加微信：642945106 发送“ 赠送 ”领取赠送精品课程
发数字“ 2 ”获取众筹列表

013 | 精读2017年KDD最佳应用数据科学论文

2017-11-01 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 09:34 大小 4.38M

周一我们讲了 2017 年 KDD 最佳研究论文，今天我们继续来聊今年的 KDD 最佳应用数据科学论文。

与研究类论文不同的是，KDD 的应用类学术论文更加强调论文所描述的方法或者系统在实际应用中发挥的作用。比如，很多论文都是对现有的已部署的系统进行总结，对工业界的很多研究人员和工程师往往都有不小的借鉴意义。和研究类论文一样，从阅读经典文献和学习最新研究成果的角度，我们都应该认真分析和探讨每年的最佳应用类论文。

2017 年 KDD 最佳应用数据科学论文题目是，《HinDroid：基于结构性异构信息网络的智能安卓恶意软件检测系统》（HinDroid: An Intelligent Android Malware Detection System Based on Structured Heterogeneous Information Network）。可以说 2017 年是信息安全备受关注的一年，2016 年美国大选过程中传出了种种关于俄罗斯利用黑客入

侵大选候选人的新闻，让整个社会对信息安全的话题变得异常敏感。这是一篇有关如何智能地分析安卓恶意软件的论文，真是非常应景。

作者群信息介绍

文章的第一作者和第二作者都来自西弗吉尼亚大学（West Virginia University）的计算机科学与电气工程系。第一作者 Shifu Hou 是该系的博士生，先后发表过多篇论文。第二作者叶艳芳（Yanfang Ye）是该系的助理教授。叶艳芳 2010 年从厦门大学博士毕业，先后在金山公司和科摩多（Comodo Security Solutions）从事信息安全方面的研究和开发工作。2013 年，她加入西弗吉尼亚大学任教。这篇 KDD 论文因为第一作者也是在读学生，因此也是最佳学生论文。

第三作者宋阳秋（Yangqiu Song）是来自香港科技大学的计算机系助理教授。宋阳秋有丰富的学术和工业界经历。2016 年加入香港科技大学，在这之前曾经在西弗吉尼亚大学任教。2012 年到 2015 年之间他曾在伊利诺伊大学香槟分校、香港科技大学、华为诺亚方舟实验室等地访问。2009 年到 2012 年曾在微软亚洲研究院和 IBM 研究院工作。2009 年于清华大学博士毕业。

最后一位作者是土耳其企业家米勒夫·阿杜勒哈尤格鲁（Melih Abdulhayoğlu）。他是科摩多（Comodo）的 CEO，于 1998 年创立了公司。这篇论文挂了他的名字是因为使用了科摩多的数据。

论文的主要贡献

我们首先来看一下这篇文章的主要贡献。类似地，按照我们周一分析最佳研究论文的思路，首先必需弄明白，这篇文章主要解决了什么场景下的问题。

这篇文章希望解决的问题描述起来很直观，那就是**如何有效地监测安卓手机系统下的恶意软件**。经预测，到 2019 年，全球的手机市场中，77.7% 将是智能手机，这里面安卓系统的市场占有率至少是 80%。由于安卓系统的开放性以及分散的各类安卓手机软件市场，对安卓软件进行监控和分析存在很大难度。各类恶意软件在安卓生态系统中可以说层出不穷，比如 Geinimi、DroidKungfu 以及 Lotoor 等等。更悲观的统计来自赛门铁克（Symantec）的《互联网安全威胁报告》，认为五分之一的安卓软件是恶意软件。

之前很多恶意软件的分析和检测都是基于某种“**指纹签字**”技术，然而这种技术常常被恶意软件开发者的新手段绕过。因此，寻找更加复杂有效的检测方式就成了各种信息安全公司所

追逐的目标。

这篇论文的主要贡献是根据安卓的 API，提出了一种新的基于结构性异构信息网络的方法，来对安卓程序的 API 模式进行更加复杂的建模，从而能够理解整个安卓程序的语义。作者们还采用了**多核学习**（Multi-Kernel Learning）的方法，在结构性异构信息网络的基础上对程序语义模式进行分类。

最后，文章提出的方法在科摩多的真实数据上达到了非常高的准确度，远远优于现在的一些主流方法。并且，科摩多已经在产品中部署了这个方法。

论文的核心方法

了解了这篇文章的目的和贡献，接下来，我就来剖析一下作者们提出的方法。

首先，需要**将安卓的程序代码转换为可以分析的形式。**一般来说，安卓的软件被打包为后缀名为 Dex 的 Dalivik 执行文件，这个执行文件无法被直接分析。于是，需要把这个执行文件通过一个叫 **Smali** 的反汇编器解析成 Smali 代码。这个时候，软件的语义就能够通过 Smali 代码来解析了。作者们从 Smali 代码中提取所有的 API 调用，通过对 API 的分析来对程序行为建模。

防止断更 请务必加
首发微信：17161436665

下一步，就是要**从繁复的 API 调用中摸索出这里面的规律。**作者们这个时候构建了**四类矩阵**来表达 API 和某个 App 之间的基本特征：

1. 某一个 App 是否包含了某一个 API；
2. 某两个 API 是否同时出现在某一段代码中；
3. 某两个 API 是否出现在同一个 App 中；
4. 某两个 API 是否使用了相同的调用方法。

可以看出，这些矩阵可以抓住 API 和 App 之间的一个基本信息，还可以抓住一系列 API 同时出现进行某种操作的特征信息。这些矩阵成了发现高阶规律的基础。

为了发现更加复杂的规律，作者们在这里引入了一个工具叫**异构信息网络。**异构信息网络的概念最早由伊利诺伊大学香槟分校的数据挖掘权威韩家炜（Jiawei Han）和他当时的学生孙怡舟（Yizhou Sun，目前在加州大学洛杉矶分校任教）提出。异构信息网络的核心思想就是希望能够表达一系列实体（Entity）之间的复杂规律。

传统的方法把实体表达成图（Graph）的节点，实体之间的关系表达成节点之间的链接。这样的方式忽略了实体本身的不同以及关系的类型也有所不同。异构信息网络就是更加完整和系统地表达多种不同实体和实体关系的一种建模工具。在这篇文章中，有两类实体：App 和 API 调用，有四类关系（和刚才定义的矩阵相同）。而刚才定义的矩阵其实就是这四类关系所对应的图的邻接矩阵。

把 App 和 API 的关系描述成为异构信息网络以后，下面的工作就是**定义更高阶的规律关系**。为了更好地定义这些复杂关系，作者们使用了一个叫**元路径**（Meta-Path）的工具。元路径其实是提供一个描述性的模板语言来定义高阶关系。

比如，我们可以定义一个从 App 到 API 再到 App 的“路径”，用于描述两个 App 可能都含有相同的 API 调用。这个路径就可以帮助我们从最开始的四个矩阵推出更加复杂的矩阵来表达一些信息。那么，根据人们的领域知识（这里就是安全领域），作者们就定义了多达**16 种元路径**，从而全面捕捉 App 和 API 之间的各种关系。

利用异构信息网络和元路径构建了程序的语义表达后，下一步就是**进行恶意软件的判别**。这里，作者们采用了多核学习的思想。简而言之，就是把之前通过元路径所产生的新矩阵看作一个“核”。这里的多核学习就是要学习一个线性的分类器，特征是每个 App 到某一个核的一个非线性转换，这个转换是在学习过程中得到的。换句话说，这个多核学习的流程要同时学习一个分类器来判断一个程序是不是恶意程序，还需要在这个过程中学习从 App 到核的转换。

方法的实验效果

作者们使用了科摩多的数据集，收集了 2017 年两个月里 1834 个 App 的信息。正常程序和恶意程序几乎各一半。另外还有一个数据集包含 3 万个 App 信息，也几乎是正例负例各一半。从实验结果来看，结合了 16 个定义好的元路径的多核学习能够**实现高达 98% 的 F1 值**。F1 值可以认为是精度和召回的一个平衡，同时**准确率也是高达 98%**。

文章还比较了一些其他比较流行的方法，比如神经网络、朴素贝叶斯（Naïve Bayes）分类器、决策树以及支持向量机，这些方法基本的 F1 值都在 85% 和 95% 之间，和文章提到的方法有较大差距。另外，文章还和现在的一些商业软件，比如 Norton、Lookout、CM 做了比较。这些商业软件的准确度也在 92% 上下徘徊。因此，文章所采用的方法的确比之前的很多方法都更有效果。

小结

今天我为你讲了 KDD 2017 年的最佳应用类论文。这篇论文提出了，如何来分析安卓手机软件的行为进而检测手机应用是否是恶意软件。一起来回顾下要点：第一，简要介绍了这篇文章的作者群信息。第二，详细介绍了这篇文章要解决的问题以及贡献。第三，简要分析了文章提出方法的核心内容 。

总结一下，文章解决的问题就是如何有效监测安卓手机系统下的恶意软件，主要贡献是提出了一种新的基于结构性异构信息网络的方法，来理解安卓程序的语义。使用元路径的工具定义复杂关系，还采用了多核学习的方法完成恶意软件的判别。论文使用科摩多的数据集，验证了所提出的方法比当下流行的一些其他方法都更加有效。

最后，给你留一个思考题，文章中提到的多核学习方法这个步骤，是不是必需的？能否换成其他方法呢？

欢迎你给我留言，和我一起讨论。

拓展阅读：HinDroid: An Intelligent Android Malware Detection System Based on Structured Heterogeneous Information Network

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　012 | 精读2017年KDD最佳研究论文

下一篇　014 | 精读AlphaGo Zero论文

精选留言 (1)

写留言

huan
2017-11-01

3

既然作者使用非线性的转换，而且我理解此阶段的响应变量已经是常规的向量了，那么应该可以通用的分类器都可以做分类，比如SVM, NN和决策树都行。如果没有很好的分类数据，是否可以直接使用无监督的kNN来聚类完成。(初级的AI爱好者，请拍砖)

拼课微信：1716143665