计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190600156

一种基于漏洞威胁模式的网络表示学习算法

黄　易[1,2]　申国伟[1,2]　赵文波[1]　郭　春[1,2]
1 贵州大学计算机科学与技术学院　贵阳 550025
2 贵州大学贵州省公共大数据重点实验室　贵阳 550025
(yHuang_Addy@163.com)

摘　要　威胁情报分析可为网络攻防提供有效的攻防信息,而细粒度的挖掘即网络威胁情报数据中的安全实体及实体间的关系,是网络威胁情报分析研究的热点。传统的机器学习算法,在被应用到大规模网络威胁情报数据分析中时,面临着稀疏、高维等问题,进而难以有效地捕获网络信息。为此,针对网络安全漏洞的分类问题,文中提出了一种基于漏洞威胁模式的网络表示学习算法——HSEN2vec。该算法旨在最大限度地捕获异构安全实体网络的结构和语义信息,并从中获得安全实体的低维向量表示。该算法首先基于漏洞威胁模式获取异构安全实体网络的结构信息,随后通过 Skip-gram 模型建模,并通过负采样技术进行有效预测进而得到最终的向量表示。实验结果表明,在国家安全漏洞数据上,与其他方法相比,利用所提算法进行漏洞分类的准确率等评价指标有所提升。

关键词:网络表示学习;异构安全实体网络;威胁模式;漏洞
中图法分类号　TP393.0

Network Representation Learning Algorithm Based on Vulnerability Threat Schema

HUANG Yi[1,2],SHEN Guo-wei[1,2],ZHAO Wen-bo[1] and GUO Chun[1,2]
1 Department of Computer Science and Technology,Guizhou University,Guiyang 550025,China
2 Guizhou Provincial Key Laboratory of Public Big Data,Guizhou University,Guiyang 550025,China

Abstract　Threat intelligence analysis can provide effective attack and defense information for network attack and defense,and fine-grained mining,that is,the relationship between security entities and entities in network threat intelligence data,is a hotspot of network threat intelligence analysis research. Traditional machine learning algorithms,when applied to large-scale network threat intelligence data analysis,face sparse,high-dimensional and other issues,and thus it is difficult to effectively capture network information. To this end,a network representation learning algorithm based on vulnerability threat schema——HSEN2vec for the classification of network security vulnerabilities is proposed. The algorithm aims to capture the structure and semantic information of the heterogeneous security entity network to the maximum extent,and obtains the low-dimensional vector representation of the security entity. In the algorithm,the structural information of the heterogeneous security entity network is obtained based on the vulnerability threat schema,and then modeled by the Skip-gram model,and the effective prediction is performed by the negative sampling technique to obtain the final vector representation. The experimental results show that in the national security vulnerability data,compared with other methods,the learning algorithm proposed in this paper improves the accuracy of vulnerability classification and other evaluation indicators.

Keywords　Network representation learning,Heterogeneous security entity network,Threat schema,Vulnerability

1　引言

随着互联网技术的发展,利用网络安全漏洞实施攻击的手段和形式日益复杂。为了确保网络信息安全,国家及企业部署了大量网络安全威胁检测设备,产生了海量的安全数据。此外,国家网络安全保障部门、安全厂商等发布了大量的安全漏洞数据,形成了网络安全威胁情报大数据[1-2]。因此,如何从碎片化、海量化的威胁情报大数据中挖掘出细粒度信息之间的关联关系,成为了威胁情报分析研究的焦点。

在细粒度的威胁情报分析中,漏洞是一类关键的网络安

到稿日期:2019-06-19　返修日期:2019-11-24　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61802081);贵州省科技重大专项计划项目(20183001);贵州省科技计划(20161052,20171051)
This work was supported by the National Natural Science Foundation of China (61802081),National Science and Technology Major Project of the Ministry of Science and Technology of Guizhou Province,China(20183001) and Guizhou Provincial Science and Technology Plan (20161052,20171051).
通信作者:申国伟(gwshen@gzu.edu.cn)

全实体[3]。图 1 给出了典型的结构化漏洞数据与非结构化文本数据之间的关联分析实例。通过漏洞信息关联，“SQL 注入”漏洞演变成了“二阶 SQL 注入”漏洞。因此，基于漏洞的数据挖掘分析可为网络威胁检测、漏洞利用等提供关键信息[4]。

为了实现细粒度的威胁情报大数据分析，本文在网络安全实体、实体关系抽取的基础上，通过异构安全实体网络模型对其进行表示。该模型能够表示不同类型安全实体之间的关联关系，能够保存结构信息和实体本身的属性信息等，为大规模漏洞数据挖掘分析奠定了基础[5]。然而，在对大规模异构安全实体网络进行分析时，传统的机器学习算法面临维度高、数据稀疏等问题[6]。

针对高维、稀疏的网络数据，网络表示学习将有价值的信息从原始数据中编码到一个低维、稠密的向量空间中[7]，以数值化的方式表示网络的语义信息。目前，为大规模网络中的每个节点自动学习到合适的表示[8]已成为学者们研究的热点。

近年来，大量学者提出了一系列网络表示学习模型和算法。2013 年，谷歌发布了词嵌入模型 Word2vec，掀起了嵌入表示学习的研究热潮。随着神经网络在计算机视觉和自然语言处理等领域的迅猛发展，研究者们开始不断尝试利用深度模型来对大规模的数据进行建模[9]。2014 年，DeepWalk 算法[10]将深度学习技术应用于网络表示学习领域。DeepWalk 算法采用随机游走的方式，一方面降低了模型的计算复杂度，另一方面有效解决了邻接矩阵的数据稀疏问题[11]。基于神经网络的算法可以有效地对大规模数据网络进行建模，能够获取网络结构本身携带的信息及网络节点所隐藏的复杂语义信息[12-13]。

传统的网络表示学习算法，主要针对的是同构网络表示学习问题[14]。然而，异构安全实体网络中包含多类实体和多种类型的实体关系，倘若按照同构网络处理，则可能会丢失重要的、有价值的语义信息[15]。从异构网络的角度对网络节点进行建模[16]，固然能够提升网络表示学习的区分能力和效果[17]，但是结构信息和异构实体间的语义信息却难以同时获得有效保留。

通过上述分析可知，大规模异构安全实体网络中的漏洞数据挖掘分析任务，仍然面临以下难点：

(1)针对漏洞数据分析任务，漏洞本身存在特定的威胁模式，且漏洞威胁模式通常都是多种类型安全实体之间的关联语义信息，因此漏洞本身的威胁模式难以融入到表示学习模型中。

(2)在大规模异构网络安全实体网络表示学习中，同一个模型难以同时保存异构安全实体网络的结构和语义信息。

为了学习大规模异构安全实体网络中漏洞节点的低维嵌入表示，本文提出了基于漏洞威胁模式的表示学习算法，最大限度地捕获异构安全实体网络的结构和语义信息。本文的主要贡献如下：

(1)针对异构安全实体网络表示学习，对异构安全实体网络进行问题抽象及形式化描述。对于大规模的漏洞数据，采用异构安全实体网络进行建模，并将漏洞威胁模式融入到异构网络中。

(2)提出了基于漏洞威胁模式的异构网络表示学习算法——HSEN2vec，最大限度地捕获了异构安全实体网络的语义和结构信息。

(3)在真实的漏洞数据集上，围绕漏洞节点分类任务开展实证研究，实验结果表明，本文方法在实际的应用场景下能取得较好的成果。

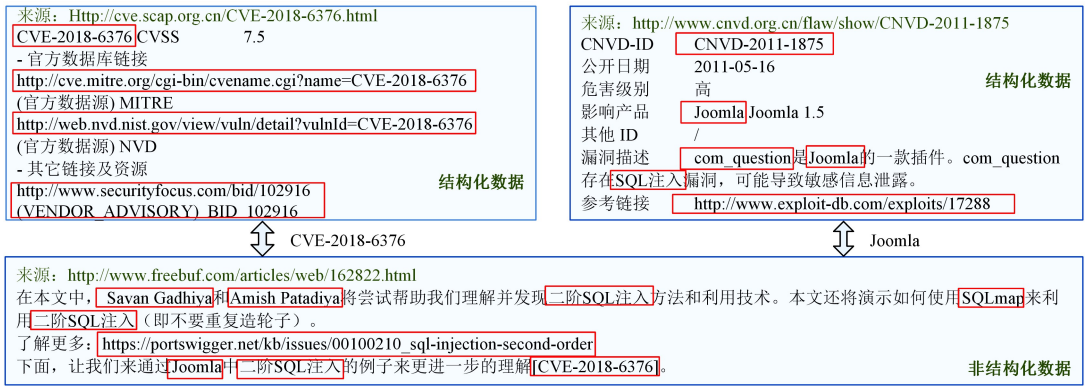


图 1　漏洞数据关联分析

Fig. 1　Vulnerability data association analysis

2　问题定义

网络威胁情报中的漏洞数据包括漏洞名称、危害等级、漏洞类型、威胁类型、漏洞描述信息以及该漏洞的补丁信息等。数据中包含不同类型的安全实体，且不同类型安全实体间存在着一定的联系，由此，构建出一个异构安全实体网络(Heterogeneous Security Entity Network，HSEN)。

异构安全实体网络形式化表示为 $HSEN=(V,E,\Phi,\Psi)$，其中，V 是安全实体节点集合，每个节点代表一个安全实体对象；E 是安全实体间的链接集合，每条链接代表两个安全实体间的关系；$\Phi:V\to A$ 和 $\Psi:E\to R$ 分别是安全实体节点和链接的类型映射函数。其中，每个安全实体 $v\in V$ 属于 A 中的一个特定节点类型，即 $\Phi(v)\in A$，且每个链接 $e\in E$ 属于 R 中的特定链接类型，即 $\Psi(e)\in R$。

本文对如何更加有效地对异构安全实体网络中的漏洞分类进行研究，重点从国家信息安全漏洞数据中抽取出漏洞名、危害等级、厂商以及软件名等安全实体，并以此构建一个异构安全实体网络，为复杂网络漏洞检测提供丰富的

网络安全威胁信息。

图 2 给出了一个异构安全实体网络，其中，S，B，C，L 分别表示软件、漏洞、厂商和危害等级类型节点，链接表示共存漏洞关系（B-B）、生产关系（S-B，C-S）及所属关系（L-B），虚线则表示共存漏洞关系。

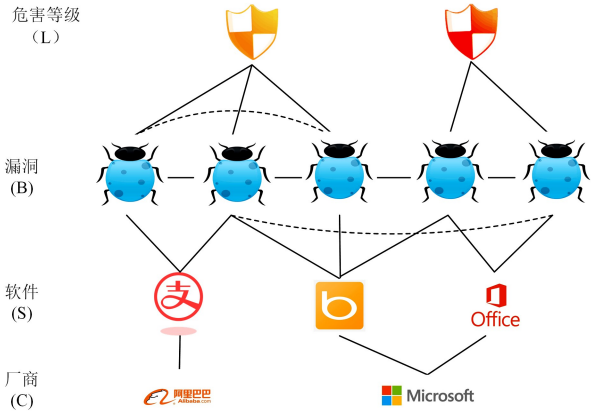


图 2 异构安全实体网络

Fig. 2 Heterogeneous security entity network

通过上述分析可知，漏洞的分类可以在大规模异构安全实体网络表示学习得到低维向量表示后，进行实体分类。给定一个异构安全实体网络 $HSEN=(V,E,\Phi,\Psi)$，其中，V 是节点集合，E 是边集合，边 $e=(V_i,V_j)\in E$ 表示节点 V_i 到 V_j 的一条边。异构安全实体网络表示学习，旨在将语义信息融入到 $HSEN$ 中，以便将安全实体映射到低维空间。每一个安全实体 $v\in V$ 学习一个低维向量，其中向量的维度 d 远远小于安全实体的总个数，即 $d\ll|V|$，保留了安全实体间的邻近性。

3 基于漏洞威胁模式的网络表示学习

本文引入了异构 Skim-gram 模型，同时通过威胁模式（Threating Schema，TS）将异构安全实体的网络结构和语义信息捕获到异构 Skim-gram 模型中进行训练，进而得到节点的向量表示，最终实现异构安全实体网络的低维嵌入表示学习。

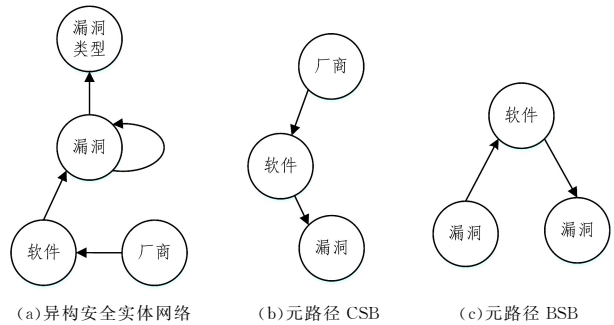
3.1 威胁模式的生成算法

以往针对同构网络的表示学习，研究者采用随机游走的方式遍历网络中所有的节点作为训练模型的输入，例如 DeepWalk 和 Node2vec 模型[18]。受此启发，在对异构安全实体网络进行向量表示时，本文也通过随机游走的方式在异构安全实体网络中构造多种类型安全实体节点的威胁模式，借此捕获到网络结构作为 Skip-gram 模型的输入，并最终得到安全实体网络的节点表示。

然而，通过 Sun 等的实验证明发现，在异构网络中进行随机游走时，会有节点偏向于高度可见的节点类型[19]。且在异质网络中，控制游走下一步的条件概率 $P(V^{i+1}|V^i)$ 不会像 DeepWalk 那样直接在节点的所有邻节点上做标准化，否则就忽视了上下文的节点类型信息，从而无法获取到网络的整个结构。

此外，在异构安全实体网络中，由于链接类型的多样性，实体节点的链接也更为复杂化。例如，链接关系“漏洞-软件-漏洞”表示一个软件中是否存在着不同的漏洞，而“漏洞-软件-厂商-软件-漏洞”则表示两个漏洞是否存在于同一厂商开发的软件中，这两种链路关系表示了不同节点间的不同关系，相互间存在着不同的语义。因此，在异构安全实体网络中，两个安全实体可以由不同的路径连接，形式化地，这些路径连接方式被称作元路径[20]，其定义如下。

元路径 P 是存在于网络中的一条路径，它的形式为 $V_1\xrightarrow{R_1}V_2\xrightarrow{R_2}\cdots\xrightarrow{R_l}V_{l+1}$，定义了类型 V_1 和类型 V_{l+1} 间的复合关系 $R=R_1\circ R_2\circ\cdots\circ R_l$，其中“$\circ$”表示关系上的复合运算。例如，对于图 3(a)所示的异构安全实体网络模式，得出了图 3(b)和图 3(c)所示的元路径例子，其中箭头的指示表明了关系的方向。由此，元路径“厂商→软件→漏洞(CSB)”表示一个厂商开发的软件存在的漏洞；“漏洞→软件→漏洞(BSB)”则表示一个软件下的漏洞间存在的关联。



(a)异构安全实体网络 (b)元路径 CSB (c)元路径 BSB

图 3 异构安全实体网络及其元路径

Fig. 3 Heterogeneous security entity network and meta-path

在现有的很多基于图的网络表示学习研究中，验证了元路径在数据的分类、聚类等任务中表现良好，因此本文将采用基于元路径的方式引导整个异构安全网络的随机游走[21]，进而基于威胁模式进行随机游走，从而生成安全实体网络的结构和语义关系的异构信息，并将其作为异构 Skip-gram 模型的输入，最终输出安全实体的向量表示。为了将威胁模式归并到异构 Skip-gram 模型中，将进一步定义威胁模式 TS。

给定一个异构安全实体网络 $HSEN=(V,E,\Phi,\Psi)$ 以及威胁模式 TS 的随机游走方案 $\theta:V_1\xrightarrow{R_1}V_2\xrightarrow{R_2}\cdots V_t\xrightarrow{R_t}V_{t+1}\cdots\xrightarrow{R_{t-1}}V_t$，$TS=R_1*R_2*\cdots*R_l$，那么，第 i 步的转移概率可定义为：

$$P(V^{i+1}V_t^i,\theta)=\begin{cases}\dfrac{1}{|N_{t+1}(V_t^i)|}, & (V^{i+1},V_t^i)\in E,\Phi(V^{i+1})=t+1\\ 0, & (V^{i+1},V_t^i)\in E,\Phi(V^{i+1})\neq t+1\\ 0, & (V^{i+1},V_t^i)\notin E\end{cases}\tag{1}$$

其中，$v_t^i\in V_t$，$N_{t+1}(v_t^i)$ 表示网络安全实体 V_t^i 对应 V_{t+1} 种类型安全实体中的邻居实体。同时，为了对随机游走有一个递推引导，第一个节点 V_1 的类型会和最后一个节点 V_l 的类型相同，也就是说，将威胁模式 TS 设置为对称元路径[22]：

$$P(V^{i+1}V_t^i)=p(V^{i+1}V_1^i),\text{if } t=l\tag{2}$$

基于对称类型的威胁模式使得安全实体间的语义和结构信息合并到异构安全实体网络中，并用作异构 Skip-gram 模型训练的输入。

在异构安全实体网络中,包含了大量的如漏洞名、软件名以及厂商等信息,这些信息包含了大量已知的威胁模式,由图 2 所示的异构安全实体网络可得到图 4 所示的威胁模式。

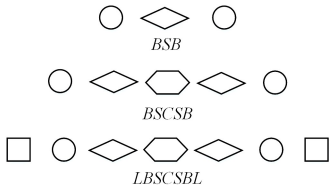


图 4　异构安全实体网络威胁模式

Fig. 4　Heterogeneous security entity network threat schema

通过图 2 构建出的威胁模式的含义分别解释为:“漏洞→软件→漏洞(*BSB*)”“漏洞→软件→厂商→软件→漏洞(*BSCSB*)”及“漏洞类型→漏洞→软件→厂商→软件→漏洞→漏洞类型(*LBSCSBL*)”。综合各因素考虑,由于本文主要研究漏洞实体的向量表示,因此采用“漏洞→软件→厂商→软件→漏洞”的威胁模式。将通过该威胁模式捕获到的安全实体节点的上下文,作为下游训练模型的输入。

3.2　异构 Skip-gram 算法

为了能够捕获一个节点的异构上下文结构和语义信息,本文引入了异构 Skip-gram 模型,并采用基于威胁模式的随机游走方式,将节点的网络结构归并到异构 Skip-gram 模型中进行训练。

对于一个异构安全实体网络 $HSEN=(V,E,\Phi,\Psi)$,$|T_V|>1$,给定一个节点 v,为了使其异构上下文结构最大化,可得到:

$$\arg\max_{\theta}\sum_{v\in V}\sum_{t\in T_V}\sum_{c_t\in N_t(v)}\log p(c_t\mid v;\theta) \tag{3}$$

其中,$N_t(v)$表示 t 种漏洞安全实体类型对应的邻节点,$p(c_t\mid v;TS)$为 softmax 函数,本文 softmax 函数的定义如下:

$$p(c_t\mid v;TS)=\frac{\exp(f(c_t,v,TS))}{\sum_{u\in V}\exp(f(c_t,v,TS))} \tag{4}$$

与已有 Skip-gram 模型之间的最大区别在于,异构 Skip-gram 模型不再将节点的上下文作为一组多项分布输出,而是为该节点的上下文中的每种类型节点分别指定一组多项分布。

3.3　异构负采样算法

在式(4)中构造节点 v 的上下文 $N_t(v)$时,忽略了函数 $softmax$ 中的节点类型信息,为此,本文将 $softmax$ 函数相对于上下文 c_t 的节点类型进行了归一化处理。因此,有:

$$p(c_t\mid v;TS)=\frac{\exp(f(c_t,v,TS))}{\sum_{u_t\in V_t}\exp(f(u_t,v,TS))} \tag{5}$$

其中,V_t 是网络中类型为 t 的一个节点集合。在 DeepWalk 和 Node2vec 中,输出多项分布的维数等于网络中节点的数量,然而在本文的异构 Skip-gram 中,类型为 t 的多项分布维度由类型为 t 的节点数决定,如图 5 所示。

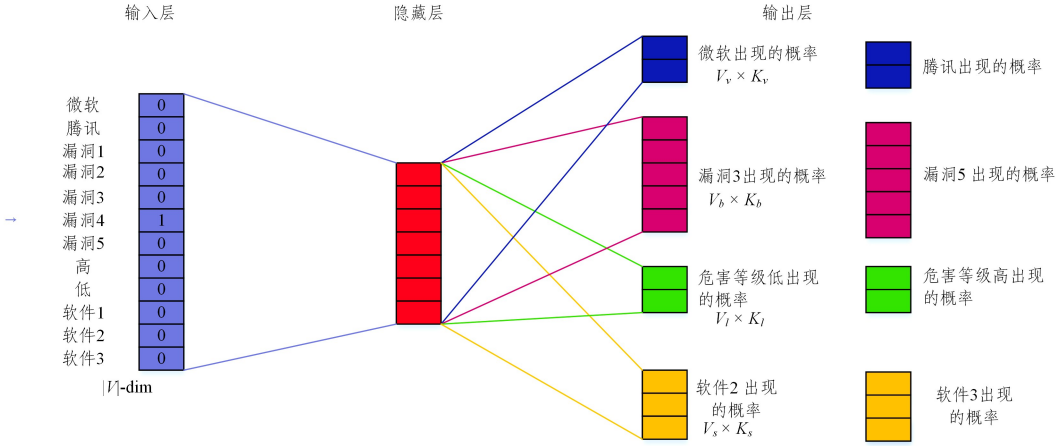


图 5　异构 Skip-gram 模型

Fig. 5　Heterogeneous Skip-gram model

最终,本文的采样分布由目标预测的邻节点的类型(即$(P_t\cdot)$)指定[23],因此得到如下函数:

$$O(X)=\log\sigma(X_{c_t}\cdot X_v)+\sum_{m=1}^{M}E_{u_t^m}\sim P_t(u_t)[\log\sigma(-X_{u_t^m}\cdot X_v)] \tag{6}$$

其梯度的来源如下:

$$\frac{\partial O(X)}{\partial X_{u_t^m}}=(\sigma(X_{u_t^m}\cdot X_v-I_{c_t}[u_t^m]))X_v \tag{7}$$

$$\frac{\partial O(X)}{\partial X_v}=\sum_{m=0}^{M}(\sigma(X_{u_t^m}\cdot X_v-I_{c_t}[u_t^m]))X_{u_t^m} \tag{8}$$

给定负样本大小 M,其中 $P(u)$是预定义的分布,将负节点 u^m 从该分布中抽出 M 次。$I_{c_t}[u_t^m]$是一个指示函数,用于指示 $m=0$,$u_t^0=c_t$ 时,u_t^m 是否为节点 c_t 的上下文节点,最后模型采用随机梯度下降算法来进行优化。整个表示学习算法如算法 1 所示。

算法 1　基于漏洞威胁模式的表示学习算法 HSEN2vec

输入:异构安全实体网络 HSEN=(V,E,Φ,Ψ),随机游走方案 θ,节点随机游走数目 w,随机游走长度 l,嵌入维度 d,邻节点窗口大小 k

输出:安全实体对应嵌入表示 $\mathbf{X}\in\mathrm{R}^{|V|*d}$

```
1. 初始化 X;
2.     for i=1 → w do
3.         for v∈V do
4. TS[1]=v;
5.             for i=1 → l-1 do
6.                 根据式(1)绘制 u;
7. TS[i+1]=u;
8.             end
```

9. return TS;
10. for i=1 → l do
11. v=TS[i];
12. for j=max(0,i−k) →min(i+k,l) &j ≠ i do
13. c_t=TS[j];
14. $X^{new}=X^{old}-\eta \cdot \frac{\partial O(X)}{\partial X}$
15. end
16. end
17. end
18. end
19. return **X**.

4 实验及分析

为了证明基于漏洞威胁模式的网络表示学习算法的有效性，本文从多类型安全漏洞信息分类应用中开展实验，展示了基于漏洞威胁模式的网络嵌入在异构安全实体网络上的学习效果。

4.1 实验设置

针对漏洞节点的分类任务开展对比实验，对比算法主要包括以下 3 种。

(1)Word2Vec：该方法中的 Skip-gram 模型根据给定的语料库，快速、有效地将一个词语表达成向量形式。

(2)DeepWalk：该方法通过在图中进行随机游走来得到节点序列，将序列看作一个特殊的“句子”输入 Skip-gram 模型，为每个节点学习到一个 d 维的向量表示。

(3)Node2Vec：该方法基于 DeepWalk 模型，通过参数 p 和 q 控制游走时邻节点选择的策略，最后将序列输入 Skip-gram 模型学习到节点的 d 维向量表示。

基于以往网络表示学习模型训练结果的比较，本文算法的主要参数分别设定为表示向量维度 $d=128$、节点游走步数 $w=100$、游走步行长度 $l=10$，算法设定节点信息表示学习窗口大小 k 为 7，负采样大小为 5。为保证公平实验，各对比算法的向量维度都设置为 $d=128$。

实验采用 CNNVD 数据集，该数据集来源于国家信息安全漏洞库，将抽取到的漏洞、软件名、漏洞类型和厂商作为网络节点，利用漏洞所属类型等关系构建异构安全实体网络。其中包含节点 84642 个、连边 77626 条以及对应不同漏洞节点类型的 7 类节点标签。

在构造威胁模式时，本文采用了最常用的威胁模式方案“漏洞→软件→厂商→软件→漏洞(*BSCSB*)”，“*BSCSB*”表示同一厂商的软件出现漏洞的异质语义。通过实际运用的实验证明，基于元路径构造的威胁模式可以运用于各种网络安全挖掘任务，这为以后的安全实体网络挖掘分析提供了极大的便利。

4.2 基于不同训练模型的向量表示分类对比实验

为了将引入威胁模式的数据同其他方法进行对比，本文首先在异构安全实体网络中构造出基于漏洞威胁模式的异构信息，即先对从国家安全漏洞数据库采集到的数据进行预处理，对漏洞、软件以及厂商的信息进行单独抽取，再根据构造威胁模式的方案“*BSCSB*”将 3 类安全实体间的关联包含进去，在捕获到网络的异构信息后将其放入异构 Skip-gram 模型进行训练，从而得到相应的词向量；与此同时，将原始的异构安全实体网络放入对比模型中进行训练，并得到相应的词向量；随后，将各模型的向量表示作为随机森林以及决策树算法的输入进行分类对比实验，分别计算并比较不同模型训练得出的准确率、召回率和 F1 值。

由图 6 所示的训练结果可知，在同等条件下基于威胁模式训练出的词向量经过分类实验对比后，节点分类的准确率、召回率和 F1 值都有所提升。同时，由图 7 所示的结果可以发现，在使用不同的分类算法进行测验时，基于威胁模式的网络表示学习能够很好地应用于节点分类任务。这是由于在引入威胁模式时，异构安全实体网络的结构和语义信息也会引入其中，使得训练模型可以有更多的信息进行链接和比较。另外，在训练过程中，模型将上下文的节点类型进行了归一化处理，降低了时间计算复杂度，更好地保留了上下文的结构和语义信息，进而训练出了更多准确且有价值的词向量。

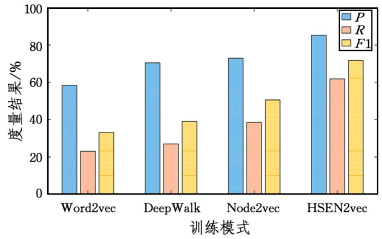


图 6 不同训练模式的对比

Fig. 6 Comparison of different training modes

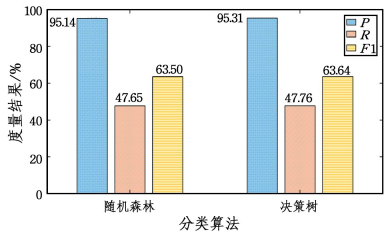


图 7 不同分类算法的对比

Fig. 7 Comparison of different classification algorithms

总而言之，引入基于威胁模式构造出的异构信息后的训练能力是有效的，这是由于该算法能够很好捕获将多种类型节点及其上下文的关系信息。

4.3 基于 TS 的向量表示的数据属性维度对比实验

本文将比较不同比例数据的输入对分类实验结果的影响。首先，仍采用上述不同训练模式对比实验中使用的基于漏洞威胁模式训练得到的词向量，将得到的词向量的 30% 作为测试集，剩余的 70% 作为训练集；然后，将训练的数据集从 60% 变到 100% 进行训练；最后，得到不同比例训练数据集下的准确值、召回率和 F1 值的对比结果。

表 1 列出了不同比例训练数据集的对比结果。通过对比实验的数据值可以看出，随着训练数据集的增加，分类实验的准确率等都有所增加。在进行分类实验时，尽管给出的训练集的数量不多，但其最终的分类结果仍较理想，例如，当只给出 60% 的训练数据时，最终分类的准确率仍然可达到近 60%；当训练数据集每增加 10% 时，最终的分类准确率也有所提升，因此，在放入所有的训练集时，分类的准确率高达 95%。由此可见，倘若只有一份测试集，尽管训练的数据数量

不够多，但依旧可以训练出较为理想的结果；另一方面，若掌握的数据足够丰富，则可为研究者们在实验分析时挖掘到更多有价值的信息。

表 1　数据属性维度的对比

Table 1　Comparison of data attribute dimension

数据比/%	准确率 P/%	召回率 R/%	F1 值
60	57.57	28.67	38.28
70	62.95	32.90	43.90
80	75.01	37.49	49.99
90	86.29	43.15	57.53
100	95.86	47.93	63.90

总之，通过节点分类实验可以看出，基于漏洞威胁模式训练得到的词向量适用于异构安全实体网络中的表示学习，这是由于基于威胁模式的随机游走方式能够尽可能多地包含网络中的信息。

4.4　基于 TS 的参数灵敏度对比实验分析

在训练词向量的过程中，本文采用了一种能从大规模非结构化的网络数据中学习到高质量的向量表示的异构 Skip-gram 模型，它可以将句子或文本中的词汇间的共现几率最大化。Skip-gram 模型由输入层、映射层以及输出层构成，在输出层采用了构造 Huffman 树计算概率值的方法，替换了原始模型利用 softmax 计算概率值的方法[24]。在 Skip-gram 模型中，输出层对词进行训练时，它会将词汇表中的所有词看作是 Huffman 树的叶子节点，将每个单词的词频看作节点的权重，从而构造出一个 Huffman 树当作 Skip-gram 模型的最终输出结果。当明确 Huffman 树的叶子节点及叶子节点权重时，Huffman 树的带权路径长度最短，若 Huffman 树的叶子节点权重越大，则该节点离根节点越近，且从根节点出发，只有唯一的路径可抵达该节点。因此，在 Skip-gram 模型中，词频越高离根节点越近，而其中存在的参数“min-count”则控制了训练词向量的最低词频。本文通过调整该参数的值，观察不同参数值对分类实验结果的影响。

图 8 给出了不同的参数 *min-count* 值对分类实验结果的影响。总体来说，随着 *min-count* 值的增加，节点分类的结果逐渐提升，但是当 *min-count* 达到一定的大小后，分类的结果变化开始趋于平缓。这是由于，在模型训练过程中，随着 *min-count* 逐渐增大，放入模型训练的词汇量也相应增加，从而延长了模型训练的时间，增大了空间复杂度。因此，综合各方面的因素，本文决定将 *min-count* 的值设为 6，这样不仅得到了理想的分类结果，也减小了模型训练的时间复杂度等。

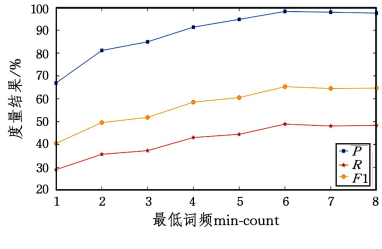


图 8　参数灵敏度的对比

Fig. 8　Comparison of parameter sensitivity

通过上述分析可知，模型的训练结果对于参数 *min-count* 值的设置还是较为敏感的，为了使整个模型的训练达到高效能，本文需要根据实际应用设定其相应的有效值。另外，由于在先前的研究中已经表明这些参数在成本有效的选择下便能达到相应的高性能，且这些参数表明了异构网络和同构网络表示学习的不同功能，展示了异构网络表示学习的不同思路和解决方案的要求，因此对于训练模型其他参数的敏感度分析不再做详细陈述。

4.5　基于 TS 的词云图实际应用分析

本文通过一个实际运用展示了本次实验的有效性。在经过异构 Skip-gram 模型训练之后，本文会得到漏洞相应的词向量表示，为了能直观观察训练结果，根据漏洞类型的不同，将训练结果绘制成一幅词云图进行分析，如图 9 所示。

(a)跨站脚本类型　(b)缓冲区类型

(c)资料不足类型

图 9　词云图实际应用分析

Fig. 9　Analysis of actual application of word cloud map

从图 9 可以发现，倘若预先知道该漏洞属于某一类型时，通过词云图的分析，可以发现在这个类型下分别存在着哪些主要的漏洞问题。如图 9(b)所示，可以发现所有的漏洞都是归属于缓冲区溢出这一类型，但是经过分析之后会发现，其下存在的主要漏洞问题有基于栈的缓冲区溢出漏洞、远程缓冲区漏洞等，这使得研究者可以更加具体地针对漏洞出现的问题来指定解决方案；另外，针对图 9(c)所示的这种无法预知的漏洞类型，在经过模型和词云图软件的分析之后，同样能很快知道互联网中存在的主要问题，这为网络安全的防护措施节约了很多挖掘成本。

总而言之，经过模型训练之后的词向量在通过词云图分析之后，可以清晰地找到存在的主要漏洞，这进一步证明了基于漏洞威胁模式的表示学习能力。此外，通过词云图的直观展示，进一步表明了本文算法能很好地建模并捕获到异构安全实体网络中节点间的繁杂结构和语义关系。

结束语　针对异构安全实体网络的表示学习问题，本文提出了一种基于漏洞威胁模式的表示学习算法。首先，制定一个随机游走的威胁模式构造方案，以此捕获网络中的节点及其上下文结构信息；然后，为了降低模型的时间计算复杂度，采用负采样方法来实现有效的优化。实验结果表明，基于漏洞威胁模式的表示能学习到异构安全实体网络中的节点嵌入，并能够应用于实际的网络挖掘任务。

下一步，计划将该算法应用于更多的应用程序，其中一个方向就是多视图的异构安全实体网络表示。在这样的网络中，每个威胁模式表示了节点间的上下文类型关系，各种威胁

模式的产生则具有了多个视图的网络。

参 考 文 献

[1] YANG P A, WU Y, SU L Y, et al. Overview of Threat Intelligence Sharing Technologies in Cyberspace[J]. Computer Science, 2018, 45(6): 9-18, 26.

[2] LI C, ZHOU Y. Analysis on Threat Intelligence in Big Data Environment[J]. Journal of Intelligence, 2017, 36(9): 24-30.

[3] QIN Y, SHEN G W, ZHAO W B, et al. Research on the method of network security entity recognition based on deep neural network[J]. Journal of Naning University(Natural Science), 2019, 55(1): 29-40.

[4] ZHANG Y C, WEI Q, LIU Z L, et al. Architecture of vulnerability discovery technique for information systems[J]. Journal on Communications, 2011, 32(2): 42-47.

[5] LI J H. Overview of the technologies of threat intelligence sensing, sharing and analysis in cyber space[J]. Chinese Journal of Network and Information Security, 2016, 2(2): 16-29.

[6] TU C C, YANG C, LIU Z Y, et al. Network representation learning: an overview[J]. Scientia Sinica Informationis, 2017, 47(8): 980-996.

[7] GAO H, HUANG H. Deep Attributed Network Embedding [C]// IJCAI. 2018: 3364-3370.

[8] LIU Z M, MA H, LIU S X, et al. A Network Representation Learning Algorithm Fusing with Textual Attribute Information of Nodes[J]. Computer Engineering, 2018(11): 165-171.

[9] YIN B C, WANG W T, WANG L C. Review of Deep Learning [J]. Journal of Beijing University of Technology, 2015, 41(1): 48-59.

[10] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online learning of social representations[C]// Proceeding of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 701-710.

[11] SHI C, SUN Y Z. Research Progress of Heterogeneous Network Representation Learning[J]. Communications of the CCF, 2018, 14(3): 16-20.

[12] SHI C, SUN Y Z, PHILIP S Y. Research Status And Future Development Of Heterogeneous Information Network [J]. Communications of the CCF, 2017, 13(11): 36-42.

[13] WANG X, CUI P, ZHU W W. On the Basic Problems in Network Representation Learning[J]. Communications of the CCF, 2018, 14(3): 12-15.

[14] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in neural information processing systems. 2013: 3111-3119.

[15] SHEN W, HAN J, WANG J, et al. Shine+: A general framework for domain - specific entity linking with heterogeneous information networks[J]. IEEE Transactions on Knowledge Data Engineering, 2018, 30(2): 353-366.

[16] YANG C, LIU M, HE F, et al. Similarity Modeling on Heterogeneous Networks via Automatic Path Discovery[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2018: 37-54.

[17] LIU Y F, LI R F. Graph Regularized Semi-Supervised Learning on Heterogeneous Information Networks[J]. Journal of Computer Research and Development, 2015, 52(3): 606-613.

[18] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks[C]// Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2016: 855-864.

[19] SUN Y, HAN J, YAN X. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003.

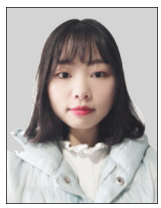
[20] DU Y P, LIU J X, ZHANG J L. Multi-semantic Metapath Based Classification Method in Heterogeneous Information Network [J]. Pattern Recognition and Artificial Intelligence, 2017, 30(12): 1100-1107.

[21] HUANG L W, LI D Y, MA Y T, et al. A Meta Path-Based Link Prediction Model for Heterogeneous Information Networks[J]. Chinese Journal of Computers, 2014, 37(4): 848-858.

[22] DONG Y, CHAWLA N V, SWAMI A. metapath2vec: Scalable representation learning for heterogeneous networks[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM: 135-144.

[23] TANG J, QU M, MEI Q. Pte: Predictive text embedding through large-scale heterogeneous text networks[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1165-1174.

[24] RONG X. word2vec parameter learning explained [J]. arXiv: 1141. 2378.



HUANG Yi, born in 1997, postgraduate, is a member of China Computer Federation. Her main research interests include representation learning and network security.

SHEN Guo-wei, born in 1986, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include cyberspace security and big data.