基于 GAN-DLNA 深度学习算法的计算机软件缺陷预测分析研究

韩 倩[1]
HAN Qian

摘 要 针对计算机软件缺陷预测存在的数据类不平衡、冗余或无关数据过多等问题，基于贝叶斯网络（bayesian network）算法、深度学习网络算法（GAN deep learning network algorithm, GAN-DLNA）的软件缺陷预测提取方案，由贝叶斯网络的随机森林模型构建有向无环图，以软件缺陷目标变量节点为中心、用贝叶斯网络分类器（GBNC）感知及处理缺陷的度量元数据样本，利用 GAN-DLNA 深度学习生成对抗网络算法，使用基于均方误差、交叉熵的损失函数对多层深度学习网络进行扩展训练，通过增大源数据训练的误分类权重、减小目标数据训练的误分类权重，以此大大提升软件缺陷数据分类、缺陷数量预测质量。

关键词 大数据深度学习算法；计算机软件；缺陷预测分析；分类与集成

doi：10.3969/j.issn.1672-9528.2025.03.018

0 引言

软件缺陷预测为发现计算机软件开发问题的重要方式，常见的缺陷预测模型分为有监督、无监督和半监督模型，通常缺陷样本数据集的数量少、集中度高，更适宜使用无监督聚类算法作出预测分析。如学者崔梦天等人针对报告库提交的软件缺陷报告文件，提出一种自然程序双模态语言的 LDA-BERT 缺陷报告检测模型，多次抽取上下文语义的特征向量作出识别分析。任晓莹等人利用 ReliefF 和 Fisher 判别比度量的特征数据集选择方案。本文面对计算机软件缺陷数据分布不平衡的实际特征，引入应用涵盖判别器和自编码器的深度学习网络算法（GAN deep learning network algorithm, GAN-DLNA）深度生成对抗网络算法，利用自编码器高斯混合模型（gaussian mixture model, GMM）去除被测数据噪声，基于 APE 均方误差、APCE 交叉熵的损失函数作出软件缺陷二分类模型训练，在降低深度学习算法迭代水平同时提高软件缺陷的预测与分类精度。

1 传统神经网络算法的不平衡数据集分类错误问题

计算机软件缺陷预测是对软件开发的不同组件模块代码、网络代码作出分析，在基于深度学习神经网络算法进行软件缺陷预测过程中，通常涉及软件代码度量、训练集特征提取、预测分析模型建构、缺陷预测等工作执行流程，即先需要将被测的软件代码转换为容易理解的标准化特征度量元，再进行正常或缺陷训练数据集的特征提取、缺陷预测分析。

当前常用的软件缺陷预测神经网络算法，分为过采样算法、欠采样算法两种技术模式，由于存在软件缺陷的代码数据量、远远小于正常代码的数据量，使用以上两种算法作出数据提取、容易出现不平衡数据集的分类错误问题。如基于随机过采样算法、SMOTE 算法等过采样技术，从软件缺陷的少数类数据集中随机抽取样本作为新样本、重复抽取多次使少数类样本数量接近多数类，但该模式可能存在少数类样本过拟合（重复样本采集）、噪音数据采集的情况，导致模型本身的噪音数据干扰严重、泛化能力下降；而基于随机欠采样、Near Miss 欠采样算法等技术，是从正常的多数类数据中选取部分样本，与少数类样本聚合后作出预测分析，该模式缺点在于丢弃的多数类样本数据中可能包含重要信息，丢弃后的模型预测训练的召回率、精准度都严重下降，由此发生欠拟合情况，软件缺陷预测的召回率、精准度[1]计算公式为：

$$\text{Recall}=\frac{TP}{TP+FN} \qquad \text{Precision}=\frac{TP}{TP+FP} \tag{1}$$

式中：TP 和 TN 分别表示多数类（正常）数据、少数类（缺陷）数据；FP 和 FN 分别表示预测为多数（正常）的少数类数据、预测为少数 (缺陷) 的多数类数据。由此可见，采用“输入 - 输出”映射关系模型的神经元欠采样和过采样算法，作出不平衡软件缺陷数据集的预测分析计算，往往更多关注多数类样本数据、忽略少数类样本数据的分类，在数据特征提取与

1. 茂名职业技术学院 广东茂名 525000

分类中容易陷入局部收敛，且泛化能力过差。

2 基于大数据深度学习算法的计算机软件数据转换与特征提取分类

2.1 被测软件代码数据的度量元转换

软件度量是将计算机软件开发产品属性作出数据映射量化，分为软件程序开发过程度量、程序代码度量，涉及代码行数（line of code, LOC）、McCabe 组件模块复杂度、Halstead 代码运算符数量等度量指标，用于反映计算机软件代码的集成度、耦合度和内聚度等属性特征。如基于 NASA MDP 计算机软件缺陷数据集，从 13 个不同数据集中选取 7 个作为待测样本，分别为 CM1、JM1、KC1、KC3、KC4、PC1、PC2 数据集，按照 LOC、McCabe、Halstead 的度量指标设置作出代码数据的度量元转换，得到表 1 的模块度量元统计结果 [2]。

表 1 NASA MDP 软件缺陷数据集

数据集项目	模块总数 / 个	度量元数 / 个	缺陷模块占比 /%
CM1	308	38	9.50
JM1	7 456	2 214	17.31
KC1	2 032	325	16.24
KC3	267	40	48.17
KC4	101	15	9.77
PC1	946	73	7.35
PC2	1 029	124	8.44

2.2 基于贝叶斯网络算法的计算机软件缺陷数据样本特征提取

面向计算机软件开发的少数类、多数类样本数据，先将特定软件项目度量元的数据集、按照变量节点的父子关联关系构建有向无环图，其中根节点为软件项目缺陷数据集 X，子节点为具有多个类别属性的度量元$\{x_1,x_2,\ldots,x_n\}$，可用 CF=(LS, SS, n) 的三元组公式表示某一软件项目缺陷数据集 X 下的度量元样本集总和、平方和 [3]，计算公式为：

$$\begin{aligned} \mathrm{LS} &= \sum_{i=1}^{n}\left(x_1+x_2+\ldots+x_n\right) \\ \mathrm{SS} &= \sum_{i=1}^{n}\left(x_1+x_2+\ldots+x_n\right)^2 \end{aligned} \tag{2}$$

式中：LS 和 SS 分别表示包含 i 软件缺陷对象的度量元样本集总和、平方和。假设度量元 x_i 样本的类别属性为 C_j，且 $C=\{C_1,C_2,\ldots,C_j,\ldots,C_m\}$中包含着软件缺陷模块（incorrect）、无缺陷模块（Correct）的度量元属性，可用贝叶斯网络算法、对软件项目缺陷数据集 X 的不同度量元作出属性提取与划分 [4]，具体的计算公式为：

$$\begin{aligned} C_j(x_i) &= \arg\max P(C_m \mid x_i) \\ &= \arg\max \frac{P(C_m)\prod_{j=1}^{m} p(C_j \mid C_m)}{\sum_{j=1}^{m} P(C_m)} \end{aligned} \tag{3}$$

式中：$C_j(x_i)$ 表示将度量元 x_i 样本划分至类别属性 C_j 的概率；$P(C_m)$ 和 $p(C_j|C_m)$ 分别表示度量元 x_i 样本出现的先验概率、条件概率。在计算机软件缺陷测试集的 n 个度量元样本分属于 m 个不同的分类属性情况下，基于加权贝叶斯网络的随机森林模型，将 n 个度量元样本分配至 m 个属性类，则在单个根节点下的度量元样本数据 x_i 属于 C_j 的纯度计算公式为：

$$\mathrm{Gini}(C_j,x_i)=\frac{|x_i|}{\sum_{i=1}^{n}|x_1+x_2+\ldots+x_n|}\mathrm{Gini}(C_j) \tag{4}$$

式中：Gini(C_j, x_i) 为将度量元样本数据 x_i 分配至 C_j 属性类的基尼系数；Gini(C_j) 为将度量元样本数据 x_i 分配至 C_j 属性类的 Gini 基尼指数。Gini(C_j, x_i) 基尼系数用于评估软件缺陷度量元样本数据不为同一类的概率，Gini(C_j, x_i) 值越小则表明度量元样本属于同一属性类的纯度越高，若 Gini(C_j, x_i)=0 则表明单个根节点下的度量元样本均属于 C_j 属性类，也即某一软件项目缺陷只被归为一个类别，而后作出软件缺陷度量元样本的正则化评估分析，得到不同属性类下样本数据出现的概率，计算公式为：

$$P_{\mathrm{laplace}}(x_i)=\frac{P(C_m \mid x_i)+\alpha}{N+\alpha P(C_m)} \tag{5}$$

式中：$P_{\mathrm{laplace}}(x_i)$ 为基于拉普拉斯平滑公式，计算软件缺陷度量元样本在特定属性类下的出现概率结果；$P(C_m|x_i)$ 为软件缺陷度量元样本在所有属性类中出现的概率；N 为总的分析评估次数；α 为用于控制平滑程度的正则化参数。

3 基于 GAN-DLNA 深度神经网络算法的软件缺陷去噪与分类结果预测策略

3.1 基于 GAN-DLNA 网络算法判别器的样本数据去噪训练

具体 GAN-DLNA 网络算法结构 [5] 如图 1 所示。

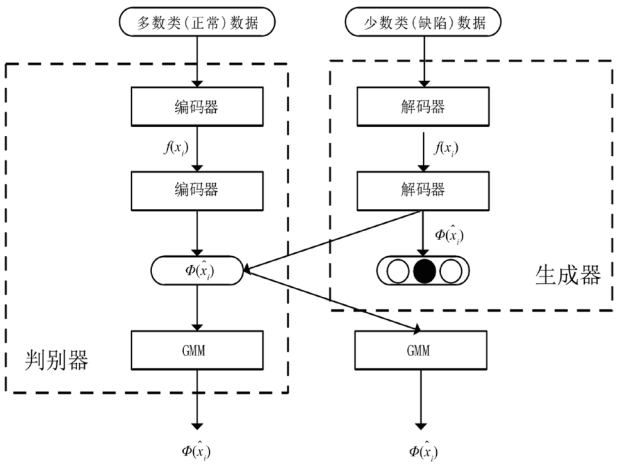

图 1 GAN-DLNA 网络算法结构分类结果预测流程

GAN-DLNA 深度神经网络算法为存在多个隐含层的神经网络结构，根据输入的计算机软件缺陷样本数据集多

少，设置输入 / 输出隐含层节点总数$V=\{v_1,v_2,\ldots,v_i,\ldots,v_n\}$、以及不同层级之间的神经元连接权重。而后基于深度生成对抗网络的生成器、判别器，作出不平衡数据分类中被测训练集样本的噪声去除，判别器分为自编码器、GMM 高斯混合模型的组成部分，其中自编码器负责对被测训练集样本内噪声的去噪处理。

若计算软件缺陷度量元样本可用 $x_i=(\hat{x}_i,\varepsilon)$ 表示，其中$\hat{x}_i$和 ε 分别表示不包含噪声的原始代码样本数据、噪声数据，那么将度量元样本 x_i 输入自编码器作出去噪过程中，可引入深度生成对抗网络的 f_θ 激活函数、数据去噪激活值 $h_{w,b}(\tilde{x})$ 完成度量元样本的原始数据提取计算，计算公式为：

$$J_{\mathrm{Dae}}(w,b)=\frac{1}{k}\sum_{i=1}^{n}\left(\frac{1}{2}\left\|h_{w,b}(x_i)-\hat{x}_i\right\|^2\right)$$
$$h_{w,b}(x_i)=f\left(\sum_{i=1}^{n}wx_i+b\right) \tag{6}$$

式中：$J_{\mathrm{Dae}}(w,b)$ 和 $h_{w,b}(x_i)$ 分别为深度网络算法隐含层的自编码器代价函数、去噪激活值；w 和 b 分别为深度网络算法参与度量元样本去噪的神经元权重、偏置，k 为参与自编码器噪声去除训练的度量元样本数，$J_{\mathrm{Dae}}(w,b)$ 的值越小表明模型能够更好地从含噪声的输入度量元样本数据中恢复出原始数据。

3.2 基于 GAN-DLNA 网络算法的软件缺陷分类结果预测分析

将自编码器去噪后的输出值作为隐含层的输入值，基于编解码器的均方误差（measquare error, MSE）、交叉熵（cross entropy error, CEE）的损失函数公式，作出不平衡数据集原始数据训练值、预测值之间的误差计算，用 pred 和 real 表示输出层神经元度量元样本分类预测期望、真实结果[6]，计算公式为：

$$f(x_i)_{\mathrm{MSE}}=\frac{1}{k}\sum_{i=1}^{n}\frac{1}{O}\left(\mathrm{pred}\_\hat{x}_i-\mathrm{real}\_\hat{x}_i\right)^2$$
$$f(x_i)_{\mathrm{CEE}}=-\frac{1}{k}\sum_{i=1}^{n}\frac{1}{O}\left(\mathrm{real}\_\hat{x}_i-\log\left(\mathrm{pred}\_\hat{x}_i\right)\right) \tag{7}$$

式中：O 为输出隐含层的网络神经元个数；$\mathrm{pred}_\hat{x}_i$和 $\mathrm{real}\_\hat{x}_i$ 分别为输出层神经元度量元样本分类预测期望、真实结果；$f(x_i)_{\mathrm{MSE}}$ 和 $f(x_i)_{\mathrm{CEE}}$ 分别为错分计算的均方误差、交叉熵。由于不平衡分布数据集中多数类样本远远多于少数类样本的数量，因而通过调整分类器的分类阈值，可使损失函数倾向于少数类软件缺陷样本的错分误差计算，得到少数类 (缺陷) 数据分类的错分率 BER[7] 公式为：

$$\mathrm{BER}=1-\frac{1}{2}(\mathrm{Recall}+\mathrm{Precision}) \tag{8}$$

式中：Recall 和 Precision 分别为计算机软件缺陷样本预测的召回率、精准度；BER 表示少数类（缺陷）数据分类的平均错分率，则少数类（缺陷）数据分类的平均准确率为 1-BER。在基于隐含层编解码器完成少数类（缺陷）数据的多次预测分类扩展训练后，使用 GMM 高斯混合模型对多数类（正常）、少数类（缺陷）样本数据作出经验分布训练，得到计算机软件缺陷度量元作出正确属性类划分的结果[8]，计算公式分别为：

$$\mathrm{argmax}\left[E_{\hat{x}_i\sim\chi^+}\log\phi f(x_i)_{\mathrm{MSE}}+E_{\hat{x}_i\sim\chi^-}\log\left(1-\phi f(x_i)_{\mathrm{MSE}}\right)\right] \tag{9}$$

$$\phi\left(\hat{x}_i\right)=\frac{1}{\left|\hat{x}_i\right|}\sum_{i=1}^{n}\log\left(\omega_i\,\hat{x}_i(\mu_i,\sigma_i)\right) \tag{10}$$

式中：$E_{\hat{x}_i\sim\chi^+}$和$E_{\hat{x}_i\sim\chi^-}$分别为多数类（正常）、少数类（缺陷）样本数据的经验分布；$\phi(\hat{x}_i)$ 为提取软件缺陷度量元训练集作出正确属性类划分的概率；ω_i 为第 j 个数据所属属性类向量的高斯权重；$\hat{x}_i(\mu_i)$、$\hat{x}_i(\sigma_i)$分别表示训练集属性分类向量的多维高斯函数均值、向量对角协方差矩阵，用于描述训练集属性类的分布情况。

在不平衡分布数据集的多数类（正常）、少数类（缺陷）数据量相差不大情况下，可将自编解码器的分类阈值设为 0.5，但实际情况是多数类（正常）数据量远高于少数类（缺陷）数据量，此时需设置编解码器的分类阈值趋近于 1、$\sum_{i=1}^{n}\omega_i=1$，以此增加错分样本的迭代训练权重值 e。在基于构造的多个分类器完成计算机软件缺陷数据集的训练后，将 $\phi_1(\hat{x}_i),\phi_2(\hat{x}_i),\cdots,\phi_7(\hat{x}_i)$ 等多个训练子集的正确属性类划分概率作出汇总，若其满足式（11）的判定规则、则表明少数类 (缺陷) 样本数据的正确分类精度满足需求[9]。

$$\begin{cases}\left(\phi_1\left(\hat{x}_i\right)\geq\mathrm{BER}_1\right)\cup\left(\phi_2\left(\hat{x}_i\right)\geq\mathrm{BER}_2\right)\ldots\cup\left(\phi_7\left(\hat{x}_i\right)\geq\mathrm{BER}_7\right) & \mathrm{THEN}\,\hat{x}_i\in\chi^+\\ 0.5\leq\mathrm{BER}_{K/2}\leq\ldots\leq1 & \mathrm{ELSE}\,\hat{x}_i\in\chi^-\end{cases} \tag{11}$$

式中：BER 为少数类（缺陷）样本数据分类的错分率，即在计算机软件缺陷度量元训练集正确属性分类概率 $\phi(\hat{x}_i)$ 大于错分率、至少 $K/2$ 个样本数据错分率值大于 0.5 的情况下，得到计算机软件少数类（缺陷）样本的正确分类精度最高、损失代价最小，表明使用多个分类器的被测训练集分类与集成、可减轻不平衡分类的倾斜效果。

4 仿真实验及结果分析

4.1 实验环境及指标设置

选用 NASA MDP 计算机软件缺陷数据集作为仿真实验数据，基于 Matlab2022a 仿真软件对表 1 的模块度量元统计结果作出数据模型预测分析。在 Matlab2022a 仿真软件中设置 GAN-DLNA 深度学习生成对抗网络算法的分类阈值，利

用自编码器作出训练集的去噪及基分类迭代操作后，使用判别器、生成器设置处理后的缺陷数据集高斯权重，计算少数类（缺陷）样本数据的属性类向量分布情况。引入 CNN 卷积神经网络算法作为对比算法，以均方误差 $f(x_i)_{MSE}$、BER 平均错分率等指标作出实验测试结果评估分析。

4.2 实验结果分析

基于半监督学习的 GAN-DLNA 生成对抗网络算法，按照输入表示、生成对抗、生成器和判别器半监督学习的执行流程，使用多个隐含层的网络神经元作出不平衡多数类（正常）、少数类（缺陷）数据的二分类预测评估，得到软件缺陷数据集训练值、预测值之间的误差分析结果如表 2、图 2 所示。

表 2 实验测试结果评估分析

数据集项目	GAN-DLNA 深度学习算法		CNN 卷积神经网络算法	
	$f(x_i)_{MSE}$	BER 错分率	$f(x_i)_{MSE}$	BER 错分率
CM1	0.003 57	3.45	0.011 70	12.49
JM1	0.004 83	5.27	0.016 52	19.53
KC1	0.006 25	6.30	0.023 14	26.44
KC3	0.003 84	4.86	0.025 32	30.75
KC4	0.002 29	4.09	0.019 68	23.21
PC1	0.003 77	3.24	0.017 43	22.58
PC2	0.004 98	4.15	0.016 39	20.64

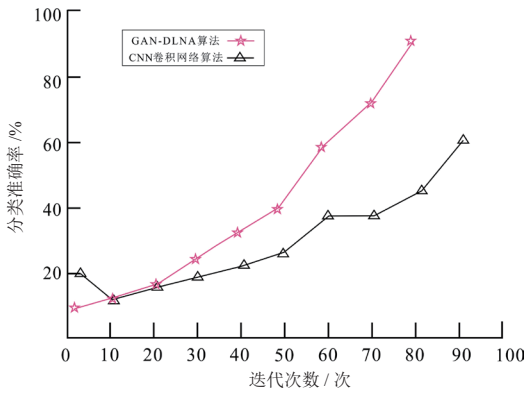

图 2 基于 GAN-DLNA 和 CNN 网络算法的仿真结果

由表 2、图 2 的仿真实验结果可得，基于 GAN-DLNA 深度学习算法的 APCE 损失函数为多数和少数类的软件缺陷样本数据赋予权值，促使输出隐藏层属性类特征提取的数据分类面朝着正确方向移动，得到的软件缺陷代码数值分类预测的均方误差 $f(x_i)_{MSE}$ 均控制在 0.001 以内，满足数据集训练值、预测值之间误差计算的需求，BER 平均错分率控制在 5% 以内，经过多次迭代的分类准确率达到 95% 以上，且多项指标均优于 CNN 卷积神经网络算法的迭代结果，表现出优秀的不平衡数据分类准确率和鲁棒性。

5 结语

“互联网 +”环境下计算机软件程序的应用环境日益复杂，软件组件运行的缺陷数据集提取、缺陷问题预测成为软件管理重点。因此基于 GAN-DLNA 深度生成对抗网络算法，对软件项目缺陷数据集 X 中不同度量元作出属性提取的结果，通过增大少数类数据训练权重、减小多数类数据训练权重，由“自编码器 - 解码器”对度量元样本的输入、输出数据之间差异作出预测分析，将包含多个基特征的数据集作出分类与合并，可使分类面朝正确方向移动，并提高软件缺陷预测分类精准度。

参考文献：

[1] 崔梦天，杨善矿，袁启航 . 基于 LDA-BERT 重复缺陷报告检测模型研究 [J]. 西南民族大学学报 (自然科学版), 2023, 49(4): 414-423.

[2] 田笑，常继友，张玉清，等 . 开源软件缺陷预测方法综述 [J]. 计算机研究与发展 , 2023, 60(7): 1467-1488.

[3] 任晓莹，陈浩，王淑琴，等 . 基于判别结构向量互补的集成特征选择方法 [J]. 天津师范大学学报 (自然科学版), 2023, 43(4): 57-63.

[4] 王越，赵静，杜冠瑶，等 . 网络空间安全日志关联分析的大数据应用 [J]. 网络新媒体技术 , 2020,9(3):1-7.

[5] 崔梦天，龙松林，谢琪，等 . 基于量子粒子群混合烟花优化支持向量机的软件缺陷预测研究 [J]. 西南民族大学学报 (自然科学版), 2022, 48(6): 653-659.

[6] 魏威，江峰 . 基于加权复杂度的 SMOTE 算法及其在软件缺陷预测中的应用 [J]. 计算机与数字工程 , 2024, 52(5): 1418-1422.

[7] 齐莉 . 云计算背景下分布式软件系统故障检测技术研究 [J]. 电子制作 , 2021(18): 88-90.

[8] 张莹，朱丽娜 . 一种基于半监督集成学习的软件缺陷预测方法 [J]. 计算机与数字工程 , 2023, 51(10): 2390-2394.

[9] 郭威，谢光伟，张帆，等 . 一种分布式存储系统拟态化架构设计与实现 [J]. 计算机工程 ,2020,46(6):12-19.

【作者简介】

韩倩（1981—），女，山东烟台人，硕士，讲师，研究方向：计算机应用技术。

（收稿日期：2024-11-01）