文章编号:1007 - 2985(2013)04 - 0062 - 05

CVE 漏洞分类框架下的 SVM 学习模型构建*

彭 华,莫礼平,唐赞玉
(吉首大学信息科学与工程学院,湖南 吉首 416000)

摘 要:在 CVE 漏洞分类框架中,构建了基于支持向量机的学习模型,实现了根据不同的分类特征对 CVE 进行分类.

关键词:支持向量机(SVM);公共漏洞和暴露(CVE);分类特征;分类准确性
中图分类号:TP39 **文献标志码**:A **DOI**:10.3969/j.issn.1007 - 2985.2013.04.014

CVE 漏洞分类框架下的 SVM 对多种漏洞数据库(如 BID,X-Force,Secunia 等)中的分类特征进行自动集成,使得该框架能够以分类特征为基础实现 CVE 漏洞分类,从而成为一个拥有分类和归纳能力的 CVE 漏洞分类器.关于基于 SVM 的 CVE 漏洞分类框架的详细内容见文献[1].笔者分析了该框架下使用 SVM 为分类特征构造学习模型的方法,设计了数据融合和清理过程,从而消除训练数据的不一致性,使用所建立的学习模型对无标记的 CVE 漏洞进行分类,最后采用 n 倍交叉验证方法对学习模型的效果进行了评估.

1 分类特征训练数据的产生

为分类特征产生训练数据的学习模型使用 $T=\{\bar{x}_i, y_i\}$ 的形式来表达,其中 $i=1,\ldots,m$. 第 i 个数据点的特征向量和其真标记表示为 $\bar{x}_i \in R^d$, $y_i \in Y=\{l_1,\ldots,l_k\}$. 显然,如果手动地对 T 进行搜集和标记,将会费时费力.因此,该框架通过使用每个 CVE 条目中的引用池来自动产生训练数据.一方面,CVE 中大量的引用为 CVE 分类器提供了搜集分类信息的丰富资源;另一方面,不同资源中私有的数据格式、冲突的分类模式以及不一致的特征含义使整个信息抽取过程复杂化.考虑到特征数据的引用次数和质量,该框架主要使用漏洞数据库 BID,X-Force 和 Secunia 作为来源产生训练数据,见表 1.

表 1 Secunia 漏洞数据库中 SA - 11066 和 SA - 24893 条目

Part	SA - 11066	SA - 24893
General	Title: Symantec Client Firewall Products Multiple Vulnerabilities Secunia Advisory: SA11066; Release Date: 2004 - 05 - 13; Critical: Extremely critical; Impact: DoS, System access; Where: From remote; CVE reference: CVE - 2004 - 0445...	Title: McAfee e-Business Server Authentication Packet Processing Denial of Service Secunia Advisory: SA24893; Release Date: 2007 - 04 - 18; Critical: Less critical; Impact: DoS; Where: From local network; CVE reference: CVE - 2007 - 2151

* **收稿日期**:2013 - 03 - 12
基金项目:湖南省科技厅科技计划资助项目(2011FJ3209);湖南省教育厅一般科学研究资助项目(11C1025)
作者简介:彭 华(1980 -),男,湖南吉首人,吉首大学信息科学与工程学院讲师,硕士,主要从事网络安全、嵌入式系统研究.

续表

Part	SA－11066	SA－24893
Vulnerable	Software: Symantec Client Firewall 5. x; Symantec Client Security 1. x,2. x; Symantec Norton AntiSpam 2004 Symantec Norton Internet Security 2002, 2003... Symantec Norton Personal Firewall 2002, 2003,2004	Software: McAfee e-Business Server 8. x
Description	eEye Digital Security has discovered multiple vulnerabilities in various Symantec firewall products, which can be exploited by malicious people to cause a DoS (Denial of Service) or compromise a vulnerable system (1) A boundary error within the "SYMDNS. SYS" driver when processing certain NBNS (NetBIOS Name Service)...	A vulnerability has been reported in McAfee e-Business Server, which can be exploited by malicious people to cause a DoS (Denial of Service) The vulnerability is caused due to an error within the administration utility service when processing authentication packets. This can be exploited to crash the service...

创建 Secunia 的条目所使用的模板由 3 个部分组成,即 General,Vulnerable 和 Description. 其中,General 部分包含了描述性标题及其基本特征,Vulnerable 部分说明了有此漏洞的系统或产品,Description 部分则是对该漏洞进行的详细说明. 与 BID 比较而言,Secunia 漏洞数据库也通过一定的特征对安全漏洞进行分类,如表 1 中 SA－11066 和 SA－24893 条目所示,分别对应着 CVE－2004－0445 和 CVE－2007－2151. Secunia 中的 Impact 部分提供了在 Vulnerability Impact 特征上的分类信息,而 Critical 则指明了漏洞的严重性,与 Vulnerability Severity 特征对应. 在该框架下,相同的分类特征可以融合,而不同的分类特征可以互补,共同对漏洞进行统一分类.

来自不同来源的互补的分类特征也许会限制构建训练数据的来源,与分类特征 Vulnerability Cause 相关的信息只能从 BID 获得,而分类特征 Vulnerability Severity 的训练数据只能从 Secunia 获得. 因此,与某些分类特征相关的训练数据数量也许对于建立一个可信赖的学习模型来说是不充足的,究其原因有如下几点:(1)并不是每个 CVE 条目都有对漏洞数据库 BID,X-Force 和 Secunia 的引用;(2)漏洞数据库 BID,X-Force 和 Secunia 中的条目也许不能提供指定分类特征上的信息;(3)来自漏洞数据库 BID,X-Force 和 Secunia 中的数据也许是雷同的、重叠的或冲突的,减少了它们的可用性.

该框架被配置成使用 CVE 进一步扩大某分类特征的训练数据集,能够匹配问题中分类特征的任一唯一性关键词,与一个分类特征相关的关键词由领域专家使用 n-gram 产生器(该框架使用自然语言处理工具)进行创建. 框架中支持使用正则表达式的模式,关键词匹配过程仅被应用于 CVE 条目,该 CVE 条目并不在 BID,X-Force 或 Secunia 搜集来的训练数据中,若一个 CVE 条目匹配任一指定的关键词模式串,该条目被放入训练数据中.

2　数据的合并与清理

X-Force 漏洞数据库中 XF－16132 和 XF－33730 条目见表 2.

表 2　X-Force 漏洞数据库中 XF－16132 和 XF－33730 条目

Part	XF－16132	XF－33730
General	Title: Symantec Firewalls DNS response packets denial of service ID: symantec-firewall-dns-dos (16132); Severity: Medium Risk; Consequences: Denial of Service	Title: McAfee E-Business Server administration utility service denial of service ID: mcafee-ebusiness-utility-dos (33730); Severity: Low Risk; Consequences: Denial of Service

续表

Part	XF－16132	XF－33730
Description	Symantec Norton Internet Security and Personal Firewall devices are vulnerable to a denial of service attack, caused by an improper validation of Domain Name System (DNS) response packets by the SYMDNS. SYS driver. By sending a specially-crafted DNS response packet from UDP port 53 that contains a compressed name pointer that points to itself, a remote attacker could cause the...	McAfee E-Business Server is vulnerable to a denial of service, caused by the improper handling of authentication packets in the administration utility service. By sending a specially crafted authentication packet with an overly large length value, a remote attacker could exploit this vulnerability to crash the affected service
Platforms affected	Microsoft Corporation: Windows Any version; Symantec Corporation: Norton AntiSpam 2004; Norton Internet Security 2002, 2002 Pro, 2003...; Symantec Client Firewall 5.01 and 5.1.1, Symantec Client Security 1.0...	McAfee, E-Business Server 8.1 McAfee, E-Business Server 8.5.1

为获得足够的对应某分类特征的已标记样例,该框架从多个漏洞数据库中取得信息.不幸的是,不同来源的数据可能在分类特征的内涵、外延或粒度上存在不一致,因此要求在构建训练数据时进行数据合并和清理.例如,在为 CVE－2004－0445 搜集与分类特征 Vulnerability Impact 的相关信息时,从表 1 中 SA－11066 的 Impact 字段可以获得 DoS 和 system access,然而,在表 2 中 XF－16132 的 Consequences 字段仅能获得 DoS.显而易见,在数据合并之前,需要解决不同来源的命名上的不一致.Secunia 和 X-Force 对相同的分类使用不同的名字.例如前者中使用 DoS 和 system access,后者中使用 gain access.对于相同的分类特征,一些漏洞数据库将其作为一元分类特征,另外一些将其作为多元分类特征.结果在 CVE－2004－0445 中描述的安全漏洞,对于 Vulnerability Impact 特征被 X-Force 赋予单一的分类 DoS,在 Secunia 中同时被描述为 DoS 和 system access.所以建立分类映射模式来解决不同来源的分类特征维度上的不一致,例如,X-Force 和 Secunia 分别为分类特征 Vulnerability Impact 定义了 10 个和 11 个分类.来自多个来源的分类信息的不兼容也对数据合并提出了更大的挑战.例如分类特征 Vulnerability Severity 上,在 Secunia 上有 5 级度量标准,而在 X-Force 上只有 3 级度量标准.另外,CVE－2004－0445 安全漏洞被 Secunia 认为是极端严重的,但在 X-Force 中被认为只是较严重的.

来自不同来源的漏洞的不同定义和扩展,使得在一个 CVE 条目中描述的一个安全漏洞被当做不同漏洞数据库中的多个漏洞,导致同一个漏洞数据库中存在与一个 CVE 条目相关的多个引用.例如,CVE－2004－0452 可能由于其一元特征 Vulnerability Severity 被放入冲突的分类,因为该漏洞同时被 Secunia 中的 SA－12991 和 SA－18517 引用,它们分别具有 less critical 和 highly critical 的严重性.显而易见,分类信息的不一致产生有噪声的训练数据,并且影响了所构造学习模型的分类准确性.在训练数据产生期间进行数据合并和清理时,该框架遵守下列原则:(1)对于多元分类特征,若取自多重来源的数据能兼容并且能够建立一个分类映射模式,这些数据可以合并在一起;(2)对于一元分类特征,如果特征的分类能够排序的话,拥有最大值的信息来源将被使用;(3)对于不同来源间以不兼容方式定义的分类特征,CVE 分类器更倾向于使用拥有最好粒度的来源或拥有最大标记数量的来源.为搜集某特征对应的训练数据的算法如下所示.

算法 1

1:T 和 Y 分别是某分类特征的训练数据集和标记集,L 是该特征的二类分类器的集合;
2:for (标记集 Y 中每个分类 c) do
3:初始化正例集 T_p 和负例集 T_n 为空集;
4:for (训练数据集 T 中的每个条目 e) do
5:若 e 有标记 c,则将 e 放入 T_p,否则放入 T_n;
6:end for
7:基于正例和负例训练集——T_p 和 T_n,为分类 c 构建一个 SVM 二类分类器;
8:将产生的二类分类器放入 L;
9:end for
10:if(分类特征是 univariate 且其维度＞2)then
11:将该学习任务作为一个约束优化问题,并建立一个多类模型.

12:返回依赖于分类准确性的该多类模型或二类分类器集合 L
13:else
14:返回二类分类器集合 L 作为学习模型;
15:end if

算法 1 中描绘的过程用来为一个分类特征产生训练数据.为了一个 CVE 搜集分类信息,算法 1 首先标识了其对 BID,Secunia 和 X-Force 的引用,然后从每个引用来源处抽取数据并根据上述数据合并和清理的原则检测所取得数据的兼容性.例如,针对分类特征 Vulnerability Impact,来自 X-Force 和 Secunia 的数据是兼容的,因为这 2 个来源的分类是可转换的.相反,针对分类特征 Vulnerability Severity,X-Force 和 Secunia 的分类模式却是不兼容的,这是因为它们使用不同的度量标准.算法 1 也试图确定分类特征是单元的还是多元的,由于一个 CVE 能被赋予多个分类,因此分类特征 Vulnerability Impact 被标识为多元的.

3 SVM 学习模型的构造

当分类特征的维度(标记集 Y 的大小)等于 2 时,学习模型就是一个 SVM 二类分类器.对于 $|Y|>2$ 的分类特征,学习问题使用多类到二类削减方法被分解成 $|Y|$ 个二类分类任务.[2] 使用一对多的训练方法建立 $|Y|$ 个二类分类器:通过将训练数据中具有 l_i 标记的数据点作为正例,剩余的数据点作为负例,将 Y 的 l_i 标记的学习任务转换为 2 个分类.如 System Access 是分类特征 Vulnerability Impact 中 11 个分类中的 1 个.在其训练数据之中,CVE-2004-0445 记为正例,而 CVE-2007-2151 虽然拥有 DoS 的真标记,仍然被记为负例.当分类特征是一元的情况下,该框架也为其创建 1 个多类分类器而不是将其削减为 2 类分类任务[1].为某个指定的分类特征构造学习模型的过程如下所示.

算法 2

1:初始化某分类特征的训练数据集 T;
2:初始化特征的分类集 Y,特征类型 type 赋为 univariate(一元特征)
3:for (CVE 字典中的每个条目 e) do
4:取得 e 的引用池,把对 BID,Secunia 和 X-Force 的引用放入集合 P.
5:初始化 e 的分类集 A
6:for (引用池 P 中的每一项 p) do
7:抽取来自 p 的特征信息,检查它与 A 中数据的兼容性,若兼容的话,则加入集合 A 中.
8:end for
9:for (每一个(关键词,分类)对(k,c)) do
10:找到与关键词 k 匹配的 CVE 条目,然后把 c 放入集合 A
11:end for
12:把(e,A)对放入训练数据集 T 中,把分类信息 A 放入 Y,若 $|A|>1$,type 赋为 multivariate(多元特征)
13:end for
14:输出训练数据集 T、标记集 Y 和特征类型 type.

某分类特征的训练数据可能不会覆盖到 CVE 字典中所有条目,这就使得一些 CVE 条目无标记.漏洞数量的快速增长要求最新发现的漏洞需要被分类[3].该框架除了使用算法 2 为所有分类特征建立学习模型外,还使用它们对无标记 CVE 条目或新发现的漏洞进行分类.标记不可见数据点的过程如下所示.

算法 3

1:S 是无标记样例的测试集;L 和 Y 分别是学习模型和分类特征的标记集;type 是特征类型(univariate 或 multivariate);
2:for (S 的每个样例 s) do
3:if(L 是一个多类学习模型)then
4:将 L 的计算结果作为 s 的标记;
5:else
6:初始化由 L 赋给 s 的标记集 B;
7:for (学习模型 L 中每个二类分类器 l) do
8:使用 l 对 s 进行分类并将输出放入 B 中;
9:end for
10:if(特征类型是 univariate)then
11:找到 B 的最大值,并将其对应的分类作为 s 的标记.
12:end if
13:end if
14:end for

对于一个多元分类特征,只要无标记 CVE 的二类分类器对该 CVE 输出正值,该无标记 CVE 就可能被赋予多个分类.而对于一个一元分类特征,算法 3 仅仅将无标记 CVE 放入有最高输出的分类中.例如,假定一元分类特征 Vulnerability Severity 包含 5 个分类,并且其学习模型由 5 个二类分类器组成,如果第 2 个二类分类器输出正值而其余 4 个分类器输出的是负值,那么 CVE-2005-1993 将归属第 2 个分类.

该框架使用 n 倍交叉验证方法来评估学习模型在分类准确性、精确性、回归和 F_β 的效果. 对于某分类特征的训练数据集 T,首先被划分为相同大小的 n 个组($T_i, i=1,2,\ldots,n$),每组假定有与 T 相似的特征分类的 CVE 分布,每个部分均包含了所有可能类的样例. 然后,交叉验证过程在第 i 次迭代时采用下列步骤进行处理:(1) 训练阶段. T_i 作为验证集而其余($n-1$)部分合并在一起形成一个新的训练集 $A=\bigcup_{j\neq i} T_j$,并通过算法 2 使用训练数据集 A 建立学习模型. (2) 标记阶段. 在算法 3 的帮助下,使用第(1) 阶段得到的学习模型对验证集 T_i 中的样例进行标记. 若学习模型赋予 T_i 中某样例的标记与其真标记相同,则该样例被正确分类. (3) 度量阶段. 学习模型的效果(如分类准确性、精度以及回归等) 通过计算进行度量.

分类准确性定义为分类准确性 = 正确分类的样例数 / 验证集的样例数,该框架还使用了 F_β 来计算精度 P 和回归 R 的加权调合平均值,其公式为 $F_\beta=(1+\beta^2)PR/(\beta^2P+R)$. 一个类的精度 P 定义为 P = 正确标记的样例数 / 归属该分类中的样例总数,回归 R 的定义为 R = 正确标记的样例数 / 实际属于该分类中的样例总数. 通过设置 $\beta=1$,精度 P 和回归 R 被认为是同等重要,可以获得 F_1,且 $F_1=2PR/(P+R)$. 通过学习模型得到的形如分类准确度和精度的度量效果是交叉验证过程的 n 次迭代获得的度量结果的平均值. 经验证,该学习模型具有较好的性能.

4 结语

在 CVE 漏洞分类框架下设计并构造了基于 SVM 的学习模型,加强和完善了该框架对 CVE 漏洞分类的能力. 为进一步完善该框架的功能和灵活性,下一步,笔者拟使用带决定性属性和特殊性属性的分类特征集合进行研究,并集成包含隐 Markov 模型和条件随机场在内的建模方法来提高分类性能.

参考文献:

[1] 彭 华,李宗寿. 基于 SVM 的 CVE 漏洞分类框架构造 [J]. 吉首大学学报:自然科学版,2013,34(1):66-71.

[2] 刘奇旭,张翀斌,张玉清,等. 安全漏洞等级划分关键技术研究 [J]. 通信学报. 2012,33(S1):79-87.

[3] 廖晓锋,王永吉,范修斌,等. 基于 LDA 主题模型的安全漏洞分类 [J]. 清华大学学报:自然科学版,2012,52(10):1 351-1 355.

Construction of a SVM Learning Model in the Categorization Framework for CVE

PENG Hua, MO Li-ping, TANG Zan-yu

(College of Information Science and Engineering, Jishou University, Jishou, 416000, Hunan China)

Abstract: In the categorization framework for CVE, this paper designs and constructs a learning model based on SVM, so that it can categorize the CVE according to the different taxonomic features. In the process of constructing a learning model based on SVM, first of all, the training data is generated according to the different taxonomic features in the several vulnerability databases, then a data fusion and cleansing process are designed to eliminate the inconsistencies of data, and finally the n-fold cross-validation method is used to evaluate the effect of the model. The learning model has been verified to have better performance of CVE classification.

Key words: support vector machine (SVM); common vulnerabilities and exposures (CVE); taxonomic feature, classification accuracy

(责任编辑 陈炳权)