面向类别不平衡的软件缺陷预测半监督深度学习方法

姚永芳[1] 丁永昌[2] 陈林君[2] 廖 珂[1] 荆晓远[1,2*]
YAO Yongfang DING Yongchang CHEN Linjun LIAO Ke JING Xiaoyuan

摘 要 针对软件缺陷预测中常见的“标注数据稀缺”与“类不平衡”两大挑战，文章提出了一种半监督类不平衡深度度量学习方法（SSDML-SDP）。通过深度度量学习构建判别性嵌入空间，使缺陷模块与无缺陷模块在度量空间中有效区分；在训练过程中引入类别权重策略，缓解多数类对模型的干扰；同时结合伪标签机制充分利用大量无标签数据，以提升模型的泛化能力。在 10 个 NASA 公共数据集上进行了实证研究，结果表明，所提出方法在 *F*-measure 和 AUC 等指标上均优于现有代表性方法，尤其在缺陷样本比例极低的数据集上表现更为突出。说明 SSDML-SDP 在有限标注和类不平衡条件下能够有效提升缺陷预测性能。

关键词 软件缺陷检测；半监督学习；类不平衡；深度学习

doi：10.3969/j.issn.1672-9528.2026.07.004

0 引言

软件缺陷预测（software defect prediction, SDP）是软件质量保障中的关键环节，其目标是利用已有的历史信息预测哪些软件模块更可能含有缺陷，从而帮助测试人员将有限资源优先分配到高风险模块。随着软件系统规模和复杂度的不断增加，仅依赖人工检查或全面测试已不现实，因此能够有效识别缺陷倾向模块对于降低维护成本、提升可靠性至关重要。

然而，当前 SDP 研究面临两大核心挑战。一是标注数据有限：缺陷标注需要依赖专家经验，获取成本高昂，而实际项目中可用的缺陷样本数量往往十分有限；相对而言，无标签模块数据易于获得；二是类不平衡问题严重：在多数软件项目中，无缺陷模块远多于缺陷模块，会导致传统分类器偏向多数类，降低对缺陷的检出能力。两方面问题明显制约了现有监督式方法的效果和泛化能力。

为缓解标注稀缺，研究者尝试引入半监督学习，利用少量有标签数据与大量无标签数据共同建模。例如，Seliya 等[1]使用 EM 算法迭代估计缺陷标签；Lu 等[2]通过迭代半监督方法发现带标签比例超过 5% 时预测效果优于监督方法；Jiang 等[3]在半监督框架中结合欠采样策略以处理不平衡；Catal 等[4]对比了多种半监督分类器，指出低密度分割方法在大数据集上表现较优。

另一条研究线索是采样与代价敏感学习。通过过采样、欠采样或合成少数类样本来缓解不平衡，例如 SMOTE 方法及其变体。Premalatha 等[5]进一步结合图学习与采样，提出了基于非负稀疏图的标签传播方法（NSGLP），在 NASA 数据集上取得了较优性能。然而，这类方法仍主要依赖于静态代码度量指标，特征表达有限，且稀疏优化带来较高计算开销，限制了其在大规模场景中的应用。

近年来，深度度量学习（deep metric learning, DML）为 SDP 提供了新的研究契机。其核心思想是通过构建判别性嵌入空间，使同类样本在空间中更加接近，不同类样本彼此远离。相比传统特征或浅层模型，DML 能够自动学习高层次表示，并且可以自然结合半监督机制（如伪标签、一致性正则、自训练）来有效利用大量无标签数据。此外，DML 还可与类不平衡处理策略结合，如焦点损失（Focal Loss）、类别重加权（Class-Balanced Loss）、难例挖掘（Hard Mining），从而在度量空间层面提升对少数类的识别能力。

基于以上背景，本文提出一种面向类不平衡的半监督深

1. 广东石油化工学院计算机学院和石化装备智能安全广东省重点实验室 广东茂名 525000
2. 武汉大学计算机学院 湖北武汉 430000

[基金项目] 国家自然科学基金面上项目“人类群体决策引导的多视图表示学习方法和石化动静装备故障诊断与剩余寿命预测应用研究”（62576115）；广东省自然科学基金项目“基于信号文本双模态融合与增强的石化大机组故障诊断方法及应用研究”（2026A1515011903）；广东省自然科学基金面上项目“软件缺陷预测关键技术和新方法研究”（2023A515012653）；广东省教育厅创新团队“石化生产网络安全与大数据分析控制创新团队”（2020KCXTD014）

度度量学习方法（SSDML-SDP）。该方法学习统一的判别性度量空间；在训练过程中引入代价敏感损失和难例挖掘机制，提升对缺陷模块的检出能力；同时结合半监督正则化策略，充分利用无标签模块数据以增强模型泛化能力。

本文的主要贡献如下：

（1）提出了一种结合深度度量学习与半监督机制的缺陷预测框架，能够在有限标注条件下有效利用大量无标签数据；

（2）设计了类不平衡优化策略（代价敏感损失 + 难例挖掘），从度量空间层面提升对少数类缺陷的识别；

（3）在多个公开 SDP 数据集上进行实证研究，结果表明本文方法在 *F*-measure 和 AUC 等指标上显著优于现有代表性半监督方法。

1 相关工作

在软件缺陷预测领域，针对有限标注与类不平衡问题的研究主要分为以下几类：

（1）半监督学习方法。早期工作将半监督学习直接引入 SDP。例如，Seliya 等 [1] 基于期望最大化（EM）算法，迭代为无标签样本分配估计标签，从而扩充训练集；Lu 等 [2] 提出迭代自训练策略，并在带标签样本比例超过 5% 时验证了优于监督方法的效果；Jiang 等 [3] 提出的 ROCUS 方法结合委员会学习与欠采样策略，同时利用无标签样本与类别不平衡信息。Catal 等 [4] 比较了多种半监督分类方法，包括低密度分割（LDS）、EM-SEMI、SVM 与类质量归一化（CMN），指出在数据量有限的情况下，LDS 方法更为适用。Lu 等 [6] 在半监督框架中嵌入多维尺度降维（MDS），提升了模型性能。

（2）类不平衡处理技术。类不平衡是 SDP 中的常见挑战。传统方法多依赖采样：过采样通过复制或合成少数类样本（如 SMOTE[7]），欠采样通过删除部分多数类样本来平衡分布；也有学者探索结合不同采样策略以提升效果 [8]。Pelayo 等 [9] 将合成过采样技术应用于缺陷预测，发现能够显著改善偏斜分布下的性能。近年来，代价敏感学习和集成学习也被引入 SDP，以降低偏向多数类的风险。

（3）图学习与稀疏表示方法。Premalatha 等 [5] 提出的非负稀疏图标签传播（NSGLP）方法是这一方向的代表。他们利用稀疏表示构建模块间相似关系，结合拉普拉斯得分采样平衡类别分布，并在 NASA 数据集上验证了其优于多种代表性方法。NSGLP 的贡献在于同时利用了稀疏性与无标签信息，但其依赖的静态代码度量特征表达能力有限，且稀疏优化计算代价较高，难以扩展到大规模系统。

综上所述，现有方法在缓解有限标注与类不平衡问题上各有探索，但仍存在两点不足：一是多数方法依赖静态度量特征，难以捕捉现代软件的语义特征与复杂结构；二是传统采样或图方法在计算效率与表示能力上受到限制。因此，本文提出的半监督深度度量学习方法（SSDML-SDP）旨在通过深度度量学习自动获取判别性表示，结合半监督机制充分利用无标签样本，并在度量空间中集成类不平衡优化策略，从而在准确性与泛化性上取得更优性能。

2 方法

2.1 问题定义

设训练集包含两部分：

少量有标签模块：

$$D_l=\{(x_i,y_i)\}_{i=1}^{l},\ y_i\in\{0,1\} \tag{1}$$

大量无标签模块：

$$D_u=\{x_j\}_{j=l+1}^{l+u},\ l+u=n \tag{2}$$

式中：y=1 表示缺陷模块，y=0 表示无缺陷模块。

由于缺陷模块远少于无缺陷模块，数据分布呈现严重的类不平衡。目标是在这种条件下，学习一个判别函数，能够在度量空间中区分缺陷与无缺陷模块。

2.2 模块嵌入与度量空间

每个模块 x_i 通过深度网络编码为嵌入向量：

$$\boldsymbol{z}_i=\frac{f_\theta(x_i)}{\|f_\theta(x_i)\|_2}\in\mathbb{R}^d \tag{3}$$

其中归一化保证在单位球面上，便于余弦相似度计算。为每一类引入代理向量（prototype）$\boldsymbol{p}_0$、$\boldsymbol{p}_1$，并约束 $\|\boldsymbol{p}_c\|_2$=1。

2.3 监督损失（类不平衡处理）

对于有标签样本 (x_i, y_i)，通过余弦相似度得到对数几率：

$$s_{i,c}=\frac{\boldsymbol{z}_i^{\top}\boldsymbol{p}_c}{\tau},\ c\in\{0,1\} \tag{4}$$

式中：τ>0，为温度参数。

为缓解类不平衡，引入类别权重 w_c：

$$\mathscr{L}_{\text{sup}}=\frac{1}{|D_l|}\sum_{(x_i,y_i)\in D_l} w_{y_i}\cdot\left(-\log\frac{\exp(s_{i,y_i})}{\sum_k\exp(s_{i,k})}\right) \tag{5}$$

2.4 半监督损失

对无标签样本 x_j，先由当前模型预测其伪标签分布：

$$q_{j,c}=\frac{\exp(\boldsymbol{z}_i^{\top}\boldsymbol{p}_c/\tau)}{\sum_k\exp(\boldsymbol{z}_j^{\top}\boldsymbol{p}_k/\tau)} \tag{6}$$

当 $\max\limits_c q_{j,c}\geq\rho$（置信度阈值）时，取伪标签 $\tilde{y}_j=\arg\max\limits_c q_{j,c}$，并定义无监督损失：

$$\mathscr{L}_{\text{unsup}}=\frac{1}{|D_u|}\sum_{x_j\in D_u} l\{\max_c q_{j,c}\geq\rho\}\cdot H(\tilde{y}_j,q_j) \tag{7}$$

式中：$H(\cdot,\cdot)$ 为交叉熵。

2.5 总目标函数

最终训练目标为监督与半监督部分的加权和：

$$\mathcal{L}=\mathcal{L}_{\text{sup}}+\lambda\mathcal{L}_{\text{unsup}} \tag{8}$$

式中：λ 代表控制无标签样本的影响。

2.6 方法直观解释

本方法的核心在于通过深度度量学习构建判别性嵌入空间，并结合伪标签机制利用无标签数据，同时通过类别权重缓解不平衡问题。

3 实验设计

3.1 数据集与预处理

本文采用公开的 NASA[10] 数据集（如 CM1、KC1、KC3、PC1、PC3~PC5、JM1、MW1、MC2），以模块级静态度量作为输入特征，数据集的具体情况如表 1 所示。

表 1 NASA 数据集的具体情况

项目	度量特征数量	缺陷模块	无缺陷模块数量	缺陷率 /%
CM1	38	42	285	12.84
KC1	22	314	878	26.34
KC3	22	36	158	18.56
PC1	40	61	650	8.58
PC3	38	134	939	12.49
PC4	38	177	1 110	13.75
PC5	39	471	1 220	27.85
JM1	22	1 672	6 083	21.56
MW1	38	27	228	10.59
MC2	40	44	81	35.20

3.2 评价指标

本次工作采用软件缺陷检测领域的主流指标 *F*-measure、AUC 指标进行评价。

具体指标计算为：

Measure（即 F_1 分数）是精确率（Precision）和召回率（Recall）的调和平均数：

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}},\quad \text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{9}$$

$$F_1=\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}}=\frac{2\text{TP}}{2\text{TP+FP+FN}} \tag{10}$$

式中：TP 代表真阳性，FP 代表假阳性，FN 代表假阴性。

AUC（area under the ROC curve）是 ROC 曲线下的面积，等价于随机选取一个正样本和一个负样本时，分类器把正样本排在负样本前的概率。用公式表示为：

$$\text{TPR}=\frac{\text{TP}}{\text{TP+FN}} \tag{11}$$

$$\text{FPR}=\frac{\text{FP}}{\text{FP+TN}} \tag{12}$$

$$\text{AUC}\approx\sum_{i=1}^{n-1}(\text{FPR}_{i+1}-\text{FPR}_i)\cdot\frac{\text{TPR}_{i+1}+\text{TPR}_i}{2} \tag{13}$$

式中：TPR 代表真正率，FPR 代表假正率。

3.3 实验协议

为模拟“少量标注 + 大量无标注”的设定，采用项目内随机划分：

在每个数据集上，先将全部样本随机分为训练 / 测试（70%/30%，分层抽样以保持类比例）；

在训练部分中，再随机抽取$\mu\in\{10\%,20\%,30\%\}$的样本保留标签（形成 D_l），其余视作无标签（形成 D_u）。

为控制变量，仅使用静态度量特征，以突出方法本身对不平衡与半监督的贡献。

3.4 对比方法

本文选取四类代表性基线：FTF（迭代半监督方法）；ROCUS（结合欠采样的半监督委员会学习）；LDS（低密度分割方法）；CMN（类质量归一化方法）。这些方法在科研与工业中充分论证了其优异性，因此作为基准更能说明本次提出方法的优异性。

3.5 训练细节

编码器采用三层全连接网络，隐藏维度分别为 256/128，激活函数为 ReLU，输出嵌入维度 d=128，并进行 L_2 归一化。优化器使用 Adam，学习率 3×10^{-4}，批大小 128，训练轮次 200，在验证集上使用早停策略。伪标签置信度阈值设为 ρ=0.8，无监督损失权重 λ=1.0。类别权重按照类别样本数量的倒数设定，以缓解不平衡问题。

4 实验结果对比

表 2 展示了本文方法与五种代表性基线方法在 10 个 NASA 数据集上的 *F*-measure 与 AUC 的结果。可以从表中看出，本文方法在大多数数据集上均取得了最优性能。

例如在 KC3、PC1、PC5、MW 上的提升尤为明显，相较于当前最佳方法 GKSLP 分别提高了 0.044、0.040、0.071 与 0.077。

在 JM1 数据集上表现与基线接近，说明在大规模且类分布复杂的项目中仍有优化空间。

表中同样给出了 AUC 指标的对比结果。可以看出，本文方法在 10 个数据集中的 9 个中取得了最优结果，平均提升幅度在 0.02~0.05 之间。其中 KC1、KC3、PC4、PC5、MC2 的增益最为明显。

这表明本文方法不仅在 *F*-measure 上表现优越，且能在综合考虑召回率与误报率的情况下保持稳健。

综合 *F*-measure 与 AUC 两个指标，本文提出的 SSD-

ML-SDP 方法在多数数据集上表现优于现有方法，尤其在 KC3、PC1、PC5、MC2 等数据集上的提升最为显著。

这说明通过深度度量学习构建判别性嵌入空间，并结合伪标签机制与类别权重策略，能够在有限标注和类不平衡条件下有效提升缺陷预测性能。

表 2 每个项目的 *F*-measure、AUC 值比较结果

Dataset	FTF[2]	ROCUS[3]	LDS[4]	CMN[4]	GKSLP[11]	本文方法
每个项目的 *F*-measure 值比较结果						
CM1	0.281	0.333	0.326	0.324	0.344	0.387
KC1	0.296	0.432	0.423	0.433	0.413	0.438
KC3	0.333	0.387	0.355	0.368	0.448	0.492
PC1	0.259	0.22	0.267	0.298	0.345	0.385
PC3	0.272	0.288	0.28	0.332	0.358	0.359
PC4	0.382	0.37	0.289	0.438	0.512	0.513
PC5	0.474	0.489	0.45	0.459	0.522	0.593
JM1	0.411	0.463	0.433	0.42	0.434	0.433
MW1	0.216	0.25	0.261	0.218	0.268	0.345
MC2	0.453	0.531	0.546	0.522	0.578	0.585
各项目 AUC 值的比较结果						
CM1	0.634	0.688	0.713	0.665	0.684	0.743
KC1	0.723	0.768	0.766	0.765	0.789	0.822
KC3	0.746	0.733	0.754	0.726	0.812	0.847
PC1	0.688	0.688	0.745	0.615	0.784	0.793
PC3	0.719	0.728	0.733	0.748	0.726	0.76
PC4	0.789	0.8	0.824	0.82	0.825	0.855
PC5	0.878	0.928	0.915	0.91	0.92	0.943
JM1	0.657	0.736	0.711	0.663	0.713	0.719
MW1	0.665	0.677	0.635	0.654	0.691	0.725
MC2	0.841	0.918	0.923	0.88	0.921	0.94

5 总结与未来工作

本文提出了一种面向软件缺陷预测的半监督类不平衡深度度量学习方法（SSDML-SDP）[12]。该方法通过深度度量学习构建判别性嵌入空间，并结合伪标签机制充分利用无标签样本，同时在训练过程中引入类别权重以缓解类不平衡问题。在多个 NASA 公共数据集上的实验结果表明，SSDML-SDP 在 *F*-measure 和 AUC 等指标上均优于现有代表性方法，尤其在缺陷样本稀少的数据集上表现更为突出。这说明在有限标注与严重不平衡的现实条件下，本方法能够有效提升缺陷预测性能。

尽管如此，本研究仍存在一些局限性，值得在未来工作中进一步探索。首先，本文主要聚焦于单项目（within-project）场景，而跨项目（cross-project）缺陷预测更贴近真实应用场景，未来可将方法扩展至跨项目场景，并结合迁移学习与领域自适应技术来增强模型的鲁棒性；其次，实验中仅使用了静态度量特征，而在实际软件工程中，还可以利用代码语义信息、版本演化历史以及缺陷报告等多源特征，未来可以将这些信息融入深度度量空间以提升预测准确性；再次，本文采用固定的类别权重策略来处理不平衡，未来可探索基于样本难度的自适应加权方法，或引入对抗训练机制，以进一步缓解多数类的干扰；最后，本文方法尚未在工业级软件项目中进行验证，未来可在实际开发环境中部署，以评估其在真实场景下的实用性和可解释性。

参考文献：

[1]Seliya N, Khoshgoftaar T M. Software quality analysis of unlabeled program modules with semisupervised clustering[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2007, 37(2): 201-211.

[2]Lu Huihua, Cukic B, Culp M. An iterative semi-supervised approach to software fault prediction[C]//Proceedings of the 7th International Conference on Predictive Models in Software Engineering. NewYork: ACM, 2011: 1-10.

[3]Jiang Yuan, Li Ming, Zhou Zhihua. Software defect detection with ROCUS[J]. Journal of Computer Science and Technology, 2011, 26(2): 328-342.

[4]Catal C, Diri B. A systematic review of software fault prediction studies[J]. Expert Systems with Applications, 2009, 36(4): 7346-7354.

[5]Premalatha H M, Srikrishna C V. Effort estimation in agile software development using evolutionary cost-sensitive deep belief network[J]. International Journal of Intelligent Engineering & Systems (IJIES) 2019, 12(2): 261-269.

[6]Lu Huihua, Cukic B, Culp M. Software defect prediction using semi-supervised learning with dimension reduction[C]// ASE'12: 27th IEEE/ACM International Conference on Automated Software Engineering, NewYork: ACM, 2012: 314-317.

[7] 曾子安，李英梅．基于 PCA-Smote-XGBoost 的软件缺陷预测研究 [J]. 软件工程与应用 ,2024, 13 (3): 346-357.

[8]García S, Herrera F. A study of the behavior of several methods for balancing machine learning training data[J]. Attern Recognition Letters, 2004, 25(8): 1017-1034.

[9]Pelayo L, Dick S. Applying novel resampling strategies to software defect prediction[C]//NAFIPS 2007-2007 Annual meeting of the North American fuzzy information processing society.Piscataway:IEEE, 2007: 69-72.

[10] 王馨煜，崔艺凝，段盈盈．基于 ExtraTree 的软件缺陷预

基于边缘计算的拌和站自动控制系统实时性优化研究

韦相安[1]
WEI Xiang'an

摘 要 面对传统拌和站控制系统在高实时性场景下暴露的通信延迟以及控制响应不稳定问题，文章基于边缘计算开展控制架构的搭建，系统开展了集中式控制链路时延放大以及资源竞争机理分析。凭借提出的边缘节点分布式部署方式、多级任务优先级调度机制以及数据压缩以及同步策略，搭建起边缘—本地—云端协同控制体系。将 1:4 缩比实验平台当作依托来开展 72 h 连续压力测试工作，揭示出边缘侧在高负载以及网络干扰条件下对系统响应以及控制精度的提升效果。研究结果显示，该架构可以极大程度上降低控制响应时延并提升计量稳定性，让系统连续稳定运行的目标得以实现。相关研究为工业现场实时控制系统的结构优化以及性能提升提供了可靠的技术支撑。

关键词 边缘计算；实时控制；任务调度；数据压缩；协同控制

doi：10.3969/j.issn.1672-9528.2026.07.005

0 引言

拌和站作为混凝土生产环节中的核心设备，其自动控制系统需满足复杂工况下对精度与响应时效的双重要求。当前传统拌合站多采用“局域网络 + 现场 PLC”本地控制模式，在实现远程监控与调度过程中需依赖集中式管理平台，受限于公网通信质量、设备兼容性差异与多系统融合改造协调等问题，已难以支撑毫秒级闭环控制任务的时序稳定。边缘计算作为一种贴近数据源、具备低延迟特性的分布式处理模式，为重构控制链路提供新思路[1]。本文聚焦于拌和站生产过程中的实时性优化难题，旨在打破传统架构的性能瓶颈。区别于现有的通用型工业边缘网关研究，本研究的创新性体现在物理工艺特性与计算架构的深度解耦与重构：传统研究多侧重于边缘侧的数据协议转换或通用的流量负载均衡，未能解决拌合过程中“强机械振动干扰”与“毫秒级配料精度”之间的强耦合矛盾。本文通过在边缘侧嵌入具有物理特征感知的动态窗口滤波算法，并配合 PREEMPT_RT 内核的硬实时调度，实现了控制逻辑从“数据驱动”向“工艺特征驱动”的转变，解决了通用架构在极端工况下确定性时序丢失的问题。

1. 南宁兴典混凝土有限责任公司 广西南宁 530000

测方法研究 [J]. 智能计算机与应用 , 2022,12(3):139-142.

[11]Zhu Xiaojin, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions[C]// Proceedings of the 20th International conference on Machine learning (ICML-03). NewYork: ACM, 2003: 912-919.

[12] 陶显 , 侯伟 , 徐德 . 基于深度学习的表面缺陷检测方法综述 [J]. 自动化学报 , 2021, 47(5): 1017-1034.

【作者简介】

姚永芳（1975—），女，湖南常德人，硕士，助理研究员，研究方向：模式识别、人工智能、故障诊断、软件工程。

丁永昌（1997—），男，浙江金华人，博士研究生，研究方向：软件缺陷检测、人工智能。

陈林君（1996—），男，江西九江人，博士研究生，研究方向：类不平衡、模式识别。

廖珂（1997—），男，湖南衡阳人，硕士，研究方向：类不平衡、模式识别。

荆晓远（1971—），通信作者（email:jingxy_2000@126.com），男，江苏南京人，博士，教授，研究方向：模式识别、人工智能、故障诊断、软件工程。

（收稿日期：2025-11-12 修回日期：2026-07-08）