软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(5):1294−1313 [doi: 10.13328/j.cnki.jos.005954]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于深度学习的安全缺陷报告预测方法实证研究*

郑 炜[1,5,6], 陈军正[1], 吴潇雪[2], 陈 翔[1,4], 夏 鑫[3]

[1](西北工业大学 软件学院,陕西 西安 710072)

[2](西北工业大学 自动化学院,陕西 西安 710072)

[3](Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia)

[4](南通大学 信息科学技术学院,江苏 南通 226019)

[5](空天地海一体化大数据应用技术国家工程实验室(西北工业大学),陕西 西安 710072)

[6](大数据存储与管理工业和信息化部重点实验室(西北工业大学),陕西 西安 710072)

通讯作者: 吴潇雪, E-mail: wuxiaoxue00@gmail.com

摘 要: 软件安全问题的发生在大多数情况下会造成非常严重的后果,及早发现安全问题,是预防安全事故的关键手段之一.安全缺陷报告预测可以辅助开发人员及早发现被测软件中潜藏的安全缺陷,从而尽早得以修复.然而,由于安全缺陷在实际项目中的数量较少,而且特征复杂(即安全缺陷类型繁多,不同类型安全缺陷特征差异性较大),这使得手工提取特征相对困难,并随后造成传统机器学习分类算法在安全缺陷报告预测性能方面存在一定的瓶颈.针对该问题,提出基于深度学习的安全缺陷报告预测方法,采用深度文本挖掘模型 TextCNN 和 TextRNN 构建安全缺陷报告预测模型;针对安全缺陷报告文本特征,使用 Skip-Gram 方式构建词嵌入矩阵,并借助注意力机制对 TextRNN 模型进行优化.所构建的模型在 5 个不同规模的安全缺陷报告数据集上展开了大规模实证研究,实证结果表明,深度学习模型在 80%的实验案例中都优于传统机器学习分类算法,性能指标 $F1$-score 平均可提升 0.258,在最好的情况下甚至可以提升 0.535.此外,针对安全缺陷报告数据集存在的类不均衡问题,对不同采样方法进行了实证研究,并对结果进行了分析.

关键词: 安全缺陷;安全缺陷报告预测;深度学习;文本挖掘

中图法分类号: TP311

中文引用格式: 郑炜,陈军正,吴潇雪,陈翔,夏鑫.基于深度学习的安全缺陷报告预测方法实证研究.软件学报,2020,31(5): 1294−1313. http://www.jos.org.cn/1000-9825/5954.htm

英文引用格式: Zheng W, Chen JZ, Wu XX, Chen X, Xia X. Empirical studies on deep-learning-based security bug report prediction methods. Ruan Jian Xue Bao/Journal of Software, 2020,31(5):1294−1313 (in Chinese). http://www.jos.org.cn/1000-9825/5954.htm

Empirical Studies on Deep-learning-based Security Bug Report Prediction Methods

ZHENG Wei[1,5,6], CHEN Jun-Zheng[1], WU Xiao-Xue[2], CHEN Xiang[1,4], XIA Xin[3]

[1](School of Software, Northwestern Polytechnical University, Xi’an 710072, China)

[2](School of Automation, Northwestern Polytechnical University, Xi’an 710072, China)

[3](Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia)

[4](School of Information Science and Technology, Nantong University, Nantong 226019, China)

* 基金项目: 陕西省工业科技攻关项目(2015GY073); 陕西省重点研发计划(2019GY-057)

Foundation item: Industrial Science and Technology Plan of Shaanxi Province (2015GY073); Key Research and Development Program of Shaanxi Province (2019GY-057)

本文由“系统软件构造与验证技术”专题特约编辑赵永望副教授、刘杨教授、王戟教授推荐.

收稿时间: 2019-08-31; 修改时间: 2019-10-24; 采用时间: 2019-12-24; jos 在线出版时间: 2020-04-07

[5](National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology (Northwestern Polytechnical University), Xi'an 710072, China)

[6](Key Laboratory of Big Data Storage and Management (Northwestern Polytechnical University), Ministry of Industry and Information Technology, Xi'an 710172, China)

Abstract: The occurrence of software security issues can cause serious consequences in most cases. Early detection of security issues is one of the key measures to prevent security incidents. Security bug report prediction (SBR) can help developers identify hidden security issues in the bug tracking system and fix them as early as possible. However, since the number of security bug reports in real software projects is small, and the features are complex (i.e., there are many types of security vulnerabilities with different types of features), this makes the manual extraction of security features relatively difficult and lead to low accuracy of security bug report prediction with traditional machine learning classification algorithms. To solve this problem, a deep-learning-based security bug report prediction method is proposed. The text mining models TextCNN and TextRNN via deep learning are used to construct security bug report prediction models. For extracting textual features of security bug reports, the Skip-Gram method is used to construct a word embedding matrix. The constructed model has been empirically evaluated on five classical security bug report datasets with different scales. The results show that the deep learning model is superior to the traditional machine learning classification algorithm in 80% of the experimental cases, and the performance of the constructed models can improve 0.258 on average and 0.535 at most in terms of $F1$-score performance measure. Furthermore, different re-sampling strategies are applied to deal with class imbalance problem in gathered SBR prediction datasets, and the experiment results are discussed.

Key words: security bug; security bug report prediction; deep learning; text mining

近年来,互联网技术的快速普及,在推动各行各业快速发展的同时,也助长了各种网络安全事件的发生[1].国际权威漏洞发布组织的 CVE(common vulnerabilities and exposures)[2,3]数据显示,近年来,系统安全漏洞数量急速上升:2016 年发现的安全漏洞为 6 447 个,2017 年发现的安全漏洞为 14 714 个,2018 年发现的安全漏洞则达到 16 555 个.软件安全漏洞也称为安全缺陷,是软件缺陷的一种类型,这类缺陷的存在可能会导致系统被恶意攻击者利用,从而造成信息泄露、篡改、系统瘫痪等安全问题.缺陷报告记录软件开发、运行以及维护过程中发现的各种问题,是缺陷调查与再现的主要依据.随着软件漏洞数量的增加以及安全漏洞类别的衍变,手动对缺陷报告进行分类变得越来越困难[4].除了需要一定的人力以外,手动缺陷报告分类还要求分类人员具有较为丰富的安全领域知识.已有研究指出[4−6]:实际项目中,许多安全缺陷报告往往由于开发人员或者测试人员安全知识缺乏等因素,未被及时标记为安全相关缺陷报告,从而长期滞留在缺陷报告库中,直到安全事件发生并造成严重的后果.针对此问题,研究人员尝试采用机器学习等方法进行安全缺陷报告自动预测[4−8].

安全缺陷报告预测是指通过机器学习等方法从大量的缺陷报告库中识别出与安全相关的缺陷报告.给定一个缺陷报告(bug report,简称 BR),如果该 BR 与安全相关,则称其为安全缺陷报告(security bug report,简称 SBR);如果 BR 与安全无关,则称其为非安全缺陷报告(non-security bug report,简称 NSBR).从大量的 BR 中识别出 SBR 这一问题,可以被形式化建模为二分类问题.基于该思想,研究人员开展了一系列研究工作.例如,Gegick 等人[6]将文本挖掘与分类算法相结合,采用朴素贝叶斯(Naïve Bayes)、k 最近邻(k-nearest neighbor)等分类算法对来自思科的实际工程项目的缺陷报告进行预测;Peter 等人[5]针对安全缺陷预测中样本不均衡问题,给出了一种噪音数据过滤框架 FARSEC,该框架可以用于过滤掉 NSBR 中与安全关键词相似度较高的样本(即缺陷报告);在此基础上,Shu 等人[7]进一步采用 SMOTE 进行样本不均衡问题处理,并采用超参优化方法 SMOTUNED 进行参数优化.这些研究主要采用分类算法与文本挖掘技术相结合的方法,通过提取 BR 中与安全相关的关键字和词频等特征,然后用于训练分类模型并对大量未标记样本进行预测.这些研究成果为 SBR 识别奠定了基础.但是,已有研究提出的方法在预测性能上还存在较大的提升空间.例如,Peter 等人[5]所给出的实验结果中,在性能度量指标 G-measure(论文主要采用的性能度量指标)最优得分为 53.8%~71.9%的情况下,其查全率仅为 47.6%~66.7%,而误报率却高达 3.0%~41.8%.

传统的机器学习分类算法对特征工程较为敏感,由于安全缺陷的复杂性和稀有性,缺陷报告中安全相关特征提取困难,从而导致模型预测性能难以提升.基于神经网络的深度学习技术在许多任务中被证明优于传统的

© 中国科学院软件研究所 http://www.jos.org.cn

机器学习方法,已在计算机视觉、推荐系统、机器翻译等领域内的大量任务中取得较大研究进展[9−12].据我们所知,本文首次将深度学习引入到 SBR 预测问题中,基于 BR 的文本描述信息来进行 SBR 预测,并随后使用对深度学习在 SBR 预测的效果进行实证研究.通过经典文本分类模型 TextCNN(convolutional neural networks for sentence classification)[13]和 TextRNN(recurrent neural network for text classification)[14]构建 SBR 预测模型,并采用注意力机制(attention mechanism)[15]对 TextRNN 模型进行优化.同时,就深度学习模型对 SBR 预测的性能以及不同的采样策略对预测结果的影响进行了实证研究.

我们将基于 TextCNN、TextRNN 以及注意力机制构建的深度学习模型应用于 4 个小规模数据集(每个数据集包含 1 000 条 BR)和 1 个大规模数据集(包含 41 298 条 BR),并与传统的分类算法进行比较.结果表明,本文构造的基于 TextCNN 和 Attention+TextRNN 的深度学习模型对 SBR 预测能力在 80%的情况下,其性能要显著优于 4 种传统的机器学习模型——朴素贝叶斯、逻辑回归、k 最近邻、随机森林,性能指标 $F1$-score 平均可提升 0.258,最多时甚至可以提升 0.535.

本文的主要贡献可总结如下.

(1) 本文提出了基于深度学习的 SBR 预测方法.根据 SBR 文本特征,基于经典文本分类方法 TextCNN 以及 TextRNN 与注意力机制相结合构建 SBR 预测模型.

(2) 基于实际开源项目验证了本文所提方法的有效性.实证研究共分析了来自 Ambari、Camel、Derby、Wicket 和 OpenStack 这 5 个项目累计 45 298 个 BR.最终结果表明,本文所提方法的性能要显著优于传统的机器学习分类方法.

(3) 论文深入分析了类不均衡问题、不同词嵌入方法对模型性能的影响,从而为如何有效使用本文所提方法提供了指导.

本文第 1 节介绍安全缺陷预测的研究背景和相关研究工作.第 2 节介绍深度学习模型构建具体过程和方法细节.第 3 节介绍论文的实证研究,包括研究问题、评测对象、评测指标、实验流程以及关键参数设置.第 4 节对实证研究结果进行详细分析和总结.最后总结全文,并对下一步研究工作进行展望.

1 相关工作

随着软件系统安全问题的日益突出,SBR 预测也受到越来越多的关注[4−8].软件项目一般使用缺陷跟踪系统(如 JIRA、LaunchPad、Bugzilla 等)对项目开发和运行中产生的缺陷进行记录和跟踪.在长期积累的大量 BR 中,有些 BR 已被明确标记为 SBR(比如与 CVE 数据相关的缺陷),这些缺陷报告对于识别新的 SBR 具有非常重要的参考价值.本文主要研究如何使用深度学习进行 SBR 文本特征挖掘,从而更高效、更准确地完成 SBR 预测.本节将从安全缺陷报告预测和基于深度学习的文本分类这两个角度对已有工作进行分析.

1.1 安全缺陷报告预测

SBR 预测是及早发现系统安全缺陷的关键方法,已成为当前软件工程领域一个研究热点[4−8].安全缺陷预测可以尽早地从缺陷库中的大量缺陷报告中发现隐藏的安全缺陷,因此开发团队可以更为合理地分配资源,使得安全问题尽早得以解决,从而有效避免后期可能造成的重大危害.近些年来,研究人员针对该问题提出了多种解决方法.

Gegick 等人[6]首先提出将文本挖掘的方法应用于 SBR 描述信息分析,通过已标记的 SBR 的 term-by-document 变量矩阵对 SAS 文本挖掘器进行训练,然后对未知样本(即未标记缺陷报告)进行预测,并对来自思科实际工业项目中收集的缺陷报告数据集进行了实际验证.随后,Wijayasekara 等人[4]对 Linux 内核和 MYSQL 这两个项目在 2006 年~2011 年期间暴露的安全漏洞进行了分析,指出:Linux 中 32%的 SBR 和 MySQL 中 62%的 SBR 都属于隐藏 SBR,在缺陷数据库中未能被及时识别出,而且这一数字在接下来的两年仍有上升趋势.他们同时给出了基于文本分类的隐藏 SBR 的识别方法.

最近,Peter 等人[5]提出了一种噪音数据过滤方法 FARSEC.该方法首先使用 TF-IDF(term frequency-inverse document frequency)从 SBR 提取安全相关关键词,然后计算 NSBR 与安全关键词表之间的相似度得分,通过得

© 中国科学院软件研究所 http://www.jos.org.cn

分排序,将 NSBR 中与安全关键词表相似度较高的样本筛除.该方法一方面可以移除掉可能标记错误的样本,另一方面可以对 NSBR 进行欠采样处理,从而有效减小正负样本间的数量差距.FARSEC 在 4 个小规模数据集(Ambari、Camel、Derby 和 Wicket)和 1 个大规模数据集(Chromium)上进行了实验验证,结果表明,FARSEC 可以在一定程度上提高 SBR 预测性能.随后,Shu 等人[7]通过 SMOTUNED[16]超参优化方法对其进行改进,并对分类算法参数优化和数据预处理方法参数优化对 SBR 预测性能的影响进行实验对比,结果表明,数据预处理方法(即类不均衡处理方法 SMOTE)的参数优化要比分类算法参数优化对预测结果的影响更大,更有助于提高 SBR 预测的查全率.

与已有 SBR 预测研究工作不同,本文首次从深度学习入手,基于面向自然语言处理的经典文本分类模型 TextCNN 和 TextRNN 构建 SBR 预测模型.

1.2 基于深度学习的文本分类

文本分类属于自然语言处理的一个分支,随着深度学习在自然语言处理中的长足发展和不断应用,基于深度学习的文本分类技术也取得了显著进展[17].基于深度学习的文本处理模型包括 TextCNN、TextRNN、FastText、TextRCNN、HAN 等.基于深度学习的文本分类中,一个非常重要的问题就是确定分类的网络.某些情况下,分类的文本包含多种语言,如果针对每一种语言都训练一个网络,会让问题变得更为复杂,并降低了网络的可迁移性.Yang 等人提出了一种基于分词注意力模型 HAN(hierarchical attention networks for document classification),可以有效解决上述问题.当样本数据量较小或者正负样本数差距较小时,传统的卷积神经网络模型会因为得不到足够的样本训练或者池化层丢失了大量信息导致对输入变化不敏感,胶囊网络有效地解决了这些问题[18,19].胶囊网络通过基于挤压函数的卷积来感受和传递特征,并用胶囊(向量)来描述,使得比传统卷积神经网络获取和保留了更多的信息.

词嵌入算法是深度学习在文本处理中的一个核心,良好的词嵌入算法可以很好地反映词语的相似关系,而这些关系又很难被传统的文本分类网络学习到,所以良好的词嵌入算法可以给模型性能带来大幅度提升.在确定词向量的维度时,大部分研究工作偏向于经验主义,因此缺乏理论依据.Yin 等人[20]提出了 PIP(pairwise inner product)损失来描述词嵌入矩阵的维度理想度,并通过实验进行了验证.较早的词嵌入算法是 Mikolov 等人[21]提出的 word2vec,他们提出两个新的模型:连续词袋模型(continuous bag of words,简称 CBOW)和连续 skip-gram 模型来训练词向量.实验结果表明,相对于基于潜在语义(latent semantic analysis,简称 LSA)具有更好的表现;而相对于潜在狄利克雷分配(latent Dirichlet allocation,简称 LDA)[22]算法,word2vec 方法可以大幅度减少工作量.但是在后续的使用中,研究人员发现 word2vec 依赖于大量的文本数据,在较小的数据集上一般表现不佳.因此,Bojanowski 等人提出了基于 word2vec 方法的字符级 *n*-grams[23],将每个单词表示为字符的 *n*-gram.每个字符和一个向量相关联,而单词则表现为这些向量的和.该方法的优点是对于训练中未出现的单词也能计算出比较好的表示.他们在 9 种不同语言的数据集上测试,证明这种方法达到了当时最好的性能.Pennington 等人[24]则提出了 Glove 方法,他们关注于词和词之间共现的统计数据,核心思想是:两个词共同出现的频率越高,则这两个词的相关性越高.即如果词 a 和词 b 的关联度大于词 a 和词 c,则应有 $P_{ab}>P_{ac}$,其中,P 代表一定距离内(窗口词)两个词共同出现的概率.但是有时候,同一个词在不同的语句中有着不同的意义(即多义词),因此,Peters 等人提出了 ELMo(embeddings from language models)[25],给出了预训练一个双向的语言模型,再根据语言模型的每一层计算出每个单词的词向量.

ELMo 的成功,让研究人员意识到,大家发现了在非任务大语料文本中预训练的重要性.随后,研究人员设计出两个重要网络:一个是 BERT[26],另外一个是 GPT[27].GPT 使用 Transformer[28]代替传统的 LSTM 网络,并且使用了非监督预训练和监督微调相结合的方式,在多项任务中(包括文本分类任务)达到了当时最高水平.并且该论文通过对比证明了预训练对分类任务有着显著的提升.GPT-2 作为一个通用的语言模型[27],在高达 800 万个网页上训练而成,在很多任务上达到了专用模型的水平,并且部分任务的表现已经达到了人类的水平.可见,在大量文本上预训练可能是未来的趋势.Google 提出的 BERT 是一个基于微调的表示模型,同样也使用了 transformer,并且使用上下文双向预测[29]来预训练一个语言模型(masked language model,简称 MLM),其刷新了

© 中国科学院软件研究所 http://www.jos.org.cn

11 项 NLP 问题的记录.此外,数据类别不均衡是影响分类准确率的因素之一[30,31].Huang 等人对经典深度学习分类模型在不均衡数据集上的表现进行了实验验证,表明,通过实施深层网络来维持集群间和边际,可以学习更具辨别能力的深层模型.Song 等人[32]提出基于聚类的双向采样方法来进行数据类别不均衡问题处理.

本文将深度学习引入 SBR 预测问题,根据 SBR 文本描述信息差异性大、安全特征稀缺的特征、SBR 预测数据集中正样本(SBR)较少导致类别严重不均衡(例如在 Peter 等人[5]搜集的 5 个数据集中,其正样本所占比例为 0.5%~9.0%)等问题,采用 word2vec 的 skip-gram 模型(本文将窗口大小设置为 3)来训练词向量,通过设置全连接层偏置项(bias)提高模型泛化能力;同时,结合实际项目经验进行句长设置,构建出符合 SBR 文本特征的分类模型.

2 基于深度学习的安全缺陷报告预测方法

2.1 方法的整体框架

基于深度学习的 SBR 预测方法框架如图 1 所示,其共包含 4 个阶段.

(1) 阶段 1:执行 BR 数据集的收集与划分,具体细节见第 2.2 节.

(2) 阶段 2:执行文本预处理.该阶段将文本数据预处理成数字序列,深度学习模型可以方便地根据序列获得到单词的表示(词向量),具体细节见第 2.3 节.

(3) 阶段 3:进行深度学习模型的构建.我们一共选择了两种适合文本挖掘的深度学习模型,分别是 TextCNN 和 TextRNN[33].这两个模型分别衍生自经典的 CNN 模型和 RNN 模型,选用这两个具有代表性的模型可以确保我们实证研究的结论更有代表性.具体细节见第 2.4 节.

(4) 阶段 4:借助已经构建的深度学习模型,从未标记的缺陷报告中预测中 SBR.

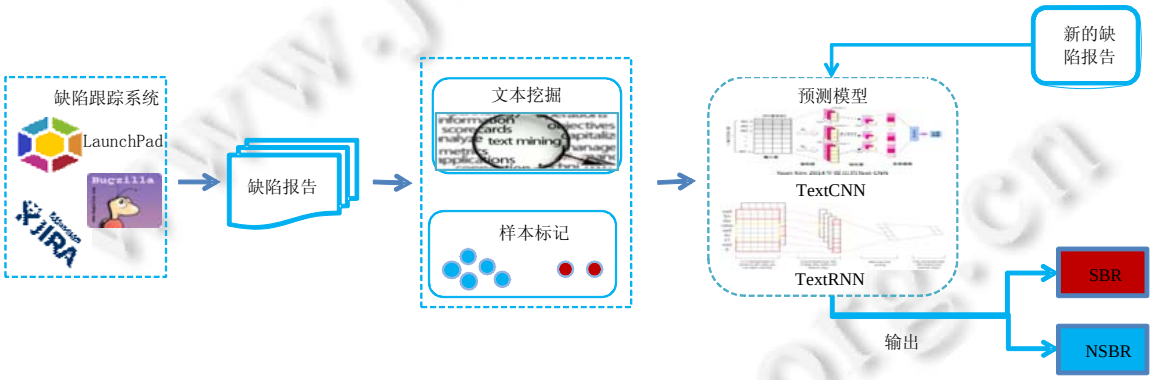


Fig.1 Framework of deep-learning-based security bug report prediction

图 1 基于深度学习的安全缺陷报告预测方法框架

2.2 阶段I:缺陷报告的收集

项目开发和运维过程中产生的缺陷,都会使用特定的缺陷跟踪系统(例如 JIRA、Bugzilla、LaunchPad 等)进行管理.BR 从多个维度对所发现的问题进行记录,表 1 是 LaunchPad 系统中 OpenStack 项目的一个 BR 示例,该示例展示了 BR 中的关键字段信息(在 LaunchPad 中,OpenStack 项目使用 Tag 标识特殊的 BR 类型,如 security, performance 等;同时,由于 Description 信息较长,本示例只显示了前面一部分文字信息).在 BR 的多个字段中,缺陷描述信息(bug description)是对缺陷最详尽的描述,一般包括缺陷触发的条件、问题现象、预期结果等文本描述信息.如其他 SBR 预测的研究[5,7]一样,本文也使用 BR 的描述信息进行 SBR 预测.虽然开源缺陷跟踪系统收集了大量的 BR,可以提供丰富的数据资源,但是提交缺陷报告的人员质量和知识背景差异较大,因此在收集数据集过程中,我们需要对一些低质量缺陷报告(比如没有缺陷描述信息的)进行过滤.Chaparro 等人[34]开发出一种过滤信息不完整缺陷报告的工具,该工具认为,缺陷报告描述信息中,OB(observed behavior) 、S2R(steps to

© 中国科学院软件研究所 http://www.jos.org.cn

reproduce)和 EB(expected behavior)是缺陷报告描述的基本内容,如果这三者缺少任何一个,则认为该缺陷报告的描述信息不完整,因此可以通过自然语言处理和解析方式自动进行判断和过滤.本文考虑的数据集一部分来源于已有的广泛使用的安全缺陷报告数据集,另一部分是在作者前期工作[35]的基础上进一步完善得到的 OpenStack 安全缺陷预测数据集.在该数据集的收集过程中,我们同样采用了 Chaparro 等人[34]的方法对描述信息不完整的缺陷报告进行了过滤.

Table 1 A bug report example from OpenStack project

表 1 来自 OpenStack 项目中的一个缺陷报告示例

字段	内容
Key	1218977
Title	DOS by passing an ephemeral or swap of arbitrary size
Status	Fix released
Importance	Critical
Tag	Security
Description	Due to a previous bug that was never caught and the fact that we can now pass ephemeral and block devices through the API, it is possible to ask nova to create an arbitrarily large ephemeral block device - which nova will happily do (and by default make it raw). The bug was introduced in commit 0ef7e15e225efcce3e02098cb1d57f9f40181f82 as before that commit the ephemeral device size will be defaulted to whatever was in the instance_type - due to a bug this defaulting was not done anymore (see compute.api.API._update_block_device_mapping). Steps to reproduce: ndipanov@localhost devstack]$ nova flavor-show 1 …

2.3 文本预处理

本文在文本预处理时依次执行如下 4 个步骤.

(1) 步骤 1:移除停用词.

停顿词是指对问题分析没有实际价值的一些词语,如“this”“that”,“we”,“then”,“a”等.移除停顿词一方面可以降噪,另一方面可以减小数据量,降低计算开销和空间开销.我们首先使用 NLTK[36]提供的通用的停用词集.但是 NLTK 提供的默认停顿词对于缺陷报告问题的分析并不完善,我们通过人工分析,在此基础上进一步对本文所使用的停顿词进行了补充.

(2) 步骤 2:进行句子填充(sentences padding).

深度学习模型在进行文本数据处理的时候,需要执行句子填充操作(即统一句子长度).一般文本处理采用最长句子的长度,但是由于 BR 的描述信息长度差异较大,如果全部依据最长句子进行填充,一方面会大幅度增加模型的空间开销和时间开销,另一方面也会引入大量噪音.因此,我们根据实际预测效果,选择平均长度的 1.5 倍作为最大句子长度,其计算公式为

$$L=ceil(L_{avg}\times 1.5) \tag{1}$$

其中,L_{avg} 代表句子平均长度,$ceil$ 代表向上取整.长度大于该计算结果的句子会被按序截取,即第 L 个词以后的都会被截断.长度不够的句子会自动补充无意义符号(如 0),使其长度达到 L.

(3) 步骤 3:建立词汇-索引映射表.

词表是一个包含所有出现过的词和一个特殊符号的集合,每种出现过的词和词表的一个元素有着唯一的对应,并且所有未出现的词都将会和词表中的特殊符号对应.

(4) 步骤 4:建立词嵌入(word embedding)矩阵.

词嵌入矩阵是一个 $S\times d$ 大小的矩阵,其中,S 是词表中的元素个数,d 是每个词的词嵌入向量的维度.矩阵的每一个行向量都代表了一个词,并且按顺序和词表中的每一个词相对应.Word2vector 提供 skip-grams 和 CBOW 两种词嵌入矩阵构建方法,本文中,我们使用 skip-grams 方法,并根据安全缺陷文本特征设定窗口大小为 3,完成词嵌入矩阵的构建.

这样,我们数据集中的每一条记录的描述信息都成为了长度为 L 的向量.所有的缺陷描述构成了一个 $T=L\times N$ 的矩阵,其中,L 是句子长度,N 是样本数.样本的标签构成了一个长度为 N 的列向量 La.设存在一个理想

© 中国科学院软件研究所 http://www.jos.org.cn

模型 f,则有 $La=f(T)$.

2.4 深度学习模型

本文主要考虑了两种深度学习模型 TextCNN 和 TextRNN,并采用注意力机制对模型进行优化.

2.4.1 TextCNN

TextCNN 是一个由经典的卷积神经网络(CNN)衍生出来的用于文本分类的神经网络,其结构如图 2 所示.

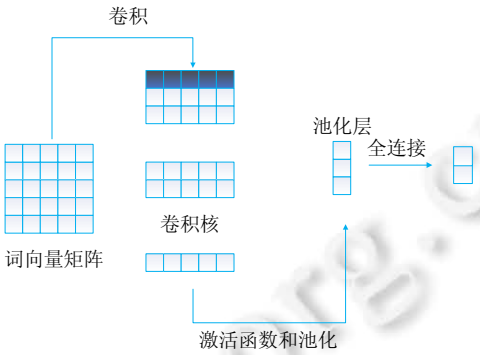


Fig.2 Network Structure of TextCNN

图 2 TextCNN 网络结构

模型先在词向量矩阵上用不同大小的卷积核卷积,每个卷积核类似于一个可训练的滤波器,提取输入的部分特征,然后将卷积结果经过激活函数(这里我们用的是 ReLU)处理后输入到池化层,池化层会生成更高维的向量.然后将池化层的结果通过 dropout 层随机选择后(这样可以一定程度地避免过拟合)通过全连接层连接,最后得到分类结果.

CNN 的空间问题复杂度如下所示,其与输入数据本身的大小无关,主要涉及模型的参数数量和每层输出的特征图大小.

$$Space \sim O\left(\sum_{l=1}^{D} K_l^2 \times C_{l-1} \times C_l\right) \tag{2}$$

其中,K 为卷积核的尺寸,C 为通道数,D 为网络的深度.

针对安全缺陷报告本身特征,我们对模型做了一些针对性的设计.首先,我们设置了多个卷积核,大小分别是 $i\times d$,i=1,2,3,其中,d 是每个词的词嵌入向量的维度.这种处理方式可以把卷积视野抽象成 i(i=1,2,3)个单词,其核心公式为 $y=f(C_1+C_2+C_3)$,其中,C_1,C_2,C_3 是不同大小卷积核的输出结果.

2.4.2 TextRNN

TextRNN 是本文选择的另一种深度学习网络模型,也是经典的用于文本处理的深度学习模型.基于 RNN 的基本思想是利用给定序列中存在的信息(如文本):给定一系列单词,将单词进行数字化表示(GloVe 或 Word2Vec)馈送到神经网络并计算输出.在计算下一个单词的输出时,还会考虑前一个单词的输出.RNN 之所以被称为循环,是因为它们使用先前计算的输出对序列的每个元素执行相同的计算.在任何步骤,RNN 执行以下计算.

$$RNN(t_i)=f(W\times x_{t_i}+U\times RNN(t_{i-1})) \tag{3}$$

其中,W 和 U 是模型参数,f 是非线性函数.$RNN(t_i)$是第 i 个 timestep 的输出,可以按原样使用,也可以再次输入参数化结构,如 softmax[37],具体取决于当前执行的任务.训练是通过根据全部或部分 timestep 的输出制定损失目标函数,并尽量减少损失来完成的.我们使用基于 BiLSTM(bi-long short-term memory)的 TextRNN 网络,通过在经典的 RNN 上加入 3 扇门(即输入门、输出门和遗忘门)而实现,其结构如图 3 所示.

输入门、输出门和遗忘门可以让信息选择性的通过,各层主要通过一个 sigmoid 的神经层和一个逐点相乘的操作实现.其中,sigmoid 层的输出是一个向量,向量的每个元素取值介于[0,1]之间,表示让对应的信息通过的

© 中国科学院软件研究所 http://www.jos.org.cn

权重(例如,0 表示不允许通过,1 表示允许所有信息通过).

输入门:决定下一步允许多少新的信息加入到 cell 状态中,由“input gate layer”的 sigmoid 层决定哪些信息要更新,其公式为

$$i_t=\sigma(W_i\times[h_{t-1},x_t]+b_i) \tag{4}$$

其中,σ是激活函数,h_{t-1} 是前一时刻隐层状态,x_t 是当前时刻的输入,W_i 和 b_i 是线性关系的系数.

输出门:确定最终输出值.输出状态是一个过滤后的状态.运行一个 sigmoid 层来确定细胞状态的哪个部分将会输出,然后把细胞状态通过 tanh 进行处理,得到一个[−1,1]之间的值,并将它和 sigmoid 门的输出相乘,最终仅仅会输出确定要输出的部分:

$$h_t=\sigma(W_o[h_{t-1},x_t]+b_o)\times\tanh(C_t) \tag{5}$$

其中,σ是激活函数,h_{t-1} 是前一时刻隐层状态,x_t 是当前时刻的输入,W_o 和 b_o 是线性关系的系数.

遗忘门:LSTM 第 1 步决定从细胞中遗忘的信息内容,通过遗忘门来实现.该门会读取 h_{t-1} 和 x_t,输出一个[0,1]之间的数值给每个在细胞状态:

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{6}$$

其中,W_f 和 b_f 是线性关系的系数,σ是激活函数,h_{t-1} 是前一时刻隐层状态,x_t 是当前时刻的输入.

LSTM 有多个变种[37],有着不同的适应场合和性能.从图 3 中可以看到,我们将词向量输入到双向的 LSTM 中,然后将结果进行连接,最后通过全连接层(full connection,简称 FC)得到分类结果.我们使用双向 LSTM 的原因是,双向 LSTM 可以有效缓解越往后的单词对结果影响越大的问题.

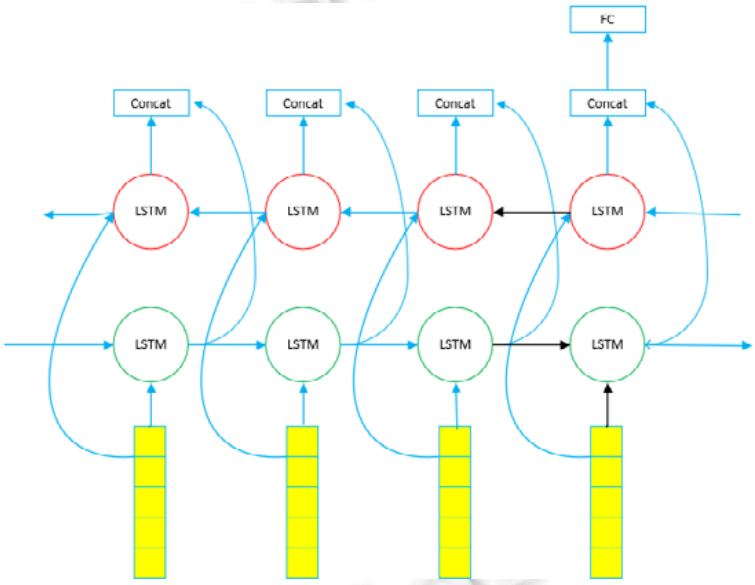


Fig.3 Structure of TextRNN

图 3 TextRNN 的结构图

2.4.3 注意力机制

本文进一步借助注意力机制对 TextRNN 模型进行优化.SBR 的识别主要依赖 BR 中的描述信息(description),采用自然语言描述形式,其安全相关文本特征较为稀疏,给 SBR 预测准确性带来一定困难.注意力机制源于对人类视觉的研究,使训练重点集中在输入数据的相关部分,忽略无关部分[12,15].注意力机制最早由 Bahdanau 等人[15]提出,并应用于机器翻译任务中.其基本思想是,基于 encode-decode 模式,首先定义 RNN 模型中的条件概率,其计算方法如公式(7)所示.

$$p(y_i|y_1,\ldots,y_{i-1},x)=g(y_{i-1},s_i,c_i) \tag{7}$$

© 中国科学院软件研究所 http://www.jos.org.cn

其中,S_i 为 RNN 在 i 时刻的隐藏状态,其计算方法如公式(8)所示.

$$S_i=f(s_{i-1},y_{i-1},c_i) \tag{8}$$

c_i 为上下文向量,其计算方法如公式(9)所示.

$$c_i=\sum_{j=1}^{T_x}a_{ij}h_j \tag{9}$$

a_{ij} 为每个注解 h_j 的权重,计算方法如公式(10)所示.

$$a_{ij}=\frac{\exp(e_{ij})}{\sum_{k=1}^{T_x}\exp(e_{ik})} \tag{10}$$

其中,$e_{ij}=a(s_{i-1},h_j)$是所得到的对齐矩阵,用于记录位置 j 的输入和位置 i 的输出的匹配程度.简单而言,注意力机制把源语言端的每个词学到的表达和当前要预测的词联系了起来,在模型训练好后,根据注意力矩阵得到源语言和目标语言的对齐矩阵.

3 实验设计

为了支持我们的结论,我们设计了如下 4 个研究问题(research question,简称 RQ)来指导我们的实验设计.

- RQ1:深度学习方法与已有基准方法相比是否可以提升安全缺陷报告预测的性能?
- RQ2:针对安全缺陷报告预测问题,不同的正负样本比例对深度学习模型性能的影响如何?
- RQ3:不同的深度学习模型的迭代表现是否相似?
- RQ4:词嵌入方法对基于深度学习的安全缺陷预测方法的性能影响程度如何?

随后,该节对实证研究中所需要的数据集、性能评价指标进行介绍,最后给出实验流程及关键参数的具体设定.

3.1 评测对象

本文在实证研究中使用了两部分数据集.

第 1 部分是 Ohria 等人[38]提供的 4 个小规模数据集,分别来自 4 个不同的项目:Ambari[39]、Camel[40]、derby[41]和 Wicket[42].这 4 个项目都来自 Apache,并使用 JIRA 作为缺陷跟踪系统.Ambari 给 Apache Hadoop[43]集群提供管理服务和可视化的用户接口.Camel 简化了特定领域语言的路由和中介规则的建立.Derby 是一个完全由 Java 编写的轻量级的数据库.Wicket 是一个基于 Java 的 web 开发框架,这 4 个项目代表 4 类典型应用系统.Ohria 等人[38]从 4 个项目中各选择了 1 000 个具有较大影响的缺陷,并采用人工检查和确认的方式进行了标注,标注类型包括 6 类:surprise、dormant、blocking、security、performance 和 breakage.Peter 等人[5]在此基础上根据 security 标签提取了 4 个用于安全缺陷预测的数据集,并将其公开.因此,本文直接使用了他们处理之后的 4 个小规模数据集.

本文使用的另外一个数据集是我们自己基于 OpenStack 项目缺陷报告构建的一个大规模数据集,包含 41 298 多个来自 OpenStack 项目的缺陷报告.Openstack 是一个由 NASA 和 Rackspace 合作发起的开源云管理平台[44],包含多个用于构建云平台的组件设施(如 Nova、Neutron、Swift、Cinder 等).我们在之前的研究工作中[35]对 OpenStack 的 41 298 个缺陷报告从不同维度进行了深入分析,包括缺陷类型分布、缺陷、解决措施等.本文是在原有工作基础上,进一步结合 OpenStack 的 CVE 数据进行了安全样本标记,从而构建了用于安全缺陷报告预测的 OpenStack 数据集.OpenStack 的缺陷报告使用 LaunchPad[45]进行跟踪管理,至今已积累了超过 14 万条缺陷报告.在之前的研究工作[35]中,我们确定了如下的缺陷报告筛选条件.

- Importance 字段仅考虑 Critical、High 和 Medium.
- Status 字段仅考虑 Fix Committed 和 Fix Released.
- Report data 字段的时间要早于 2018 年 5 月 1 日.

我们采用这些筛选条件,一方面因为符合这些条件的数据在实际项目中开发人员的关注度更高;另一方面,

© 中国科学院软件研究所 http://www.jos.org.cn

这些缺陷报告都是已修复完成的,并且经过了分析过程,其结论可信度较高.根据这 3 个条件,总共检索到 48 255 条缺陷报告.在此基础上,我们采用 Chaparro 等人[34]提供的工具对描述质量较低的缺陷报告进行了进一步过滤,最终得到 41 298 条符合条件的缺陷报告.最后,在此基础上,根据以下两条规则进行安全相关缺陷报告的标记.

规则 1. BR 与 CVE 记录相关[3].

CVE 是国际权威漏洞跟踪与发布组织,OpenStack 系统目前在 CVE 系统中约有 180 条记录.为了便于了解漏洞详细记录,对于开源系统,CVE detail[3]网站会将每个漏洞所对应的原始缺陷报告进行关联.因此,当 OpenStack 在 LaunchPad 中的缺陷报告与 CVE 数据关联时,则表明该缺陷是安全相关缺陷.

规则 2. 缺陷报告在 LaunchPad 系统中,本身被标记了 security 的 label.

在 LaunchPad 系统中,OpenStack 的 SBR 通过将“label”字段设置为“security”来进行标记的,因此,本文将该部分 BR 也作为 SBR 的一部分.

当满足这两条规则的任意一条时,我们认为这个缺陷报告是安全相关缺陷报告.最终,本文实证研究采用的数据集的统计信息见表 2,主要包括数据集的名称、安全相关的缺陷报告数和安全无关的缺陷报告数.

Table 2 Statistics of security bug report prediction datasets

表 2 安全缺陷报告预测数据集的统计信息

数据集名称	安全相关的缺陷报告数	安全无关的缺陷报告数
Ambari	29	971
Camel	32	968
Derby	88	912
Wicket	10	990
Openstack	239	41 059

3.2 评测指标

给定测试集中的任意一个 BR,其可能的结果有 4 种:TP(true positive)、TN(true negative)、FP(false positive)、FN(false negative),反映了预测结果和实际结果的关系,可用混淆矩阵来表示,具体见表 3.

Table 3 Confusion matrix of security bug report prediction

表 3 针对安全缺陷报告预测问题的混淆矩阵

实际结果	预测结果	
	SBR	NSBR
SBR	TP	FN
NSBR	FP	TN

基于上述混淆矩阵,我们依次介绍本文使用的评测指标.

准确率(accuracy):反映了模型预测的正确率,其计算公式如下.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

但是,在类别不均衡的问题中使用 *accuracy* 可能无法客观反映模型性能.假如某个模型在大部分情况下都将少数类识别为多数类,但是因为少数类占比过少,所以即使准确率很高,这样的模型在实际任务中很难对少数类做出精确的预测.

查全率(recall):预测为 SBR 的样本数量占实际 SBR 样本数量的比率,其计算公式如下.

$$recall = \frac{TP}{TP + FN} \tag{12}$$

查准率(precision):预测正确的 SBR 数量占预测为 SBR 的数量的比率,其计算公式如下.

$$precision = \frac{TP}{TP + FP} \tag{13}$$

*F*1-score:查全率和查准率可以说是一对相反的指标:查全率高,说明模型更加倾向于将一个样本认为是正样本;而查准率高,则说明模型在预测为 SBR 时更加谨慎.*F*1-score 则综合考虑查全率和查准率,其计算公式

© 中国科学院软件研究所 http://www.jos.org.cn

如下.

$$F1\text{-}score=\frac{2\times precision\times recall}{precision+recall} \tag{14}$$

3.3 统计分析方法

本文使用 Wilcoxon rank-sum 检验[46]来分析深度学习方法与 4 种传统分类器之间在 SBR 预测上性能是否具有统计差异性.Wilcoxon rank-sum 是一个非参假设检验方法,用来检测两个数据集是否来自于相同分布的总体.对于相比较的两个数据集,零假设是两种方法之间不存在显著差异.如果由 Wilcoxon rank-sum 检验得到的 p-value 小于显著性水平,则拒绝原假设.也就是说,两种方法之间的差异被确定为具有统计显著性.本文使用 0.05 作为显著性水平,当 p-value<0.05 时,认为两组数据集具有统计显著性[10].本文计算深度学方法与传统分类方法性能指标 $F1$-score 值的 Cliff's delta 来度量二者之间差异性.Cliff's delta 计算公式如下.

$$d=2W/mn-1 \tag{15}$$

其中,W 是 Wilcoxon rank-sum 检验统计值,m 和 n 分别为两组参与对比数据集的大小.$W=R-n(n+1)/2$,其中,R 为 Cliff's delta 值的范围.Cliff's delta 的级别和取值范围见表 4,总共分为 4 个级别.

Table 4 Level and range of Cliff's delta[46]

表 4 Cliff's delta 的级别与取值范围对应关系[46]

级别	Cliff's delta 的取值范围
Negligible	$\lvert d\rvert<0.147$
Small	$0.147\leqslant\lvert d\rvert<0.33$
Medium	$0.33\leqslant\lvert d\rvert<0.474$
Large	$0.474\leqslant\lvert d\rvert$

3.4 基准方法

为了验证深度学习对于安全缺陷报告预测的性能表现,我们将本文构建的深度学习预测效果与 Peter 等人[5]近期发表于 TSE 的方法 FARSEC 进行对比.该方法使用文本挖掘技术与传统分类算法相结合进行安全缺陷报告识别,并给出一个框架 FARSEC 用于噪音数据过滤.该方法首先使用特征提取方法 tf-idf 从安全缺陷报告中识别与安全相关的关键词;之后,通过计算缺陷报告与安全关键词之间的相似度得分来对缺陷报告进行排名,取出排名最高的缺陷报告为噪音数据,将其从数据集中删除.在实验验证中,其所选定的关键词数量设置为 100 个,阈值 0.75 用于过滤噪声错误报告的错误报告得分.他们评估了 FARSEC 对 5 个数据集(Ambari、Camel、Derby、Wicket 和 Chromium)的有效性.该方法所使用到的分类算法包括朴素贝叶斯、逻辑回归、k 最近邻、随机森林.

(1) 朴素贝叶斯.朴素贝叶斯是基于贝叶斯法则的概率模型,其假设各个特征之间相互独立,所以逻辑简单、计算开销小.但是有时候变量之间并不是完全随机的,因此会损失一部分准确度.

(2) 逻辑回归(logistics regression,简称 LR).逻辑回归在线性回归模型的基础上经过激活函数(非线性函数,一个典型的例子就是 sigmod),使得回归的结果变成了分类结果.

(3) k-近邻.k-近邻算法基于一个简单的假设:如果一个样本的 k 个最相似的样本的大多数属于某一类别,则这个样本也属于那一类别.

(4) 随机森林(random forest,简称 RF).随机森林是多棵分类树得到的,其中,每棵分类树都由随机的样本和随机的特征训练而成.由于其引入了两个随机因素,所以不容易过拟合,并且对不均衡样本的处理效果较好.

3.5 实验流程和方法参数设置

- 数据选取和划分

实验主要涉及两组:第 1 组是在数据不做任何采样情况下,将深度学习模型与传统方法对 SBR 预测的性能进行对比.首先,在 5 个数据集上分别执行论文考虑的 2 种深度学习模型和 4 种基准方法,通过实验对比验证深

© 中国科学院软件研究所 http://www.jos.org.cn

度学习模型对 SBR 预测的有效性.第 2 组实验是验证不同样本比例对深度学习模型的影响,通过对训练数据集进行重新采样,探索 SBR 数量与 NSBR 数量呈不同比例情况下对深度学习预测性能的影响.首先,每个数据集中的样本按照对正负样本进行分层采样的方法随机分为 3 部分:训练集、验证集和测试集,分配比例根据工业项目模型使用经验进行设置,训练集样本数、验证集样本数及测试集样本数的比例为 6:2:2.我们使用 N_{sbr} 表示 SBR 的数目,N_{nsbr} 表示 NSBR 数目,对于 4 个小规模数据集 Ambari,Camel,Derby 和 Wicket,我们使用“欠采样”方法,根据拟定正负样本比例设置随机选择相应数量的负样本.4 个小数据集在训练样本中的比例设置为 N_{nsbr}:N_{sbr}=x, x=1,5,10.对于大规模数据集 OpenStack,因为其正负样本数量差距较大(约 1:200),若按照欠采样方法进行处理,大量的负样本将会被丢弃,从而很可能丢失大量有用的特征文本信息,并影响预测结果.因此,针对 OpenStack 的样本分类不均衡问题,本文采用“过采样”方法进行处理,通过随机复制正样本信息来提高训练数据集中正样本分布比例.

- 模型参数的选择

本文基于开源机器学习框架 PyTorch[12]进行代码实现.对于 4 个小规模数据集,因为样本量比较少,综合考虑,每一批(batch)30 个样本,并且训练 1 000 次迭代(即 epoch),而大数据集 OpenStack,每一批(batch)100 个样本,训练 1 000 次迭代.然后,使用训练集对模型进行训练,并通过验证集评价模型,最终确定学习率为 0.01.由于 SBR 预测数据集中正样本(SBR)数量较少,因此存在严重类别不均衡问题(如本文所选择的 5 个数据集中,正样本所占比例在 0.58%~9.00%之间).此外,4 个小规模数据集由于正样本数量较少,导致所构建模型较小.因此,模型实现中,我们对全连接层偏置项(*bias*)进行了调整,以提高模型泛化能力.全连接层公式为 $y=wx+bias$,其中,w 是权重,*bias* 为偏置项,是全连接层的一种重要参数.我们测试了多组 *bias* 值,发现当 *bias*=0.2 时,其构建的模型在 4 个小规模数据集上都有较好表现,因此在本文实验中,我们将 *bias* 的取值设置为 0.2.

- 模型训练和结果输出

对两个模型(TextCNN、Attention+TextRNN)在 5 个不同的数据集进行实验,每训练一个迭代(epoch)就使用测试集进行测试,获得各个度量指标值.

4 结果分析

4.1 针对RQ1的结果分析

RQ1:深度学习方法与已有基准方法相比是否可以提升安全缺陷报告预测的性能?

目前,研究人员针对安全缺陷报告预测问题已经提出了基于各种传统分类算法与文本挖掘相结合的方法,本文中,我们以 Peter 等人[5]的工作作为基准方法,将本文给出的深度学习模型的预测性能与其工作中的最优分类算法的性能进行比较,分别分析深度学习方法在小规模数据和大规模数据情况下对安全缺陷报告预测的性能表现.我们使用 *F*1-score 作为主要判断指标,同时也给出 Accuracy、Recall 和 Precision 的值.为了使实验结果更为公正、客观,我们对每一组实验运行 10 次,然后求平均值.表 5 是 5 个数据集不进行任何采样时传统分类算法和深度学习模型取得最佳 *F*1-score 指标的结果对比.第 2 列“传统分类算法”表示 4 个 baseline 分类器中表现最好的 *F*1-score,第 3 列“深度学习”取 TextCNN 和 Attention+TextRNN 在该数据集中得到的 *F*1-score 的较大值,最后一列给出了深度学习模型最优结果相对于 4 个传统分类模型对 *F*1-score 的提升值.

Table 5 Comparision of the best *F*1-score for traditional classifier and deep learning based method

表 5 传统分类算法和深度学习方法的最优 *F*1-score 对比

数据集名称	传统分类算法		深度学习方法		*F*1-score 提升
	算法名称	*F*1-score	算法名称	*F*1-score	
Ambari	NB	0.125	Attention+TextRNN	0.615	0.490
Camel	LR	0.108	TextCNN	0.643	0.535
Derby	NB	0.360	Attention+TextRNN	0.464	0.104
Wicket	KNN	0.048	TextCNN 或 Attention+TextRNN	0.000	–0.048
OpenStack	NB	0.201	Attention+TextRNN	0.410	0.209

© 中国科学院软件研究所 http://www.jos.org.cn

在 5 个数据集中,有 4 个数据集中深度学习模型的性能优于传统分类算法.在 Wicket 数据集上,尽管深度学习传统分类算法差,但是实际上,传统分类算法也只是正确识别出了一个正样本(TP=1,F1-score=0.048),因此两者性能差异较小.

为了验证深度学习方法的统计显著性,我们将表 3 中所给出的每个数据集上表现最佳的传统分类算法和深度学习算法分别运行 10 次,通过深度学习模型所得到的结果与传统分类算法所得结果进行对比,计算 Wilcoxon rank-sum 检验所得 p-value 和 Cliff's delta,所得结果见表 6.

Table 6 Comparison results in terms of p-value and Cliff's delta

表 6 基于 p-value 和 Cliff's delta 的比较结果

数据集名称	p-value	Effect size
Ambari	<0.05	1.000 (large)
Camel	<0.001	1.000 (large)
Derby	<0.05	0.721 (small)
Wicket	<0.05	–0.745 (large)
OpenStack	<0.05	1.000 (large)

通过表 5 和表 6 可以发现,深度学习对 SBR 的预测性能在 80%的场景中优于传统分类算法,无论是在小规模数据集,还是大规模数据集 OpenStack,深度学习模型的预测性能都要显著优于传统分类算法.

4.2 针对RQ2的结果分析

RQ2:针对安全缺陷报告预测问题,不同的正负样本比例对深度学习模型性能的影响如何?

在文本分类问题中,不同类型样本的比例会影响最终预测结果准确率[29].因此在这个 RQ 中,我们拟分析采用深度学习模型进行 SBR 预测问题中,不同正负样本比例对模型的预测性能可能产生的影响.我们对 4 个小规模数据集和 1 个大规模数据集采用不同的采样策略,并计算使用 TextCNN 和 Attention+TextRNN 预测模型时所得的 F1-score 值.对于 4 个小规模数据集 Ambari、Camel、Derby 和 Wicket,我们采用欠采样方法,随机减少训练样本中的负样本数量,使其样本比例达到拟定比例设置 N_{nsbr}:N_{sbr}=x,x=1,5,10,其结果见表 7~表 10(N_{nsbr} 表示参与实验的 NSBR 数量,N_{sbr} 表示参与实验的 SBR 的数量).为了分析不同深度学习模型的迭代表现,我们打印了两个模型在整个实验过程中(共 1 000 次迭代)F1-score 随着迭代进行的变化情况,表中所给出的是 1 000 次迭代中所取得的最优 F1-score.同时,我们在表格第 3 列给出了取得最优值的迭代次数.

在大规模数据集 OpenStack 上,我们采用"过采样"方法进行正负样本不均衡问题处理,根据拟定比例对正样本进行随机复制.我们选择的复制比例分别为 1 倍、2 倍和 3 倍,其执行结果见表 11.

Table 7 Results on Ambari dataset

表 7 Ambari 数据集结果

算法名称	样本比例(N_{nsbr}:N_{sbr})	最佳迭代次数	准确率(accuracy)	查全率(recall)	查准率(precision)	F1-score
TextCNN	1	50	0.873	0.640	0.271	0.381
	5	444	0.957	0.400	0.833	0.541
	10	147	0.966	0.440	1.000	0.611
Attention+TextRNN	1	13	0.922	0.480	0.387	0.429
	5	52	0.959	0.440	0.786	0.564
	10	6	0.966	0.440	1.000	0.611

Table 8 Results on Camel dataset

表 8 Camel 数据集结果

算法名称	样本比例(N_{nsbr}:N_{sbr})	最好结果出现的迭代次数	准确率(accuracy)	查全率(recall)	查准率(precision)	F1-score
TextCNN	1	933	0.866	0.529	0.084	0.145
	5	794	0.976	0.412	0.438	0.424
	10	185	0.981	0.529	0.563	0.545
Attention+TextRNN	1	937	0.870	0.471	0.078	0.134
	5	12	0.981	0.412	0.583	0.483
	10	9	0.968	0.353	1.000	0.522

© 中国科学院软件研究所 http://www.jos.org.cn

Table 9 Results on Derby dataset

表 9 Derby 数据集结果

算法名称	样本比例 (N_{nsbr}:N_{sbr})	最好结果出现的迭代次数	准确率 (accuracy)	查全率 (recall)	查准率 (precision)	F1-score
TextCNN	1	271	0.884	0.432	0.390	0.410
	5	31	0.927	0.405	0.682	0.508
	10	10	0.909	0.351	0.520	0.419
Attention+ TextRNN	1	18	0.886	0.432	0.400	0.416
	5	12	0.929	0.459	0.680	0.548
	10	8	0.898	0.486	0.462	0.474

Table 10 Results on Wicket dataset

表 10 Wicket 数据集结果

算法名称	样本比例 (N_{nsbr}:N_{sbr})	最好结果出现的迭代次数	准确率 (accuracy)	查全率 (recall)	查准率 (precision)	F1-score
TextCNN	1	449	0.078	1.000	0.011	0.022
	5	958	0.781	0.625	0.029	0.055
	10	47	0.988	0.125	0.333	0.182
Attention+ TextRNN	1	0	0.023	1.000	0.010	0.020
	5	0	0.023	1.000	0.010	0.021
	10	41	0.978	0.125	0.091	0.105

Table 11 Results on OpenStack dataset

表 11 OpenStack 数据集结果

算法名称	正样本数量	最好结果出现的迭代次数	准确率 (accuracy)	查全率 (recall)	查准率 (precision)	F1-score
TextCNN	复制 1 倍	86	0.997	0.607	0.773	0.680
	复制 2 倍	330	0.997	0.571	0.842	0.681
	复制 3 倍	23	0.997	0.571	0.800	0.667
Attention+ TextRNN	复制 1 倍	12	0.996	0.321	0.818	0.462
	复制 2 倍	36	0.994	0.250	0.467	0.326
	复制 3 倍	6	0.993	0.036	0.250	0.063

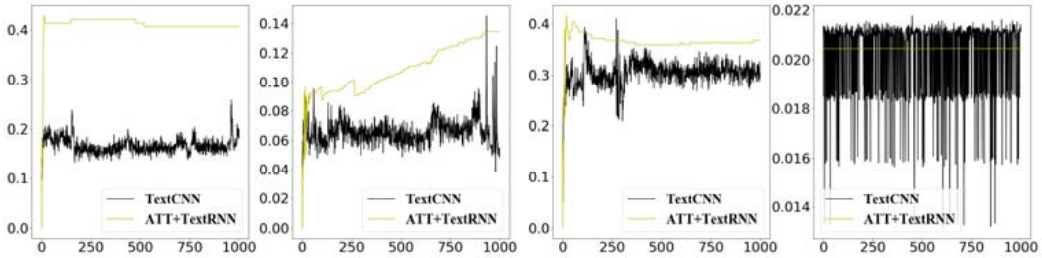
从表 7~表 10 我们可以发现,在 4 个小规模数据集中,N_{nsbr}:N_{sbr} 比例越小(即不均衡程度越低),性能指标 Recall 的取值越大;但是 Precision 指标的取值却恰恰相反,并最终造成 F1-score 指标的取值也呈减小趋势.与原始数据集(未进行采样)执行结果(即表 5)相比,“欠采样”方法对 4 个小规模数据集的性能贡献非常有限,只有在数据集 Wicket 上对 F1-score 有所提高.对于大规模数据集 OpenStack,对比表 11 和表 5(最后一行)数据可以发现,所采用的“过采样”方法可以明显提高 F1-score 指标取值,尤其在模型 TextCNN 中,F1-score 指标的最大取值可以从原始数据集中的 0.410 提高到 0.681(复制 2 倍正样本情况下),即提高了 66.10%;但是在正样本复制 3 倍的情况下,F1-score 指标的取值则呈下降趋势.此外,在 4 个小规模数据集中,“欠采样”方法对 TextCNN 和 Attention+TextRNN 模型的影响类似,所得到的性能指标值非常接近(例如数据集 Ambari 中,在 N_{nsbr}:N_{sbr}=10:1 的情况下,模型 TextCNN 和 Attention+TextRNN 所得的 F1-score 值都为 0.611).而在大规模数据集 OpenStack 中,对训练样本进行“过采样”处理后,模型 TextCNN 的整体表现要优于 Attention+TextRNN.

4.3 针对RQ3的结果分析

RQ3:是否不同的深度学习模型的迭代表现相似?

深度学习采用多次迭代方式进行模型训练.在每次迭代中,根据当前读取的小批量数据样本特征和标签,通过调用反向函数 backward 计算小批量随机梯度,并调用优化算法 Adam 迭代模型参数.核心就是在无限成本取得最优到有限成本取得次优之间权衡.图 4~图 7 是 4 个小规模数据集在不同样本均衡度情况下,F1-score 随迭代次数(epoch)的增加而产生的变化.

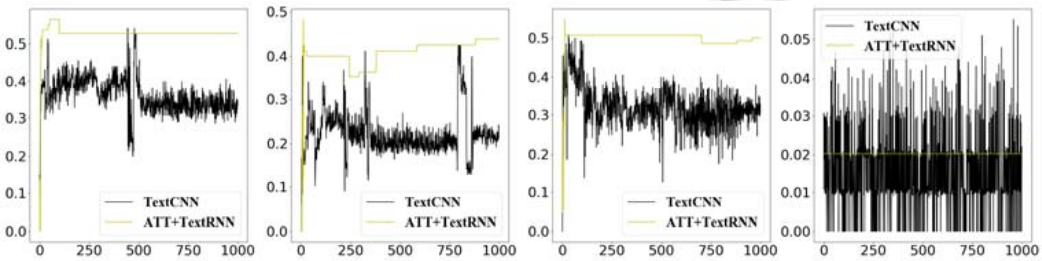
© 中国科学院软件研究所 http://www.jos.org.cn



(a) Ambari (b) Camel (c) Derby (d) Wicket

Fig.4 Trend of $F1$-score by increasing epoch (N_{nsbr}:N_{sbr}=1:1)

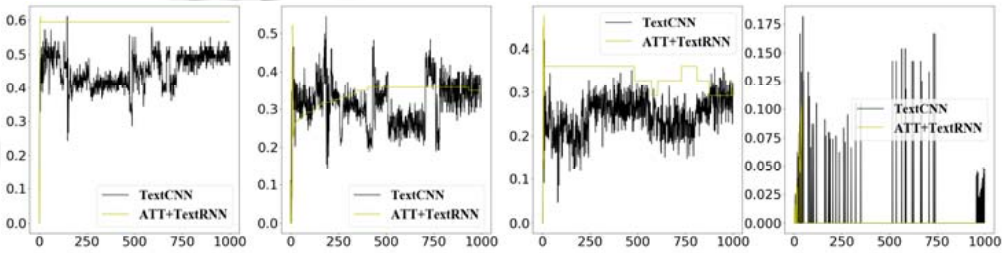
图 4 $F1$-score 随着迭代次数增加的变化趋势(N_{nsbr}:N_{sbr}=1:1)



(a) Ambari (b) Camel (c) Derby (d) Wicket

Fig.5 Trend of $F1$-score by increasing epoch (N_{nsbr}:N_{sbr}=5:1)

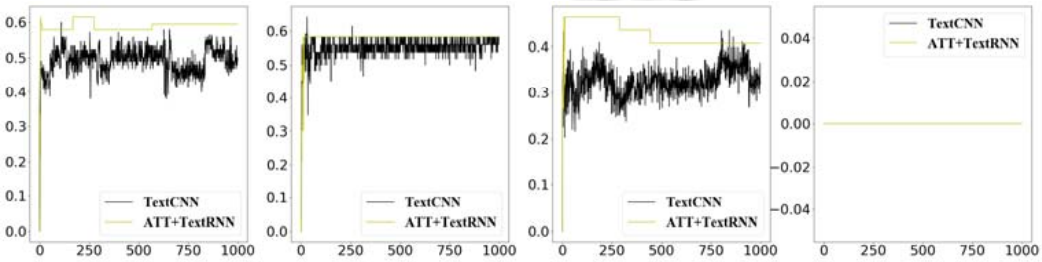
图 5 $F1$-score 随着迭代次数增加的变化趋势(N_{nsbr}:N_{sbr}=5:1)



(a) Ambari (b) Camel (c) Derby (d) Wicket

Fig.6 Trend of $F1$-score by increasing epoch (N_{nsbr}:N_{sbr}=10:1)

图 6 $F1$-score 随着迭代次数增加的变化趋势(N_{nsbr}:N_{sbr}=10:1)



(a) Ambari (b) Camel (c) Derby (d) Wicket

Fig.7 Trend of $F1$-score by increasing epoch (without sampling)

图 7 $F1$-score 随着迭代次数增加的变化趋势(未采样)

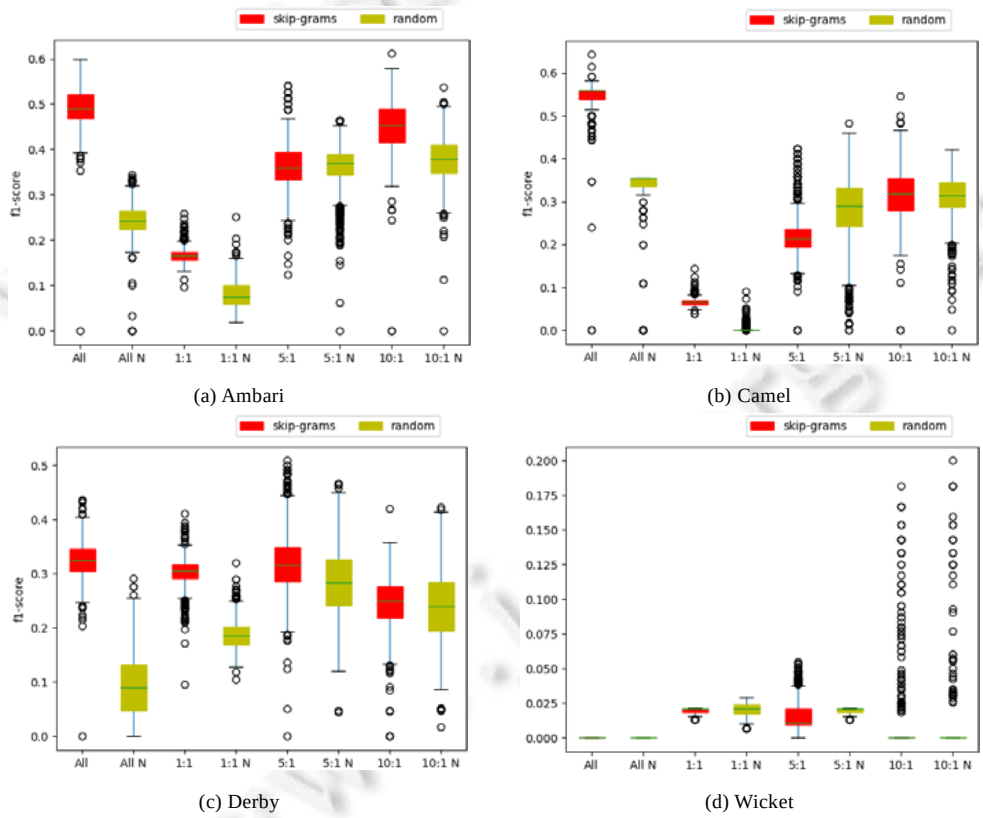
© 中国科学院软件研究所 http://www.jos.org.cn

从图 4~图 7 中可以看到,在不同的样本均衡度下,TextCNN 随着迭代的进行,*F*1-score 的变化幅度都要显著高于 TextRNN.这一方面受益于 CNN 的特性,另一方面也与 Dropout 层有关.并且随着迭代的增加(从 0 到 999,共 1 000 次),TextCNN 和 Attention+TextRNN 均获得了一定的收益.TextCNN 的收益表现为最大值的增加、震动幅度的减小和部分图上停留在 0 处的时间减少.Attention+TextRNN 则是体现在稳定性的增加和最高值的增加.但是在迭代的后期,Attention+TextRNN 几乎没有收益,而 TextCNN 仍有比较明显的收益,主要表现为震动幅度的减小和部分图上停留在 0 处的时间减少.无论是 TextCNN 和 Attention+ TextRNN 的最高值,大部分情况下均在迭代前期达到.此外,迭代变化过程中,Attention+TextRNN 的稳定性明显优于 TextCNN.

4.4 针对RQ4的结果分析

RQ4:词嵌入方法对基于深度学习的安全缺陷预测方法的性能影响程度如何?

自然语言处理中,通过词嵌入把一个维数为所有词的数量的高维空间嵌入一个维数低得多的连续向量空间中,每个单词或词组被映射为实数域上的向量[10].本文采用 word2vec 中的 skip-grams 方式构建词嵌入矩阵,我们将其与 PyTorch 默认的“随机生成”词向量方法进行对比,验证 skip-grams 在基于深度学习 SBR 预测中的性能影响.我们对 4 个小规模数据集在不同采样比例下使用 word2vec 和使用随机生成词向量的 *F*1-score 结果进行对比,图 8 和图 9 分别为 TextCNN 和 Attention+TextRNN 的结果,其中标注了 *N* 的是随机生成词向量的结果.可以看到,使用 word2vec 在绝大多数情况下,相对于随机生成词向量具有优势,因为 word2vec 生成的词向量可以很好地反映单词的相似程度,而相似程度又能辅助 TextCNN 和 Attention+TextRNN 内部参数的确定.例如,如果两个没有关系的词具有相似的词向量,会给模型带来困扰,因为模型学习不到一个参数可以拟合这个关系.

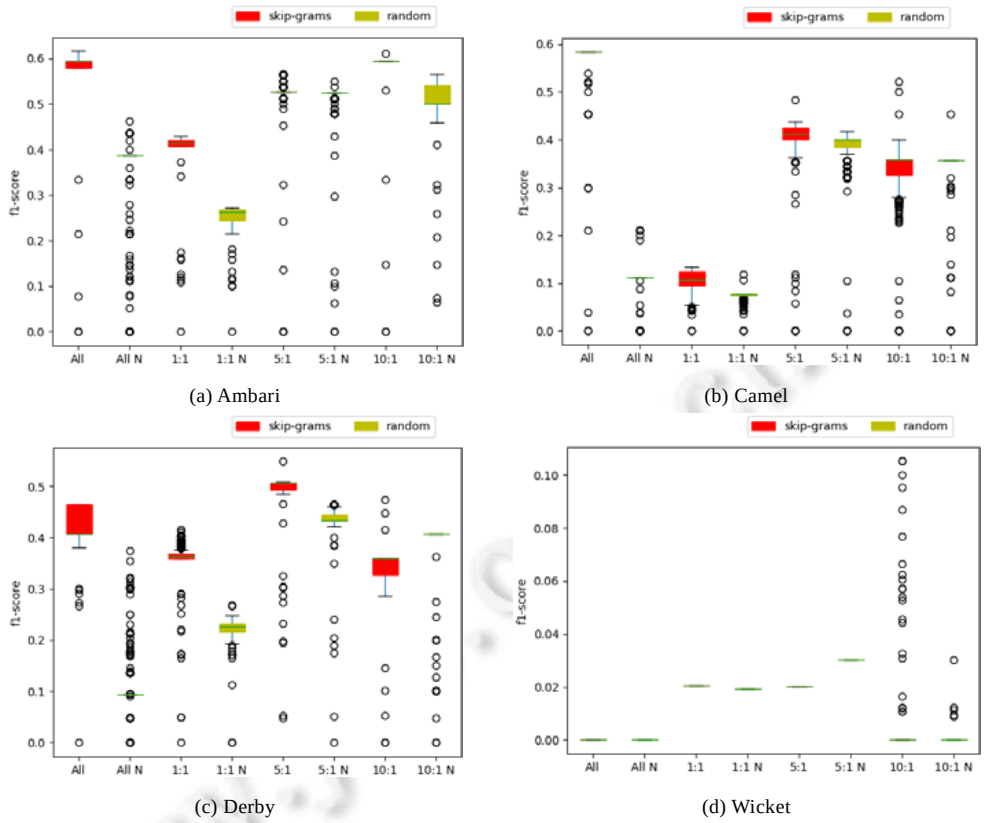


(a) Ambari (b) Camel

(c) Derby (d) Wicket

Fig.8 Comparison of word2vec and Random for embedding on 4 small-scale datasets (classification model is TextCNN)

图 8 4 个小规模数据集在不同采样比例下使用 word2vec 和随机生成词向量对比(分类模型为 TextCNN)



(a) Ambari (b) Camel

(c) Derby (d) Wicket

Fig.9 Comparison of word2vec and Random for embedding on 4 small-scale datasets (classification model is Attention+TextRNN)

图 9 4 个小规模数据集在不同采样比例下使用 word2vec 和随机生成词向量对比(分类模型为 Attention+TextRNN)

4.5 进一步讨论

4.5.1 TextCNN 模型与 Attention+TextRNN 模型的性能比较

本文设计了两种模型,即 TextCNN 和 Attention+TextRNN 模型.基于表 5 的实验结果可以看出,如果针对样本不均衡问题不进行采样处理,在数据集 Ambari,Derby 和 OpenStack 中,Attention+TextRNN 模型的性能均为最优;在 Wicket 数据集上,两个模型的性能相当;在 Camel 数据集中,TextCNN 模型性能最优.结合图 7(b)可以看出,在 Camel 数据集中,TextCNN 模型只在极个别迭代时出现优于 Attention+TextRNN 模型的情况;而在大多数迭代中,依然是 Attention+TextRNN 模型的性能更优.此外,由图 7 的曲线趋势可以看出,Attention+TextRNN 模型的稳定性要显著优于 TextCNN 模型;并且 Attention+TextRNN 模型取得性能最优值的迭代次数较小,从而可以在一定程度上减少模型训练时间.在使用采样方法的情况下,TextCNN 模型在 OpenStack 数据集中的性能表现最优,通过复制正样本 2 倍,其 $F1$-score 取值平均可以提高 65%以上;但是,其取得性能最优值所需的迭代次数也会相应增加.

4.5.2 数据集规模和正负样本比例对模型的影响

深度学习模型对训练数据量存在极大的依赖,当训练样本过少时,会出现过拟合问题,并造成模型泛化能力较弱.本文实验所采用的 4 个小规模数据集总样本数量均为 1 000,但是 Wicket 数据集中的正样本数量仅 10 条(仅占总样本数量的 1%),导致模型在正样本上的预测性能并不理想,在不进行采样情况下,TextCNN 模型和 Attention+TextRNN 模型的 $F1$-score 均为 0(如表 5 倒数第 2 行所示);而在通过“欠采样”方法使得 N_{nsbr}:N_{sbr} 为

© 中国科学院软件研究所 http://www.jos.org.cn

1:5 和 1:10 的情况下,其 *F*1-score 有小幅提升并大于 0.对于大规模数据集 OpenStack,虽然其正负样本的数量都远高于小规模数据集,但是在不进行采样的情况下,其 *F*1-score 取值要低于 3 个小规模数据集 Ambari、Camel 和 Derby.造成该问题的一个可能原因是正样本比例过低,在 3 个小规模数据集中,正样本比例约在 3%~9%之间,而在 OpenStack 数据集上的正样本比例仅为 0.6%.综上所述,数据集的规模以及正负样本比例对本文考虑的模型均存在较为显著的影响.

4.6 有效性影响因素分析

这一节分析可能影响实验有效性的影响因素.

- 内部有效性主要涉及到可能影响到实验结果正确性的内部因素.一个重要的内部影响因素就是代码的正确性.在本次实验中,为了减少人为因素,同时使我们的结论不受特殊实现的影响,使其具有普遍性,我们使用了深度学习中成熟且广泛使用的框架 PyTorch,并且使用了 PyTorch 对 CNN、LSTM 等层的默认实现.此外,我们通过相互之间进行代码评审来保证代码的质量和准确.
- 外部有效性主要涉及到实验的结论是否具有一般性.这里,我们选用了 SBR 预测研究领域常用的 4 个公开数据集 Ambari、Camel、Derby 和 Wicket,这 4 个数据集分别来自 Apache 基金会的 4 个同名的经典项目,覆盖了多种软件的典型应用领域.除此之外,为了验证本文结论是否具有一般性,我们还额外选择了来自开源项目 OpenStack 的缺陷报告作为数据集.OpenStack 缺陷报告数据量大,来自多个项目,且我们之前的研究工作[35]已对 OpenStack 的缺陷报告进行了深入分析,为本次 SBR 预测数据集构建奠定了基础.因此,可以判断实验结论具有可靠性和代表性.
- 结论有效性主要指实证研究中使用的评测指标是否合理.本文使用了 SBR 研究领域经常使用的多种指标:准确率、查准率、查全率和 *F*1-score,因此可以从多个角度对基于深度学习的 SBR 预测模型性能做出客观的评价.

5 总结和展望

本文提出了基于深度学习的 SBR 预测方法,主要使用了 TextCNN 和 Attention+TextRNN 两个模型.首先是进行了不同规模数据集中基于深度学习的方法和传统机器学习方法的性能对比,然后研究了不同的样本均衡比例对不同深度学习模型的性能影响;接着探究了不同深度学习模型的迭代表现以及不同词嵌入方法对模型 SBR 预测性能的影响.本文使用了 SBR 预测研究领域广泛使用的 4 个小规模数据集和基于开源 OpenStack 项目 BR 构建的大规模数据集最为实证对象,结果表明,深度学习模型的预测性能要显著优于传统机器学习模型,其性能指标 *F*1-score 平均提升 0.258,而最好的情况下可提升 0.535.

本文属于深度学习模型在 SBR 预测研究中的初步探索,因此还存在较大的提升空间.首先,在生成词向量矩阵的时候,假如训练集和测试集的词相差太多,会导致我们使用训练集测试的词向量对测试集不适用.尽管可以选择重新生成词向量,但是新生成的词向量如果套用原模型而不进行重新训练会造成模型性能下降.而如果不训练新的词向量的话词向量就会向随机词向量方向退化,同样影响模型性能.后续工作中,我们可以考虑在在一个大语料库中训练词向量,借鉴自然语言处理中的最新方法来改进词向量库.其次,由于 SBR 的类型较多(如拒绝服务攻击、SQL 注入、内存溢出等),不同类型的安全缺陷特征差异性较大,本文所给出的模型对不同类型的安全漏洞是否具有不同敏感性,也将是我们进一步研究的方向.另外,一些新型的网络结构(例如 transformer[28]和可变卷积核[47])是否能给 SBR 预测问题带来更好的解决方案还有待验证.

References:

[1] Amoroso E. Recent progress in software security. IEEE Software, 2018,35(2):11−13.

[2] CVE Website. 2019. https://www.openstack.org/

[3] CVE Detail. 2019. https://www.cvedetails.com/vendor/11727/Openstack.html

© 中国科学院软件研究所 http://www.jos.org.cn

[4] Wijayasekara D, *et al*. Mining bug databases for unidentified software vulnerabilities. In: Proc. of the Int'l Conf. on Human System Interaction (HIS). IEEE, 2012. 89–96.

[5] Peters F, Tun TT, Yu YJ, Nuseibeh B. Text filtering and ranking for security bug report prediction. IEEE Trans. on Software Engineering, 2019,45(6):615–631.

[6] Gegick M, Rotella P, Xie T. Identifying security bug reports via text mining: An industrial case study. In: Proc. of the 2010 7th IEEE Working Conf. on Mining Software Repositories (MSR). IEEE, 2010.

[7] Shu R, Xia TP, Williams L, Menzies T. Better security bug report classification via hyperparameter optimization. arXiv preprint arXiv:1905.06872, 2019.

[8] Goseva-Popstojanova K, Tyo J. Identification of security related bug reports via text mining using supervised and unsupervised classification. In: Proc. of the Int'l Conf. on Software Quality, Reliability, and Security. 2018. 344–355.

[9] Zhang S, Yao LN, Sun AX, Tay Y. Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys (CSUR), 2019,52(1):1–38.

[10] Lai SW, Liu K, He SZ, Zhao J. How to generate a good word embedding. IEEE Intelligent Systems, 2016,31(6):5–14.

[11] PyTorch official website. 2019. https://pytorch.org/

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the Advances in Neural Information Processing Systems. 2017. 5998–6008.

[13] Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.

[14] Liu PF, Qiu XP, Huang XJ. Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101, 2016.

[15] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[16] Agrawal A, Menzies T. Is "Better Data" Better Than "Better Data Miners"? In: Proc. of the 2018 IEEE/ACM 40th Int'l Conf. on Software Engineering (ICSE). IEEE, 2018. 1050–1061.

[17] Lopez MM, Kalita J. Deep learning applied to NLP. arXiv preprint arXiv:1703.03091, 2017.

[18] Hinton GE, Krizhevsky A, Wang SD. Transforming auto-encoders. In: Proc. of the Int'l Conf. on Artificial Neural Networks. Springer-Verlag, 2011. 44–51.

[19] Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: Proc. of the Advances in Neural Information Processing Systems. 2017. 3856–3866.

[20] Yin Z, Shen YY. On the dimensionality of word embedding. In: Advances in Neural Information Processing Systems. 2018. 887–898.

[21] Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[22] Blei DM, Ng AY, Jordan MI, *et al*. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3(1):993–1022.

[23] Bojanowski P, Grave E, Joulin A, *et al*. Enriching word vectors with subword information. Trans. of the Association for Computational Linguistics, 2017,5:135–146.

[24] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing. 2014. 1532–1543.

[25] Peters ME, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.

[26] Devlin J, Chang MW, Lee K, *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[27] Radford A, Narasimhan K, Salimans T, *et al*. Improving language understanding by generative pre-training. 2018. https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf

[28] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: Proc. of the Advances in Neural Information Processing Systems. 2017. 5998–6008.

[29] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

© 中国科学院软件研究所 http://www.jos.org.cn

[30] Krawczyk B. Learning from imbalanced data: Open challenges and future directions. Progress in Artificial Intelligence, 2016,5(4): 221−232.

[31] Huang C, Li YN, Loy CC, Tang XO. Learning deep representation for imbalanced classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 5375−5384.

[32] Song J, Huang XL, Qin SJ, Song Q. A bi-directional sampling based on *K*-means method for imbalance text classification. In: Proc. of the 15th Int'l Conf. on Computer and Information Science. 2016. 1−5.

[33] Greff K, Srivastava RK, Koutník J, *et al*. LSTM: A search space odyssey. IEEE Trans. on Neural Networks & Learning Systems, 2015,28(10):2222−2232.

[34] Chaparro O, Lu J, Zampetti F, *et al*. Detecting missing information in bug descriptions. In: Proc. of the Joint Meeting on Foundations of Software Engineering. 2017.

[35] Zheng W, Feng C, Yu TT, *et al*. Towards understanding bugs in an open source cloud management stack: An empirical study of OpenStack software bugs. Journal of Systems and Software, 2019,151:210−223.

[36] NLTK. 2019. http://www.nltk.org/

[37] Bishop CM. Pattern Recognition and Machine Learning. Springer-Verlag, 2006.

[38] Ohira M, Kashiwa Y, Yamatani Y, *et al*. A dataset of high impact bugs: Manually-classified issue reports. In: Proc. of the 2015 IEEE/ACM 12th Working Conf. on Mining Software Repositories. IEEE, 2015. 518−521.

[39] Ambari. 2019. http://ambari.apache.org/

[40] Camel. 2019. http://camel.apache.org/

[41] Derby. 2019. http://db.apache.org/derby/

[42] Wicket. 2019. http://wicket.apache.org/

[43] Apache. 2019. http://db.apache.org/

[44] OpenStack bugs in LaunchPad. 2019. https://bugs.launchpad.net/openstack/

[45] LaunchPad. 2019. https://launchpad.net/

[46] Romano J. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys. In: Proc. of the Annual Meeting of the Florida Association of Institutional Research. 2006.

[47] Dai JF, Qi HZ, Xiong YW, Li Y, Zhang GD, Hu H, Wei YC. Deformable convolutional networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 764−773.

郑炜(1975－),男,安徽合肥人,博士,副教授,CCF 专业会员,主要研究领域为软件测试,软件安全漏洞分析与检测.

陈军正(1998－),男,学士,主要研究领域为软件安全研究,高性能计算.

吴潇雪(1983－),女,博士生,CCF 学生会员,主要研究领域为软件漏洞分析与检测,软件测试.

陈翔(1980－),男,博士,副教授,CCF 高级会员,主要研究领域为软件缺陷预测,软件缺陷定位,回归测试和组合测试.

夏鑫(1986－),男,博士,讲师,主要研究领域为软件工程,软件仓库挖掘,经验软件工程.

© 中国科学院软件研究所 http://www.jos.org.cn