在*代理*数据集从2007年2月5日至2012年7月25日期间使用无关键字的Google网页搜索获取的文件中随机进行二次抽样*。Surrogate*数据集具有与*Contagio*数据集相同的大小和组成。

*3）*  *攻击数据集：*这个数据集包含100个恶意文件，被用作所有攻击的起点。为了最大限度地减少对PDFRATE的操作影响，故意选择这个数据集很小。对手可以在所有情况下访问这些文件。这些文件是从分类器中已知的*Contagio*数据集中随机抽取的，因此对攻击者来说，逃避更具挑战性。PDFRATE和WEPAWET都将它们全部分*Surrogate*类为恶意的PDFRATE，得分非常高，如图5“基线”所示。所有文件在问题和特征空间上都是不同的。

*B.分类器*

根据攻击者是否知道PDF RATE 使用的确切分类算法，他可能使用原始或不相关的替代分类器。我们的实验框架MIMICUS 通过在分类器类型已知的场景FC和FTC以及场景F和FT中的支持向量机（SVM）分类器中部署随机森林分类器来对这两种情况进行建模。

*1）*  *随机森林：*分类器的实现和参数与第IV-C节中描述的PDFRATE的原始分类器相同。

*2）*  *支持向量机：*我们选择了*支持向量机* [35]作为替代分类器，因为它在许多问题上具有很高的分类性能，包括使用PDFRATE 特征区分恶意和良性PDF文件，而且与随机森林无关。使用了SCIKIT-LEARN [36]机器学习工具包版本0.13.1的SVC实现。

SVM通过将标记的训练点映射到高维或无限维特征空间来学习，可选地将称为*核函数*的非线性变换应用于输入特征向量，以使它们更好地可分离。然后在新空间中找到具有最大可能边界的分离超平面，即两类点的凸包之间的距离，恶意和良性。由输入点的子集表示的超平面矢量，所谓的*支持矢量*以及它们的权重构成SVM模型。在分类新数据时，计算超平面与新数据点之间的距离。这个距离，称为*决定功能*分数，是一个真正的值，其含义与PDF RATE 的分类得分相似。通过为新数据点取决策函数得分的符号，可以做出二元决策。我们按照惯例将正分数分配给恶意分类。为了逃避SVM，攻击者需要修改恶意数据点，以使其决策功能得分改变。

两个核函数被评估：线性和RBF。线性核*k 线性*（*x*1*，x*2）= *x*1 · *x*2提供了两个输入向量*x*1和*x*2的线性变换，而

RBF核*ķ RBF*（*X*1*，X*2）= *EXP*（ - *γ* || *X*1- *X*2 || 2）利用高斯径向基函数作为其自变量的非线性变换。为了优化逃避结果，对两个对手的数据集进行网格搜索，优化线性的SVM参数*C*，针对RBF的*C*和*γ*

表红外线消除了一个TTACK S CENARIOS

|  |  |  |  |
| --- | --- | --- | --- |
| 脚本 | 分类 | 数据集 | 攻击（S） |
| F | SVM | *代孕* | 模仿，G D -K DE |
| FC | 随机森林 | *代孕* | 模仿 |
| FT | SVM | *Contagio* | 模仿，G D -K DE |
| FTC | 随机森林 | *Contagio* | 模仿 |

核心。使用10倍交叉验证和60％：40％训练测试分组的最高准确率在*Contagio*为98.7 ％，在*Surrogate*数据集为99.5％。

在网格搜索中找到的参数对于两个数据集都是相同的：RBF内核，*C* = 10 ，*γ* = 0 *。*01 。然而，这两种数据集学习模型的泛化能力差别很大。受过训练的SVM *Contagio*数据实现对98.5％的准确度*代用*数据，但训练上的SVM *代用*数据执行很差上*Contagio*数据，具有61％的精确度。当然，场景F和FC中的对手无法获得原始训练数据，将无法检查SVM在其上执行的效果。训练数据的错误近似强烈地影响G D 的性能-K DE 攻击。

由于使用支持向量机时特征中并作为一般规则在尺度的差异，*特征标准化*已于提取的数据点中减去特征从特征值的平均值和由特征的标准偏差除以结果执行。在*Contagio*数据集上计算所有特征的平均值和标准偏差。

*C.攻击场景*

以下小节阐述了如何在我们的实验中使用可用的数据和算法实现第二节中描述的四种主要攻击情景。表1中给出了实现的攻击场景的总结。

*1）*  *场景F：*在场景F中，攻击者知道的关于PDFRATE的 所有内容是如何读取135和修改68个特性。不过，根据第II-A部分详述的可用数据集，他可以执行两次攻击。模仿攻击使用从*Surrogate*数据集中随机采样的良性文件，因为*Contagio*数据集是未知的。类似地，使用GD-KDE对替代分类器进行训练以用于*避开Surrogate*数据集。分类器参数使用*Surrogate*数据集上的网格搜索进行优化。这两种攻击都是在离线状态下进行的，没有分类器反馈，结果上传到PDF速率进行评估。

*2）*  *场景FT：*在场景FT中，除了有限的特征知识之外，攻击者对训练数据集有完整的了解。因此，在这种情况下，使用*Contagio*数据集来训练GD-KDE 攻击的替代分类器，随机抽取来自*Contagio的*良性文件作为模仿攻击的模拟目标。这一次，使用*Contagio*数据集上的网格搜索对代理分类器进行了优化。只有最终的攻击结果才会提交给PDFRATE。

|  |
| --- |
| E:\bdfirst\2018\2018Q1\srndic-laskov-sp2014.files\image007.gif  图5. 来自*攻击*数据集的所有100个攻击样本的PDF RATE 分数（“基线”）和每次攻击之后（其余）的*总数*。攻击按场景分组。方框从第一到第三四分位数之间的中间值（粗线）。晶须延伸到离盒子四分之一距离1.5倍以内的最远的基准面，而正方形代表异常值。 |

*3）*  *场景FC：*在这种情况下，关于分类器的知识被添加到有限的特性知识中。对手知道原来的分类器，它的实现和参数。他们使用带有原始分类器的*代理*数据集来生成替代分类器，他们使用模仿攻击逃避脱机，从*代理*数据集中随机选择模仿目标。结果将提交给PDF格式的RATE进行评估。

*4）*  *场景FTC：*由于对训练数据集和分类器的特征和完整知识的了解有限，攻击者创建PDFRATE 的本地副本，并将其离线。只有最后的攻击结果在线提交。

*D.结果*

在攻击实验运行之前，*Attack*数据集中的所有100个文件都通过PDF RATE 评估。这个评估的结果，如图5所示，“基线”，为我们比较攻击结果提供了一个基准。除3个文件外，其他所有文件都获得了100％

我们的评估遵循一个简单的协议 对于每次攻击，来自*攻击*数据集的100个文件被用来生成攻击样本。通过将生成的攻击文件提交给PDF RATE ，并将接收到的分类评分与我们的基线进行比较，评估生成的攻击文件的有效性。所有的攻击样本都提交给W EPAWET，以验证修改后的恶意。

PDF RATE 对攻击文件的分数总结如图5所示。对于每次攻击，将100个分类评分的总体表示为一个盒子图，中间值表示为粗线，第25和第75个百分点“四分位数范围”或IQR）作为一个方框，从中位数的1.5 IQR到“胡须”，剩下的离群数为单个点。情节按攻击情景分组。

结果显示PDF RATE *在所有4种攻击情景中都*被避开。模拟的中位数分数下降到28-42％，G D -K DE 攻击的中位数分数下降到29-34％，这取决于情况。对于情节F中除模仿以外的所有攻击，箱型图的第75个百分点位于50％标记之下，暗示

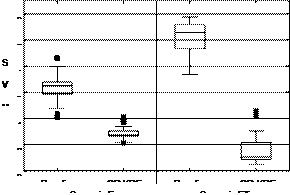


图       6.GD前后SVM决策函数值的集合D -

对于来自*攻击*数据集的所有100个攻击样本，K DE 在场景F和FT中进行*攻击*。箱形图的参数在图5中描述。

如果将50％的分数阈值用于决策，则75％的攻击将被分类为良性。这些结果的重要性得到进一步强调，只有三分之一的特征是可修改的，用于评估的文件在训练时已经被PDF RATE 知道，因此更难以回避。

在具有替代分类器F和FT的场景中的结果表明G D -K DE 优于模仿。此外，FTC情景下具有最高知识的情景下的模仿攻击仅略胜于G D K DE 。进一步的洞察行为的

G D -K DE 攻击如图6所示。它显示了两个支持向量机的决策函数的值，一个在场景F 的*代理*数据集上训练，另一个在情景FT 的*Contagio*数据集上，在攻击前后进行训练。攻击后的SVM分数表明，G D -K DE 攻击可靠地将*所有*远离决策边界的样本引导至良性区域。如果SVM分类器是通过PDF RATE 部署的，那么G D -K DE 攻击很可能已经达到完全规避，将所有分数驱动到零以下。自从

SVM只与一个随机森林的决策函数近似匹配，对PDF RATE的攻击远未达到完美，但仍然大大降低了得分。

通过仔细观察图6可以看出，场景F中超过25％的预攻击样本具有负判决函数值，即*在*攻击*之前*被SVM分类为良性（但不是PDF RATE ）。这是在情景F下操作的结果，其中对手训练使用*代理*数据集，但是使用来自*Contagio*数据集的样本进行攻击。由于在这种情况下支持向量机模型的推广性很差，正如第VI-B2部分所述，样本往往被错误分类。在场景FT中，攻击者也在*Contagio*数据集上训练，基准SVM分数严格为正值。

另一个基于图5的重要观察是随着对手对目标系统的了解的增加，攻击效率的提高。这一发现与我们对于敌手知识对分类器鲁棒性的重要性的初步猜想是一致的。然而，从F场景到FTC场景的改善并不像预期的那么戏剧化：奇怪的是，模拟效果提高了大约14％，场景F中的G D -K DE 比最好的整体攻击更胜一筹，在FTC场景中模仿，仅有6％。这是一个重要的发现，表明仅仅知道一个特征子集就可以为敌人提供比先前考虑的更多的优势。

拥有训练数据是对攻击者成功的第二大贡献，在了解了特征之后。图7比较了两个本地随机森林和PDF RATE 在场景FC中的模拟攻击，使用*Surrogate*和FTC，使用*Contagio*数据的得分。可以看出，在*代理*数据中，确切的分类器对攻击有效性进行了过分乐观的评估，实现了约18％的中值分数，而提交到PDFRATE 时相同的文件得到37％的中值分数。但是，在使用*Contagio*数据集时，攻击分数的本地估计与PDF RATE 几乎相同（分别为29％和28％）。考虑到局部分类器只是使用PDF RATE 特征的一个子集进行训练，并且随机森林的训练过程是严重随机化的，这种相似性是令人惊奇的。

作为评估的最后一步，我们调查攻击对PDF RATE 检测性能的影响。回想一下，分类评分仍然需要与某个阈值进行比较，以作出二元决策。早些时候，我们报告说，如果阈值设定在0.5的具体值，那么75％的攻击点就会落在雷达的下方。为了分析*所有可能的阈值*的检测性能*，*在图8中给出了接收机操作特性曲线（ROC），用于基线和所有的攻击。ROC曲线是从包含相同的100个攻击样本的混合数据样本和来自Contago数据库的所有1051个良性样本获得的，而不是在*Contagio*数据集。从图中可以清楚的看到，特别是在误报率较低的范围内（小于0.5％），PDF RATE 的检测性能由于攻击而显着降低，

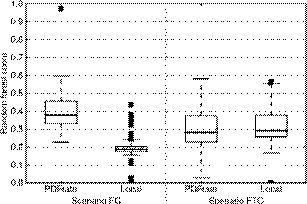


图7. 来自*攻击*数据集的所有100个攻击样本的随机森林分数由PDF RATE 和两个本地随机森林在两个模拟攻击中的分布情况。箱形图的参数在图5中描述。

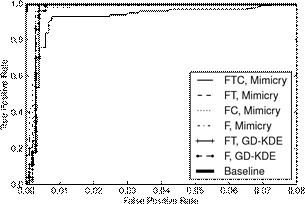


图8.基线和所有攻击的ROC曲线。

对FTC情景的模仿攻击造成了7％的误判率。攻击相对于检测性能的相对有效性与其在分类评分方面的相对有效性相似（参见图5）。

*E.防御措施*

在我们上一个实验中，我们已经研究了[23]中提出的防御机制对我们的逃避技术的鲁棒性。为了设定基线，我们使用在*Contagio*数据集上训练过的随机森林分类器，完全按照Smutz和Stavrou提出的模拟攻击和防御技术（参见第IV-D节）重现模仿攻击和防御技术。我们的分类器按照最重要的顺序排列了以下十个特征[[11]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn11" \o ")：

count\_font pos\_eof\_avg count\_endobj count\_js pos\_eof\_max producer\_len

count\_javascript len\_stream\_min pos\_box\_max count\_obj

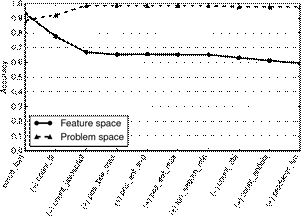


图9.在特征空间中的数据点上应用的BRN攻击的结果与问题空间中的文件的结果。攻击被应用了10次，从count\_font开始，逐步修改更多的功能。

然后，我们重新实现了Smutz和Stavrou在特征空间中的原始“良性随机噪声”（BRN）攻击，并通过为攻击的最终特征向量生成目标文件将其扩展到问题空间。这两个攻击变量的有效性的比较，作为修改的特征数量的函数，如图9所示。我们观察到，BRN攻击（实线）的合成变体的行为非常类似于[ 23]，对精度的影响稍高[[12]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn12" \o ")。但是，在应用于文件时，BRN攻击是无效的（虚线）。只有一个或两个特征的修改对检测精度有一定的影响。试图修改更多的功能会导致无意的修改，最终导致模仿样本走向良性阶层。而且，前10个特征中只有5个是可修改的。因此，在问题空间中的BRN攻击是不切实际的，与VI-D中提出的其他攻击相比，并不理想。

最后，我们评估了Smutz和Stavrou提出的“疫苗接种”机制的有效性，该机制修改了训练数据集中恶意样本的一部分，使其与*预期的*攻击样本更相似。两种情况被考虑：当防御者预期*（1）权利*和*（2）错误*一种攻击。在图10中比较了在两种情况下针对BRN攻击的接种防御的效力。我们的实验证实了当预期正确的攻击类型（即，BRN）时接种防御的有效性（虚线）。然而，接种BRN攻击的分类器对FTC情况下的模拟攻击（虚线曲线）没有任何抵触。用我们的模拟攻击重复接种疫苗的实验显示，抵抗攻击的能力得以恢复（图11）。因此可以得出结论，该疫苗接种的机制是对任何有效的*正确*

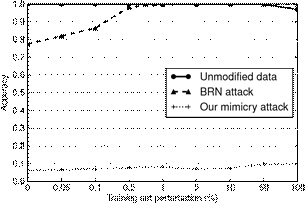


图10. [23]中提出的防御措施的性能。5次独立试验的平均结果，使用10倍交叉验证。

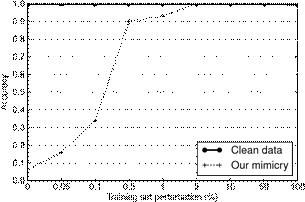


图11. [23]中提出的防御措施对我们的模仿攻击的影响。5个独立试验的平均结果。

*预期的*攻击。后一种假设在实践中是相当不切实际的。

七。我对一个攻击的认识

从操作的角度来看，了解哪些功能对所报告的攻击的成功起最大作用是至关重要的。一般来说，学习技术创造的模型的解释总是困难的。尽管随机森林分类器根据其信息性提供了对BRN攻击设计至关重要的特征排序，但这些信息只与本文提出的两种类型的攻击间接相关。因此，需要开发不同的分析技术来解释我们的攻击。

我们的解释是基于特征向量在攻击之前和之后的二进制差异。尽管声称具有最大变化的特征是最具信息性的可能是诱人的，但是由于特征值的范围大不相同，所以这种测量在我们的例子中是强烈的误导。即使将更改重新缩放到有效值范围也是不合适的，因为只能基于PDF文件的经验性样本确定特定要素的边界

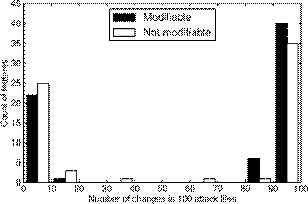


图12. 场景FT中100 G D -K DE 攻击发生变化的特征数量分布。

并容易出现异常值。此外，在我们的方法中，只有三分之一的特征是可以直接修改的，但所有这些特征都可能因其他一些变化而间接修改。

模仿攻击唯一可以想象的特征是具体特征的经验支持，即攻击改变了特定特征的文件的百分比。图12和图13分别显示了G D -K DE 和模拟攻击的特征支持直方图。可以看出，两种攻击都进行了大量的特征修改，因此无法用少量的基本特征来解释攻击。在两次攻击之间，由G D -K DE 产生的修改更统一，有几乎每次攻击都会改变的45个特征。剩下的23个特征很少被修改，很可能是由于相反的变化方向（回想一下，35个特征只能在我们的设置中增加），或者由于所请求的变化的不可行性。由模仿攻击影响的变化表现出较高的支持变异性。观察直接修改伴随着几乎平衡的间接修改量也是有趣的。这是PDF RATE 特性高度相互依赖的另一个例子。

解释攻击的一种实用方法是观察攻击产生的特征值的具体变化。虽然这个调查并没有扩大到有很多特征和档案的案件，但是这种调查能够深入了解当前袭击的*手法*。图4显示了一个特定文件的特征如何在G D -K DE 攻击中改变。回想一下，G D -K DE 通过将决策边界上的恶意数据点转向使用梯度下降的良性区域来操作，并且同时利用核密度估计来将它们推向所看到的良性样本。在所提供的例子中，这种“良性化”是显而易见的。通过比较*BEFORE*和*在*列之后，我们看到攻击已经添加了一个作者

（作者\*功能），将创建（createdatets）和修改（moddate\_ts）日期设置为最近的过去，减少JavaScript发生（count\_javascript），添加一些页面（count\_page），字体（count\_font），图像

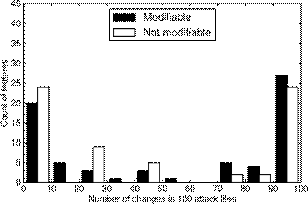


图13. FTC场景中100次模拟攻击发生变化的特征数的分布情况。

（count\_image \*）等 - 所有的变化，良性阶层。一些特征，例如大小和版本，被改变为无效值，可能是由于梯度下降分量的影响。

八。d 藏密坛城艺术和- [R 兴高采烈W¯¯ ORK

虽然我们的研究结果只适用于一个单一的系统，我们的研究结果表明几个重要的影响。这是对部署的基于学习的系统进行全面实践评估的第一次尝试，因此我们不能期望我们的结果能够被大量类似的系统完全重现。但是，我们的实验揭示的一些关键问题值得仔细考虑，因为它们指出了在设计未来数据驱动系统时需要解决的一些常见问题。

我们实验的主要信息是，如果攻击者对其特征和方法有足够的了解，攻击者可以大大降低基于学习的系统的准确性。导致这种不安全的主要因素是特征的知识。对于PDF 率我们已经观察到即使是来自我们的武器库的最简单的攻击，除了其特征之外没有对该系统的进一步知识，可以将所选代表性恶意样本集的恶意分数从100％降低到33％的中值。尽管大概三分之一的分类器特征完全不为我们所知，而另外三分之一的特征不能被我们的工具所修改，但这是可能的。这样的影响表明，即使是关于这些特征的少量知识也可以用于规避躲避攻击。诸如训练数据的知识和算法的精确类型之类的其他因素对于攻击是有帮助的，但不是关键的，因为这些信息可以通过替代来源很好地近似。

基于学习的方法不安全的根本问题在于特征的设计。机器学习在各种信息系统中越来越受欢迎 - 这远远超出了安全 - 主要是因为它能够预测，或多或少的成功，从*副作用的原因*。使机器成为这种概括性的能力

学习算法是对不确定性笼罩的问题寻找解决方案的选择方法，当人们对问题的理解不足以*设计*解决方案时，也不能从查看原始数据中找出解决方案。通过手动选择一组容易计算的副作用或“专家特征”来设计用于学习算法的特征的主流方法显然具有攻击者可能完全相同的危险。为了保护这些方法免于逃避，似乎除了隐藏关于特征的大量知识之外别无他法。否则，正如我们的研究所表明的那样，有一些原则性的方法可以自动扩展缺少的知识，从而成功地躲避攻击。

有没有其他解决方案可以使学习方法更具有规避性？一个可能的解决方案是使用至少在合理的程度上固有地表示*原因的特征*被检测。在以前关于shellcode检测和分类的工作中可以找到这样的特征的一个例子，例如[37,38,39,40]，其使用n-gram或短字节序列作为基本特征。近来已经探索了类似的方法来检测JavaScript恶意软件，同样的技术被应用于语法令牌序列[41,42]。这些方法的区分能力在于shellcode和普通包内容以及恶意JavaScript代码和良性程序之间固有的统计差异。因此，人们可以期望这种功能比“专家功能”更不容易被恶意操纵。如果事实证明，基于ngram特征的准确回避是NP完全的[20]，并且在实践中普遍认为近似解法是困难的。

另一个可能的解决方案可以通过试图在整个尽可能多的特征上统一传播“判别力”的方法来提供。最近有人提出了这种方法来学习潜在的特征删除或腐败问题[43,44]。由于训练问题的复杂性显着增加，这种方法提供了对有限量的特征噪声的合理保护，而不管特征类型如何。假设攻击者可以修改访问有限数量的特征，就像我们研究中的情况一样，当没有“内在特征”可以设计时，可以期望这样的方法为这些情况提供良好的折衷。

最后，应该提到基于多个分类器系统[45]的方法作为一个潜在的解决方案。逃避一些补充分类器可能比单个分类器困难得多。例如，我们无法使用一组异构决策函数来设计针对随机森林分类器的最佳攻击。最近已经考虑了多个分类器系统在安全和其他对抗场景中的一些应用[46,47]。

IX。C 结论和F UTURE W ORK

在本文中，我们已经提出了一个部署的学习型系统的第一次经验安全评估。我们的研究假设攻击者没有关于系统的具体内幕信息。但是，它表明，可以从各种来源收集足够的信息，并通过appriximations和自动推理算法进行扩展，以便实现成功的逃避攻击。在我们在已建立的用于检测PDF恶意软件的系统（PDF RATE）上进行的实验中，已经观察到分类评分显着下降（从几乎100％到28-33％）以及检测率的恶化。我们还观察到，只有在预期的攻击与所执行的攻击完全匹配的情况下，针对规避攻击的简单对策（例如在训练数据中包括一小部分攻击）才是有效的。

我们的研究结果表明，应该认真关注数据驱动的安全技术中使用的特征和算法的设计。我们今后的工作将试图解决在所提供的实验中发现的学习方法的局限性，特别是在理解构建攻击弹性特征的一般属性方面。我们的评估方法适用于其他具有可修改功能的基于学习的系统，我们打算扩展相关系统的安全评估方法以检测恶意JavaScript和PDF文件。通过在线发布我们的实验代码，我们希望与其他研究人员分享我们的经验，并促进实验结果在实际系统安全评估中的可重复性。

一个附录一个PDF RATE ˚F EATURE [R EIMPLEMENTATION

M IMICUS 实验框架支持读取[24]中描述的135 PDF Rate 特性（66％）。其余的202个功能中的67个没有透露。支持以下68个功能（33％）的值的修改：

•     其值只能递增的特征（33）：

count\_acroform count\_image\_xlarge count\_acroform\_obs count\_image\_xsmall count\_action count\_javascript count\_action\_obs count\_javascript\_obs count\_box\_a4 count\_js count\_box\_legal count\_js\_obs count\_box\_letter count\_obj count\_box\_other count\_objstm count\_box\_overlap count\_objstm\_obs count\_endobj count\_page count\_endstream count\_page\_obs count\_eof count\_startxref count\_font count\_stream count\_font\_obs count\_trailer count\_image\_large count\_xref count\_image\_med大小count\_image\_small

•     其值可以递增和递减的特征（35）：

author\_dot keywords\_dot subject\_dot author\_lc keywords\_lc subject\_lc author\_num keywords\_num subject\_num author\_oth keywords\_oth subject\_oth author\_uc keywords\_uc subject\_uc createdate\_ts moddate\_ts title\_dot createdatetz moddate\_tz title\_lc creator\_dot producer\_dot title\_num creator\_lc producer\_lc title\_oth creator\_num producer\_num title\_uc creator\_oth producer\_oth version creator\_uc producer\_uc

R EFERENCES

[1]      美国拜耳，P. Comparetti，C. Hlauschek，C. Kruegel和E. Kirda，“可扩展的，基于行为的恶意软件聚类，”在*网络和分布式系统安全研讨会（NDSS）*，2009年。

[2]      O. Thonnard，“用于支持网络攻击归因的多标准聚类方法”。巴黎电子通讯与电子学院，2010年。

[3]      M.Cova，C.Kruegel和G.Vigna，在2010年*WWW国际会议上*发表的“驱动下载攻击和恶意JavaScript代码的检测和分析” ，第281-290页。

[4]      MA Rajab，L. Ballard，N. Lutz，P. Mavrommatis和N. Provos，“ *网络和分布式系统安全研讨会（NDSS）* ”中的“CAMP：内容不可知的恶意软件保护” ，2013。

[5]      D.Canali，M.Cova，G.Vigna和C.Kruegel，“Prophiler：一种用于大规模恶意网页检测的快速过滤器”，在2011 *年万维网（WWW）国际会议*，第197-206页。

[6]      G.Stringhini，C.Kruegel和G.Vigna在2013 年*ACM计算机和通信安全会议（CCS）上的* “Shady paths：Leverage surfing crowds to detect malicious web pages”，pp.133-144。

[7]      M. Egele，G。Stringhini，C. Kruegel和G. Vigna，“COMPA：Detecting compromised accounts on social networks。”in *Network and Distributed System Security Symposium（NDSS）*，2013。

[8]      DM Freeman，“使用朴素贝叶斯在社交网络中检测垃圾邮件名字”，*ACM人工智能与安全研讨会（AISec）*，2013年，第3-12页。

[9]      B. Rahbarinia，R. Perdisci，A. Lanzi和K. Li，“Peerrush：挖掘不需要的P2P流量” ，2013年*入侵检测和恶意软件和脆弱性评估（DIMVA）*，第62-82页。

[10]   M. Barreno，B. Nelson，A. Joseph和J. Tygar，“机器学习的安全性”，*Machine Learning*，vol。81，没有。2，第121-148页，2010。

[11]   M. Kearns和M. Li，“在存在恶意错误的情况下进行学习”，*SIAM Journal* of *Computing*，vol。22，没有。4，pp。807-837,1993。

[12]   M.Bruckner，C.Kanzow和T.Scheffer，“静态预测游戏对抗学习问题” *，机器学习研究杂志，* 2617-2654页，2012。

[13]   M. Kloft和P. Laskov，“在线质心异常检测的安全性分析” *，机器学习研究杂志*，第一卷。13，pp.3133- 3176,2012。

[14]   B.Biggio，G.Fumera和F.Roli，“安全评估的模式分类攻击，” *IEEE知识和数据工程学报*，卷。99，没有。PrePrints，p。1，2013。

[15]   P.Fogla，M.Sharif，R.Perdisci，O.Kolesnikov和W.Lee，“Polymorphic blending attacks”，*USENIX Security Symposium*，2006，pp.241-256。

[16]   R. Perdisci，D. Dagon，W. Lee，P. Fogla和M. Sharif，“使用故意噪声注入的误导性蠕虫签名生成器”，*IEEE Symposium on Security and Privacy*，2006年，第17-31页。

[17]   B. Biggio，B.尼尔森和P. Laskov，在“对支持向量机中毒攻击，” *国际会议上机器学习*，2012。

[18]   Y. Lindell和B. Pinkas，“隐私保护数据挖掘”，在*密码学进展国际密码学会议（CRYPTO）*，2000年，第36-54页。

[19]   C. Dwork，“差别隐私：结果调查”，在*计算模型理论与应用国际会议（TAMC）*，2008年，第1-19页。

[20]   P. Fogla和W. Lee，“Evading network anomaly detection systems：formal reasoning and practical techniques，” *ACM Conference on Computer and Communications Security*，2006，pp.59-68。

[21]   D. Lowd和C. Meek在*“ACM SIGKDD国际会议论文集知识发现与数据挖掘* ”2005年第641-647页上发表的“良好的统计垃圾邮件过滤器的文字攻击” 。

[22]   B.Biggio，I.Corona，D.Maiorca，B。Nelson，N.S.Stndi C.，P.Laskov，G.Giacinto和F.Roli，“逃避攻击机器学习在测试时间”，*欧洲机器学习会议*，2013 *年数据库知识发现原则与实践*。[在线]。可用：[http](http://dx.doi.org/10.1007/978-3-642-40994-3_25) : [//dx.doi.org/10.1007/978-3-642-40994-3 25](http://dx.doi.org/10.1007/978-3-642-40994-3_25)

[23]   C. Smutz和A. Stavrou在“ *年度计算机安全应用会议（ACSAC）”*，2012年，第239-248页“使用元数据和结构特征进行恶意PDF检测” 。

[24]   - ，“使用元数据和结构特征进行恶意PDF检测” Malicious PDF detection using metadata and structural features，可在http://cs.gmu.edu，乔治梅森大学计算机科学系4400 University Drive MSN 4A5，Fairfax，VA 220304444 USA，技术。Rep。GMU-CS-TR-2012-5,2012。

[25]   D. Sculley，ME Otey，M. Pohl，B. Spitznagel，J. Hainsworth和Y. Zhou在*ACM SIGKDD国际知识发现和数据挖掘会议（KDD）上发表的* “在野外探测敌对广告” ，2011年，第274-282页。

[26]   B. Nelson，M. Barreno，FJ Chi，AD Joseph，BIP Rubinstein，U.Saini，C. Sutton，JD Tygar和K. Xia，“利用机器学习颠覆你的垃圾邮件过滤器”，*USENIX Workshop关于大规模利用和突发威胁（LEET）*，2008年，第1-9页。

[27]   C. Yang，RC Harkreader和G. Gu，“死亡或活得很难？经验性评估和针对不断演变的Twitter垃圾邮件发送者的新设计“，” *入侵检测最近的进展（RAID）“*，2011年，第318-337页。

[28]   M. Brennan，S. Afroz和R. Greenstadt，“对抗式测量：绕过作者身份识别以保护隐私和匿名”，*ACM信息系统安全报告*，第一卷。15，没有。3，第1-22页，2012年。

[29]   G. Oberreuter, G. L’Huillier, S. A. R´ıos, and J. D. Velasquez, “Outlier-´ based approaches for intrinsic and external plagiarism detection,” in *Knowledge-based and intelligent information and engineering systems*, 2011, pp. 11–20.

[30]   “Document management - Portable document format - Part 1: PDF 1.7,” http://www.adobe.com/devnet/pdf/pdf reference.html, 2008.

[31]   L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[32]   S. Jana and V. Shmatikov, “Abusing file processing in malware detectors for fun and profit,” in *IEEE Symposium on Security and Privacy*, 2012, pp. 80–94.

[33]   D. Maiorca, I. Corona, and G. Giacinto, “Looking at the bag is not enough to find the bomb: An evasion of structural methods for malicious pdf files detection,” in *SIGSAC Symposium on Information, Computer and Communications Security*, 2013, pp. 119–130.

[34]   E. Parzen, “On estimation of a probability density function and mode,” *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[35]   C. Cortes and V. Vapnik, “Support vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995.

[36]   F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[37]   K. Wang, J. Parekh, and S. Stolfo, “Anagram: A content anomaly detector resistant to mimicry attack,” in *Recent Adances in Intrusion Detection (RAID)*, 2006, pp. 226–248.

[38]   K. Rieck and P. Laskov, “Language models for detection of unknown attacks in network traffic,” *Journal in Computer Virology*, vol. 2, pp. 243–256, 2007.

[39]   J. Kolter and M. Maloof, “Learning to detect and classify malicious executables in the wild,” *Journal of Machine Learning Research*, 2006, to appear.

[40]   Z. Shafiq, S. Khayam, and M. Farooq, “Embedded malware detection using markov n-grams,” in *Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA)*, 2008, pp. 88–107.

[41]   K. Rieck, T. Kruger, and A. Dewald, “Cujo: Efficient detection and¨ prevention of drive-by-download attacks,” in *Annual Computer Security Applications Conference (ACSAC)*, 2010, pp. 31–39.

[42]   P. Laskov and N. Srndiˇ c, “Static detection of malicious JavaScript-´ bearing PDF documents,” in *Annual Computer Security Applications Conference (ACSAC)*, 2011, pp. 373–382.

[43]   A. Globerson and S. Roweis, “Nightmare at test time: Robust learning by feature deletion,” in *International Conference on Machine Learning (ICML)*, 2006, pp. 353–360.

[44]   O. Dekel, O. Shamir, and L. Xiao, “Learning to classify with missing and corrupted features,” *Machine Learning*, vol. 81, no. 2, pp. 149–178, 2010.

[45]   F. Roli, G. Giacinto, and G. Vernazza, “Methods for designing multiple classifier systems,” in *Multiple Classifier Systems*, 2001, pp. 78–87.

[46]   R. Perdisci, D. Ariu, P. Fogla, G. Giacinto, and W. Lee, “McPAD: A multiple classifier system for accurate payload-based anomaly detection,” *Computer Networks*, vol. 53, no. 6, pp. 864–881, 2009.

[47]   B. Biggio, G. Fumera, and F. Roli, “Multiple classifier systems for adversarial classification tasks.” in *Multiple Classifier Systems*, 2009, pp. 132–141.

[[1]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref1" \o ") [http://pdfrate.com/.](http://pdfrate.com/)

[[2]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref2" \o ") One of the training datasets used by PDFRATE is publicly available for the research community.

[[3]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref3" \o ") PDFRATE’s structural features describe physical rather than logical structure, and are not to be confused with the PDF document structure.

[[4]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref4" \o ") The Contagio archives are available at the following URL: [http:// contagiodump.blogspot.de/2010/08/malicious-documents-archive-for.html.](http://contagiodump.blogspot.de/2010/08/malicious-documents-archive-for.html)

[[5]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref5" \o ") A list of MD5 sums of those files was published: [http://pdfrate.com/ contagio](http://pdfrate.com/contagio_md5_class.csv) E:\bdfirst\2018\2018Q1\srndic-laskov-sp2014.files\image024.gif[md5](http://pdfrate.com/contagio_md5_class.csv) [class.csv.](http://pdfrate.com/contagio_md5_class.csv)

[[6]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref6" \o ") <http://cran.r-project.org/web/packages/randomForest/index.html>

[[7]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref7" \o ") MIMICUS – [https://github.com/srndic/mimicus.](https://github.com/srndic/mimicus)

[[8]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref8" \o ") Provided that they do not parse the injected content, but perform the direct jump prescribed by ISO 32000-1 instead.

[[9]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref9" \o ") We were unable to locate the malicious file with the MD5 hash 35b621f1065b7c6ebebacb9a785b6d69 in the archives.

[[10]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref10" \o ") VirusTotal – [https://www.virustotal.com/.](https://www.virustotal.com/)

[[11]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref11" \o ") We have used 10 instead of 6 top features for mimicry because they were ranked significantly above others.

[[12]](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftnref12" \o ") The accuracy scores might vary because of experiment randomization, different Random Forest models or different cutoff values.