软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(5):1386−1406 [doi: 10.13328/j.cnki.jos.005725]
©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

基于代价极速学习机的软件缺陷报告分类方法*

张天伦[1], 陈 荣[1], 杨 溪[1], 祝宏玉[2]

[1](大连海事大学 信息科学技术学院,辽宁 大连 116026)

[2](深圳大学 计算机与软件学院,广东 深圳 518060)

通讯作者: 陈荣, E-mail: rchen@dlmu.edu.cn

摘 要: 在所有的软件系统开发过程中,Bug 的存在是不可避免的问题.对于软件系统的开发者来说,修复 Bug 最有利的工具就是Bug报告.但是人工识别Bug报告会给开发人员带来新的负担,因此,自动对Bug报告进行分类是一项很有必要的工作.基于此,提出用基于极速学习机的方法来对 Bug 报告进行分类.具体而言,主要解决 Bug 报告自动分类的 3 个问题:第 1 个是 Bug 报告数据集里不同类别的样本数量不平衡问题;第 2 个是 Bug 报告数据集里被标注的样本不充足问题;第 3 个是 Bug 报告数据集总体样本量不充足问题.为了解决这 3 个问题,分别引入了基于代价的有监督分类方法、基于模糊度的半监督学习方法以及样本迁移方法.通过在多个 Bug 报告数据集上进行实验,验证了这些方法的可行性和有效性.

关键词: 软件 Bug 报告;有监督分类方法;半监督学习方法;样本迁移方法;极速学习机

中图法分类号: TP311

中文引用格式: 张天伦,陈荣,杨溪,祝宏玉.基于代价极速学习机的软件缺陷报告分类方法.软件学报,2019,30(5):1386−1406. http://www.jos.org.cn/1000-9825/5725.htm

英文引用格式: Zhang TL, Chen R, Yang X, Zhu HY. Approach of bug reports classification based on cost extreme learning machine. Ruan Jian Xue Bao/Journal of Software, 2019,30(5):1386−1406 (in Chinese). http://www.jos.org.cn/1000-9825/5725.htm

Approach of Bug Reports Classification Based on Cost Extreme Learning Machine

ZHANG Tian-Lun[1], CHEN Rong[1], YANG Xi[1], ZHU Hong-Yu[2]

[1](Information Science and Technology College, Dalian Maritime University, Dalian 116026, China)

[2](College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China)

Abstract: Bug is an unavoidable problem in the development of all software systems. For developers of software system, bug report is a powerful tool for fixing bugs. However, manual recognition on bug reports tends to be time-consuming and not economical. It thus becomes significant to advance the automated classification approach to provide clear guidelines on how to assign a reasonable severity to a reported bug. In this study, several algrithoms are proposed based on extreme learning machine to automatically classify bug reports. Concretely, this study focuses on three problems in the field of bug report classification. The first one is the imbalanced class distribution in bug report dataset; the second is the insufficient labeled sample in bug report dataset; the last is the limited training data available. In order to solve these issues, three methods are proposed based on cost-sensitive supervised classification, semi-supervised learning, and sample transferring, respectively. Extensive experiments on real bug report datasets are conducted, and the results demonstrate the practicability and effectiveness of the proposed methods.

Key words: software bug report; supervised classification method; semi-supervised learning approach; sample transferring approach; extreme learning machine

* 基金项目: 国家自然科学基金(61672122, 61602077, 61732011)
Foundation item: National Natural Science Foundation of China (61672122, 61602077, 61732011)
本文由智能化软件新技术专刊特约编辑申富饶教授和李戈副教授推荐.
收稿时间: 2018-08-31; 修改时间: 2018-10-31, 2018-12-14; 采用时间: 2019-02-03

在软件系统开发领域里,Bug 修复[1]是一个十分关键的环节.近几年,随着软件工程项目的规模与复杂度的提升,在软件开发过程中,不可避免地会出现大量的 Bug,为了系统可以正常运行,修复这些 Bug 是十分必要的;同时,修复这些Bug的任务量也是非常艰巨的.对于软件系统的开发人员而言,在修复Bug时,最常用的辅助工具就是 Bug 报告[2].Bug 报告是以文本的形式来描述 Bug 细节的数据,并且根据任务的不同,这些数据被标上不同的标签.但是Bug报告的质量良莠不齐,如果人工地辨别这些Bug报告的质量,无疑又为软件开发带来了沉重的工作量.所以近几年,很多研究者研究如何将 Bug 报告进行自动分类,其中,Antoniol 等人[3]利用文本挖掘技术来对 Bug 报告进行分类,将 Bug 报告数据分成需即时处理的 Bug 和非即时处理的 Bug,他们所使用的技术是决策树、对率回归以及朴素贝叶斯等方法.Menzies等人[4]利用规则学习技术将Bug报告分成严重Bug和非严重Bug.Tian 等人[5]通过构建一个基于机器学习的预测框架来预测Bug报告的优先级,在这个框架中,他们主要考虑到6个因素:时序、内容、作者、严重程度、产生原因以及相关的 Bug.除此之外,还有学者对 Bug 报告的质量进行分类.Feng 等人[6]通过提取 Bug 报告的特征,来对 Bug 报告进行高质量和低质量的区分,并且将预测得到的信息提交给软件开发者.为了降低Bug的冗余度,Runeson 等人[7]基于信息检索技术提出一种计算Bug报告相似性的方法.类似的,Sun 等人[8]提出一种多特征的信息检索模型,其目的也是为了计算 Bug 报告之间的相似度.近些年,更多的学者关注于 Bug 报告数据集的类别不平衡问题,其中,Lamkanfi 等人[9]通过人工选择的方法来补充数据集,解决数据集的类别不平衡问题,但是人工选择的方法局限性很大,同时也不适合于实际的工作;Yang 等人[10]利用解决不平衡问题的策略来处理 Bug 报告分类问题,即随机欠采样(random under-sampling,简称 RUS)、随机过采样(random over-sampling,简称 ROS)和合成少数类过采样(synthetic minority over-sampling,简称 SMOTE)方法,这些方法在解决不平衡问题时,其效率明显优于人工选择的方法.

本文的关注点也是自动分类 Bug 报告,但是我们更加关注于 Bug 报告分类中存在的一些问题.

首先,在Bug报告数据集里,不同类别的样本的数量不同,而且通常情况下,样本数量之间的差异会造成较大的不平衡度.如果一个数据集的不平衡度较大,那么通过这个数据集训练得到的分类模型的性能就会受到不良的影响.这主要是因为在训练过程中,不平衡的数据集会引导分类器更加侧重于对多数类样例的识别,在更严重的情况下,少数类的样例会被当做噪音点处理.

为了解决 Bug 报告中数据不平衡的问题,我们引入了基于代价的分类器训练策略.传统上的分类器的优化求解目标主要是最大化分类结果的精确度,而基于代价的分类器训练策略则是最小化分类器因错误分类而产生的代价.具体地,不同类别的样例被错误分类的代价应该不同,一些较易被分类的样例应该有较小的错误代价,而那些较难被分类的样例应该有比较大的错误分类代价.在最小化整体的错误分类代价时,就会使得分类器更加关注代价大的样例的分类情况.因为传统的分类器模型默认所有样例的分类代价是一样的,所以基于代价的分类器训练方法是对传统方法的一个扩展.

我们选用的分类器模型是极速学习机(extreme learning machine,简称 ELM).该方法由 Huang 等人[11]提出,因为其训练方法的高效性,目前被普遍用于模式识别领域[12].同时,为了解决上面提到的数据不平衡问题,我们用基于代价的分类器训练方法来训练 ELM 模型,训练得到的模型被称为代价敏感的 ELM 模型.

Bug 报告分类中存在的第 2 个问题是:在数据集里,被标记的样本量不充足.传统的分类器大都是通过有监督算法进行训练的,这就要求数据集里的样本既有条件属性也要有与之对应的标签属性.可是,给样本指定正确的标签需要大量的精细的人工标注工作.通常情况下,人工给数据标注类别属性是繁重且耗时的,所以为了解决这个问题,本文提出一种可以自动标注标签的半监督学习方法.该方法主要利用弱分类器对未知标签的数据进行类别标定,然后结合模糊度[13]这一指标选择不确定性小的数据来扩充原有的有标签的数据集.

在 Bug 报告分类中,第 3 个问题是:数据集的样本总量不充足.在训练分类器时,往往需要充足的训练数据.当训练数据不充分时,会导致训练得到的分类器出现严重的过拟合现象.过拟合现象主要表现为:分类器在训练集上的分类效果良好,可是在测试时或者在实际工作中,其分类效果很不理想.这主要是因为训练样本容量少导致分类器对数据的统计规律学习得不充分,最终使其泛化能力弱,不能对训练集以外的数据做出正确的判别.

为了解决这个问题,我们提出一种样本迁移[14]的方法,即用其他 Bug 数据集里的样本来补充数据量不充足

© 中国科学院软件研究所 http://www.jos.org.cn

的数据集的样本容量.自然地,如何选取样本来补充数据集,这是一个十分关键的问题.在我们的工作中,我们仍旧利用第二个问题里的方法.即先在原数据集上训练得到一个弱分类器,然后用弱分类器对被迁移的样本进行分类,得到分类结果后,结合模糊度,选择模糊度小的样本来补充原始的数据集.在这个过程中,需要注意,不同的数据集的样本维度可能不同.为了统一维度,我们提出使用受限玻尔兹曼机(restricted Boltzmann machines,简称RBM)[15]来对不同数据集的样本进行编码,即将不同数据集里的样本编码成统一维度的数据.这样做使得我们的方法具有更加广泛的适用性.注意到,用来自不同分布的样例来补充训练集合,这个方法的可行性已经在文献[16]得到证明,并且在我们的实验中会进一步被验证.

综上所述,本文的贡献点主要有 3 个方面:(1) 针对 Bug 报告中的类别不平衡问题,我们将基于代价的分类器训练方法引入到 Bug 报告分类问题里,并且使用 ELM 作为基本的分类模型;(2) 针对 Bug 报告中被标记的样本数量不充足的问题,我们将基于模糊度的半监督方法引入到 Bug 报告分类问题里,通过扩充有标签的样本容量来增强分类效果;(3) 针对 Bug 报告中数据集样本总量不充足的问题,我们将基于 RBM 和模糊度的样本迁移方法引入到 Bug 报告分类问题里,通过扩充训练集总样本容量来避免分类器出现过拟合的现象.

本文第 1 节介绍本文涉及到的相关背景知识.第 2 节详细介绍本文提出来的方法.第 3 节展示本文的相关实验.第 4 节对本文的工作进行总结与展望.

1 背景知识

在这一节中,我们将详细介绍本文主体方法所涉及到的背景知识.这些背景知识主要分为两个部分:第 1 个是 ELM 的训练策略;第 2 个是 RBM 的学习方法.

本文主要关注的是对 Bug 报告数据集的分类问题.我们选用的是基于神经网络的分类器.神经网络是一种非线性的函数模型,该模型通过调节网络参数,可以拟合输入到输出的映射关系.近几年,神经网络在模式识别领域里取得了很多令人瞩目的成果[17−19],尤其在自然语言处理的工作里,基于神经网络的方法要明显优于其他方法的效果[20].自动识别 Bug 报告的类别,是一种关于自然语言的类别识别工作,所以我们选用神经网络作为分类模型.

按照训练方式划分,神经网络可以被分为两种类型:第 1 种是基于迭代优化的方法[21]来调整网络的参数;第 2 种是基于随机赋权的方法[22,23]来学习网络的参数.其中,ELM 作为一种随机赋权网络模型,在近几年广泛受到关注.ELM 主要的优点是训练方式高效快捷,并且在大多数领域里,其性能要优于或者等于迭代寻优算法训练出来的网络.接下来,我们详细介绍 ELM 的参数学习过程.

ELM 模型是一个单隐层的前馈神经网络,如图 1 所示.其中,I 是网络的输入层(由 nus 个节点组成),H 是网络的隐层(由 ncc 个节点组成),O 是网络的输出层(由 ns 个节点组成).输入层和隐层之间的权重 $\boldsymbol{\alpha}\in\mathbb{R}^{nus\times ncc}$ 是网络输入层权重,隐层和输出层之间的权重 $\boldsymbol{\beta}\in\mathbb{R}^{ncc\times ns}$ 是网络的输出层权重,为了增加网络模型的线性表达能力,在隐层加入了偏置矩阵 $\boldsymbol{B}=[\boldsymbol{b}_1,\boldsymbol{b}_2,\ldots,\boldsymbol{b}_n]^T$,其中,$\boldsymbol{b}_i=(b_1,\ldots,b_{ncc})^T$.为了提高网络的非线性表达能力,隐层的每一个单元节点需要被非线性函数作用,这里称这个函数为激活函数,一般地,该函数选择的是 sigmoid 函数,即

$$\sigma = sigmoid(x) = \frac{1}{1+\mathrm{e}^{-x}} \tag{1}$$

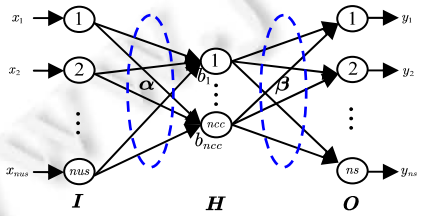


Fig.1 Feedforward neural network model with single hidden layer

图 1 单隐层前馈神经网络模型

© 中国科学院软件研究所 http://www.jos.org.cn

ELM 的训练方法是有监督的方法,因而在训练过程中需要有标签的数据集 $\boldsymbol{S}$.在 $\boldsymbol{S}$ 中,条件属性矩阵用 $\boldsymbol{X}$ 表示,决策属性矩阵用 $\boldsymbol{T}$ 表示,这两个矩阵的形式如下所示.

$$\boldsymbol{X}=\begin{pmatrix} x_{11} & \cdots & x_{1,nus} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{n,nus} \end{pmatrix}_{n\times nus},\boldsymbol{T}=\begin{pmatrix} y_{11} & \cdots & y_{1,ns} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{n,ns} \end{pmatrix}_{n\times ns}.$$

其中,n 代表数据集 $\boldsymbol{S}$ 中的样本总个数;$\boldsymbol{X}$ 中每一行是一个 nus 维的样本向量;$\boldsymbol{T}$ 中每一行是一个 ns 维的标签向量,该向量用 one-hot 形式表示,且 ns 等于数据集里的类别个数.

在实际的分类工作中,I 接收输入矩阵 $\boldsymbol{X}$,经前向传播依次得到 H 的输出 $\boldsymbol{H}\in\mathbb{R}^{n\times ncc}$ 和 O 的输出 $\boldsymbol{Y}\in\mathbb{R}^{n\times ns}$.

$$\boldsymbol{Y}=\sigma(\boldsymbol{X\alpha}+\boldsymbol{B})\boldsymbol{\beta}=\boldsymbol{H\beta} \tag{2}$$

其中,$\boldsymbol{X}$ 是已知的,$\boldsymbol{\alpha}$和 $\boldsymbol{B}$ 是被随机赋值的,只有$\boldsymbol{\beta}$一个是未知量,也就是需要优化求解的网络参数.优化的目标是最小化网络输出 $\boldsymbol{Y}$ 和期望输出 $\boldsymbol{T}$ 之间的距离:

$$\boldsymbol{\beta}^{*}=\arg\min_{\boldsymbol{\beta}}(\boldsymbol{H\beta}-\boldsymbol{T}) \tag{3}$$

$\boldsymbol{\beta}$可以用下面的等式求解.

$$\boldsymbol{\beta}=(\boldsymbol{H}^{T}\boldsymbol{H})^{-1}\boldsymbol{H}^{T}\boldsymbol{T} \tag{4}$$

很明显,通过上式求解$\boldsymbol{\beta}$是一个最小二乘问题.其中,为了更一般化地定义上式的求解过程,可以求 $\boldsymbol{H}$ 的广义逆,即 Moore-Penrose 逆.

接下来将介绍本文所涉及到的另一个数学模型,即 RBM.RBM 是一个概率图模型[24],其结构如图 2 所示.可以看出,RBM 是一个无向二分图模型.其中,$\boldsymbol{v}=(v_1,v_2,...,v_{n_v})$ 是可见层(由 n_v 个节点构成),$\boldsymbol{h}=(h_1,h_2,...,h_{n_h})$ 是隐藏层(由 n_h 个节点构成).RBM 在本文里的主要作用是将可见层的数据编码成隐藏层的表达形式.由于 RBM 的训练机制采用的是无监督的方法,所以训练数据集 $\boldsymbol{S}=\{\boldsymbol{v}^1,\boldsymbol{v}^2,\ldots,\boldsymbol{v}^N\}$里的样本 $\boldsymbol{v}^i$ 不需要标签属性,N 是样本个数.

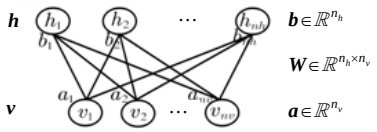


Fig.2 Network structure of RBM

图 2 RBM 网络结构

可见层和隐藏层的条件概率公式可由下面的公式来表达.

$$\begin{cases} p(h_i=1\,|\,\boldsymbol{v})=\sigma\left(b_i+\sum_{j=1}^{n_v}w_{i,j}v_j\right) \\ p(v_j=1\,|\,\boldsymbol{h})=\sigma\left(a_j+\sum_{i=1}^{n_h}w_{i,j}h_i\right) \end{cases} \tag{5}$$

其中,σ的函数形式是 sigmoid 函数.在上面的式子中,$\boldsymbol{a}=(a_1,a_2,...,a_{n_v})$ 是可见层偏置,$\boldsymbol{b}=(b_1,b_2,...,b_{n_h})$ 是隐藏层偏置,$\boldsymbol{W}=[w_{i,j}]_{n_h\times n_v}$ 是 RBM 的网络权重.$\boldsymbol{a}$,$\boldsymbol{b}$ 和 $\boldsymbol{W}$ 是需要学习的参数θ.给定训练集合 $\boldsymbol{S}$,优化目标是求解下式的极大似然.

$$\ln L_{\theta,\mathrm{S}}=\ln\prod_{m=1}^{N}p(\boldsymbol{v}^m)=\sum_{m=1}^{N}\ln p(\boldsymbol{v}^m),$$

其中,$p(\boldsymbol{v}^m)=\frac{1}{Z_\theta}\sum_{\boldsymbol{h}}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})},E_\theta(\boldsymbol{v},\boldsymbol{h})=-\sum_{j=1}^{n_v}a_jv_j-\sum_{i=1}^{n_h}b_ih_i-\sum_{i=1}^{n_h}\sum_{j=1}^{n_v}h_iw_{i,j}v_j$.通过对上式求偏导,参数更新如下:

© 中国科学院软件研究所 http://www.jos.org.cn

$$\begin{cases}\dfrac{\partial \ln L_{\theta,\boldsymbol{S}}}{\partial w_{i,j}}=\sum_{m=1}^{N}\left[p(h_i=1|\boldsymbol{v}^m)v_j^m-\sum_{\boldsymbol{v}}p(\boldsymbol{v})p(h_i=1|\boldsymbol{v})v_j\right]\\ \dfrac{\partial \ln L_{\theta,\boldsymbol{S}}}{\partial a_j}=\sum_{m=1}^{N}\left[v_j^m-\sum_{\boldsymbol{v}}p(\boldsymbol{v})v_j\right]\\ \dfrac{\partial \ln L_{\theta,\boldsymbol{S}}}{\partial b_i}=\sum_{m=1}^{N}\left[p(h_i=1|\boldsymbol{v}^m)-\sum_{\boldsymbol{v}}p(\boldsymbol{v})p(h_i=1|\boldsymbol{v})\right]\end{cases}.$$

可以看出,因为$\sum_v$ 的存在,上式的计算复杂度为 $O(2^{n_v+n_h})$.为了简化计算复杂度,可以采用 Gibbs 采样[24]的方法,经过 t 次采样后,上式可以通过下面的式子近似求解.

$$\begin{cases}\dfrac{\partial \ln L_{\theta,\boldsymbol{S}}}{\partial w_{i,j}}\approx\sum_{m=1}^{N}[p(h_i=1|\boldsymbol{v}^{(m,0)})v_j^{(m,0)}-p(h_i=1|\boldsymbol{v}^{(m,t)})v_j^{(m,t)}]\\ \dfrac{\partial \ln L_{\theta,\boldsymbol{S}}}{\partial a_j}\approx\sum_{m=1}^{N}[v_j^{(m,0)}-v_j^{(m,t)}]\\ \dfrac{\partial \ln L_{\theta,\boldsymbol{S}}}{\partial b_i}\approx\sum_{m=1}^{N}[p(h_i=1|\boldsymbol{v}^{(m,0)})-p(h_i=1|\boldsymbol{v}^{(m,t)})]\end{cases}\tag{6}$$

具体地,在本文里,我们采用 Hinton 等人[25]提出来的对比散度(contrastive divergence,简称 CD)算法来优化 RBM 模型的参数,优化过程如算法 1 所示.

算法 1. RBM 的训练算法.

(1) 输入:训练集 $\boldsymbol{S}$

(2) 初始化:迭代次数 *Iter*,学习率 η,隐层节点数 n_h;随机初始化网络参数($\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}$)

(3) 令:$\boldsymbol{W}^1=\boldsymbol{W},\boldsymbol{a}^1=\boldsymbol{a},\boldsymbol{b}^1=\boldsymbol{b}$

(4) **for** t=1 to *Iter* **do**

通过公式(6)计算参数梯度($\nabla\boldsymbol{W}^t,\nabla\boldsymbol{a}^t,\nabla\boldsymbol{b}^t$)

通过下面的公式更新参数:

$$\begin{cases}\boldsymbol{W}^{t+1}=\boldsymbol{W}^t+\eta\left(\dfrac{1}{N}\nabla\boldsymbol{W}^t\right)\\ \boldsymbol{a}^{t+1}=\boldsymbol{a}^t+\eta\left(\dfrac{1}{N}\nabla\boldsymbol{a}^t\right)\\ \boldsymbol{b}^{t+1}=\boldsymbol{b}^t+\eta\left(\dfrac{1}{N}\nabla\boldsymbol{b}^t\right)\end{cases}$$

end

(5) 令:$\boldsymbol{W}^*=\boldsymbol{W}^{Iter+1},\boldsymbol{a}^*=\boldsymbol{a}^{Iter+1},\boldsymbol{b}^*=\boldsymbol{b}^{Iter+1}$

(6) 返回($\boldsymbol{W}^*,\boldsymbol{a}^*,\boldsymbol{b}^*$)

2 方法设计

在这一节中,我们将详细介绍我们提出来的方法,并且将这个新的方法用于解决 Bug 报告分类中普遍存在的 3 个问题:第 1 个问题是数据类别不平衡问题,第 2 个是训练集里被标记的数据不充足的问题,第 3 个是数据集之间的样本迁移问题.

2.1 基于代价ELM解决数据不平衡问题

ELM 是一种随机赋权网络模型,和其他神经网络一样,是一种非线性函数.在分类问题里,ELM 将条件属性域里的样本映射为决策属性域里的类别标签.在这一点上,ELM 和其他分类器一样,都是通过学习样例到标签之间的映射关系来进行分类工作.ELM 的具体学习过程在第 1 节里已经详细介绍,这个学习过程的优化目标就是

© 中国科学院软件研究所 http://www.jos.org.cn

最小化实际输出和期望输出之间的距离,这个距离通常是二者之间的欧氏距离.在分类问题中,网络的期望输出是样例的类别标签(一般将其表达成 one-hot 形式的向量),通过最小化网络的实际输出向量与期望输出向量之间的欧氏距离,来学习得到网络的参数.

可是,有一个问题需要注意:在分类问题里,每一类的样例被错误分类的代价应该是不同的,比如在不平衡数据集里,多数类样例因为样本个数多,所以更容易被正确分类,而少数类样例则更容易被错误分类,所以一个多数类样例被错误分类的代价应该小于一个少数类样例被错误分类的代价.但是在 ELM 以及大多数分类器的前提假设里,都默认每个类别的样例被错误分类的代价是一样的.

本文主要是解决 Bug 报告的严重性分类问题.在实际工作中,严重的 Bug 需要被即时处理,不严重的 Bug 可以被延时处理.在自动分类 Bug 时,就需要准确预测 Bug 报告的严重程度,将严重的 Bug 识别出来,即时地提交开发人员处理.但是在大多数 Bug 报告数据集里,严重的 Bug 样例个数要明显多于不严重的 Bug 样例个数,因此 Bug 报告数据集通常都是类别不平衡的数据集.这样导致分类器易将不严重 Bug 误报成严重 Bug,降低了分类器对严重 Bug 的识别准确率.为此,我们对 ELM 模型在分类问题上进行扩展,得到一种考虑到错误代价的分类模型.因此,我们引入代价敏感的 ELM 分类模型[26].

为了构建代价敏感的 ELM 分类模型,首先需要构造一个错误分类的代价矩阵 $\boldsymbol{D}$,即

$$\boldsymbol{D}=\begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1,ns} \\ d_{21} & d_{22} & \cdots & d_{2,ns} \\ \vdots & \vdots & \ddots & \vdots \\ d_{ns,1} & d_{ns,2} & \dots & d_{ns,ns} \end{pmatrix}_{ns\times ns}.$$

在这个矩阵里,$d_{i,j}$ 代表第 i 类样例被错误分到第 j 类的代价大小,很明显地,$d_{i,j}$ 等于零,也就是第 i 类样例被正确分类时所产生的代价是零.我们用 c_i 代表第 i 类样例被错误分类的代价,c_i 可以用下面的公式求得.

$$c_i=\sum_{j=1}^{ns} d_{i,j} \tag{7}$$

假设数据集里的类别个数一共是 ns,那么通过等式(7),所有类别的样例被错误分类的代价可以被集合 $\boldsymbol{Cost}$ 表示:$\boldsymbol{Cost}=(c_1,c_2,\dots,c_{ns})$.

通过引入代价矩阵,代价敏感的 ELM 将原始的 ELM 扩展成了一个考虑到错误分类代价的分类模型.在优化过程中,代价敏感的 ELM 将优化目标扩展成为一个带有代价权重的欧氏距离之和.为了表示这个优化目标,首先将训练集里的样例按照类别分组,同上,这里假设类别个数为 ns,每一组样例同属于一个类别,这个组可以被表示为 $\boldsymbol{X}_i$,被分组后的数据集可以被表示为 $\boldsymbol{X}=(\boldsymbol{X}_1^T,\boldsymbol{X}_2^T,\dots,\boldsymbol{X}_{ns}^T)^T$,其对应的标签矩阵为 $\boldsymbol{T}$,对应的隐层矩阵 $\boldsymbol{H}$ 可以由下面的公式求出.

$$\boldsymbol{H}=(\boldsymbol{H}_1^T,\boldsymbol{H}_2^T,\dots,\boldsymbol{H}_{ns}^T)^T=\sigma(\boldsymbol{X}\alpha+\boldsymbol{B}) \tag{8}$$

代价敏感的 ELM 的优化目标是最小化分类错误代价,通过上面给出的一些符号化定义,这个优化目标可以被表示为

$$\boldsymbol{\beta}^*=\arg\min_{\beta}\left\|\begin{pmatrix} c_1H_1\beta \\ c_2H_2\beta \\ \vdots \\ c_{ns}H_{ns}\beta \end{pmatrix}-\begin{pmatrix} c_1T_1 \\ c_2T_2 \\ \vdots \\ c_{ns}T_{ns} \end{pmatrix}\right\|=\arg\min_{\beta}(c_1\|H_1\beta-T_1\|+c_2\|H_2\beta-T_2\|+\dots+c_{ns}\|H_{ns}\beta-T_{ns}\|) \tag{9}$$

这是一个加权的最小二乘解问题,令 $\boldsymbol{H}_c=(c_1\boldsymbol{H}_1^T,c_2\boldsymbol{H}_2^T,\dots,c_{ns}\boldsymbol{H}_{ns}^T)^T$,$\boldsymbol{T}_c=(c_1\boldsymbol{T}_1^T,c_2\boldsymbol{T}_2^T,\dots,c_{ns}\boldsymbol{T}_{ns}^T)^T$.在求解这个优化问题时,同样可以用最小二乘法来求解最优解,即

$$\boldsymbol{\beta}^*=(\boldsymbol{H}_c^T\boldsymbol{H}_c)^{-1}\boldsymbol{H}_c^T\boldsymbol{T}_c \tag{10}$$

和第 1 节介绍的 ELM 参数优化求解方法一样,这里的矩阵求逆可通过 Moore-Penrose 广义逆来求解得到.

在本文所要解决的 Bug 报告分类问题中,数据集都是二分类问题,其中多数类是严重的 Bug 报告,少数类是不严重的 Bug 报告.通常,这两类之间会存在很大的不平衡度[9,10],因此,我们在这里提出来用代价敏感的 ELM 来

© 中国科学院软件研究所 http://www.jos.org.cn

分类 Bug 报告数据集里的样例,通过给少数类样例比较大的错误分类代价,给多数类样例比较小的错误分类代价,并且最小化错误分类的总代价,来自动地解决 Bug 报告数据集在分类过程里存在的类别不平衡问题.

其中,所有数据集上的代价矩阵是一个 2×2 的方阵,对角线上元素全为 0,代表着样例被正确分类的代价是 0,其他元素则分别代表着少数类样例被错分类到多数类的错误代价 c_a 和多数类样例被错分到少数类的错误代价 c_b.在设置代价矩阵时,c_a 要大于 c_b.之后,通过最小化错误分类的总代价,可以使 ELM 模型对少数类的错误分类更加敏感,从而达到解决分类中的类别不平衡问题.用代价敏感的 ELM 来解决 Bug 报告分类中不平衡问题的过程可以用算法 2 来表示.

算法 2. 代价敏感的 ELM 的训练算法.

(1) 输入:条件属性矩阵 $\boldsymbol{X}$,标签属性矩阵 $\boldsymbol{T}$ //输入的矩阵已按类别分组

(2) 初始化:隐层节点数 ncc,代价矩阵 $\boldsymbol{D}$;随机初始化:输入权重$\boldsymbol{\alpha}$,隐层偏置 $\boldsymbol{B}$

(3) 用公式(8)计算隐层输出值矩阵 $\boldsymbol{H}$

(4) 通过公式(10)求解$\boldsymbol{\beta}^*$

(5) 返回$\boldsymbol{\beta}^*$

2.2 基于半监督方法处理训练数据不充分问题

本文之前介绍的ELM方法和代价敏感的ELM方法都是有监督学习的方法,监督学习的方法需要训练数据既有条件属性也要有与之相对应的标签属性.但是在软件测试的过程中,条件属性容易获得,与之对应的标签属性则需要精密的人工标注才可以获得.面对这种情况,我们在本节提出一种基于模糊度的半监督学习方法[26],并且用这种方法来解决软件 Bug 报告的分类问题.在文献[26]里已经证明,高模糊性的样例的分类代价要明显大于低模糊性的样例的分类代价,所以这里使用低模糊性的样例来填充原数据集可以避免整体的分类代价过高,而且 Wang 等人[13]提出,模糊性与不确定性存在着正相关的关系,低模糊性的样例具有低的不确定性,也就是被正确分类的可能性较大,这也是我们选择低模糊性样例来填充数据集的原因.

如上所述,在软件测试过程中会产生很多无标签的数据,而软件 Bug 报告分类问题主要就是通过学习条件属性到标签数据之间的映射关系,使得分类器可以自动分类这些数据.目前,在机器学习领域都是使用数据驱动的分类方法,这就需要在训练这些分类器时,可以提供大量的有标签的训练数据.本文里,这些数据的标签只有严重和非严重两类.但是要想获取这些标签,就需要大量的人工标注.人工标注的工作往往耗时而且代价昂贵,这就很自然地使人们想到用未标注的数据来解决软件测试 Bug 报告的分类问题.

在利用未标注的软件测试 Bug 报告数据时,我们提出了一个方法,即通过在有限的有标注的数据集上训练出一个弱分类器,然后用这个弱分类器对未标注的数据进行分类,得到分类结果后,我们通过模糊度来衡量这些分类结果的可信程度,选择可信程度高的数据来补充原始的训练数据集,而这些数据的标签属性则是通过弱分类器给出.

理论上,上述过程可以重复多次,每一次都从未标注的数据集里挑选出可信程度大的数据来补充训练数据集.因为训练数据集的样本容量在不断地扩充,所以训练出来的分类器的分类性能也会逐渐提高.最终的理想结果是,未标注的样例全部被自动地加上标签并且被扩充到训练集里.通过这个完备的数据集,可以训练得到一个泛化能力很强的分类器.但是需要注意:当分类器的分类精度不是很高时,错误分类的情况很可能会发生,也就是未标注的样例在这种情况下很可能被标上错误的标签,这样就会干扰分类器性能的提升.因而,要想达到之前所述的理想的情况,每次选择多少未标注的样本来扩充训练集,这是一个很关键的问题.

在实现这个方法的过程中,最关键的指标是未标注数据的模糊度的测量.模糊度的计算有多种方法,Zadeh 等人[27]总结出了一些计算模糊度的方法,其中包括基于 Hamming 距离或者欧式距离的模糊性的计算方法、针对大型多类问题的模糊性计算方法.在本文中,我们利用下面的方法来计算模糊度的大小.

$$\mathcal{F}(\boldsymbol{\mu}_i) = -\frac{1}{ns}\sum_{j=1}^{ns}(\mu_{ij}\log\mu_{ij} + (1-\mu_{ij})\log(1-\mu_{ij})) \tag{11}$$

其中,向量$\boldsymbol{\mu}_i=(\mu_{i1},\mu_{i2},\ldots,\mu_{i,ns})$,$\mu_{ij}$ 是第 i 个样本属于第 j 类的隶属度大小.

© 中国科学院软件研究所 http://www.jos.org.cn

因为考虑到了数据集里类别不平衡的因素,所以在实现本节的方法时,分类器采用的是代价敏感的 ELM 模型.但是注意到,在计算模糊度时,需要知道样例属于每一个类别的后验概率.可是在 ELM 的原始设计中,网络输出的值 $\boldsymbol{Y}=[y_{ij}]_{n\times ns}$ 不能表达这个信息.基于此,我们引入了 softmax,将该函数作为激活函数作用到网络的输出层,计算公式如下所示.

$$\mu_{ij}=\frac{\mathrm{e}^{y_{ij}}}{\sum_{k=1}^{ns}\mathrm{e}^{y_{ik}}}.$$

易证, $\mu_{ij}>0,\sum_{j=1}^{ns}\mu_{ij}=1$.对于用 softmax 的输出来表示后验概率的合理性说明,请参考附录中的相关内容.

算法 3 详细表达了上述的基于模糊度的半监督方法.

算法 3. 半监督训练算法.

(1) 输入:有标签的条件属性矩阵 $\boldsymbol{X}$,标签属性矩阵 $\boldsymbol{T}$;无标签的条件属性矩阵 $\boldsymbol{Z}$

(2) 初始化:隐层节点个数 ncc,代价矩阵 $\boldsymbol{D}$;随机初始化 ELM 的输入层权重和隐层偏置

(3) 根据算法 2,利用数据 $\boldsymbol{X}$ 和 $\boldsymbol{T}$ 训练一个代价敏感的 ELM:CELM

(4) 利用 CELM 对样本矩阵 $\boldsymbol{Z}$ 进行分类

(5) 得到 $\boldsymbol{Z}$ 上的分类结果后,通过公式 11 计算每个样本的模糊度

(6) 对模糊度进行升序排列,选择前 k 个样例 z_x 补充数据集 $\boldsymbol{X}$,
用 z_x 被 CELM 分类得到的标签 z_y 补充 $\boldsymbol{T}$

(7) 在补充后的数据集上重新根据算法 2 训练代价敏感的分类器

(8) 返回新得到的分类器的参数矩阵$\boldsymbol{\beta}$

2.3 Bug报告数据集间的样例迁移方法

第 2.2 节介绍了一种基于模糊度的半监督方法,这个方法主要为了解决训练数据中标注样本少的问题.这个方法的提出是基于一个假设,即未标注的样本和少量的有标注的样本来自同一个数据集.但是在软件测试 Bug 报告的数据集里,不仅被标注的样本量少,而且整体的样本总量也很少.在训练分类器时,训练数据集的样本量不充足会带来的问题是使得训练得到的分类器容易过拟合,也就是该分类器在训练集上表现良好,可是其泛化能力很弱.在实际工作中,泛化能力弱的分类器在处理新的待分类样本时,往往会给出错误的决策.

为了解决 Bug 报告中有些数据集的样本总量过少的问题,我们在本节提出一种基于模糊度的样例迁移方法.该方法不需要被迁移的样例与已有样例来自同一数据集,即为了扩充训练集的样本容量,可以使用该方法将别的数据集里的样例迁移到训练集里来扩充样本容量.

该方法的好处就是可以解决因为样本容量少而导致的过拟合问题,但是在选择样例上,该方法存在着和上一节一样的问题:如果样本选择不好,则同样会导致过拟合的问题出现.为了更加客观地选择样本,我们先在已有的数据集上训练得到一个弱分类器,然后用该分类器对被迁移数据集进行分类,得到分类结果后,可以计算每个样例的模糊度,模糊度越大,不确定性越大,所以选择模糊度小的样例来补充原有的数据集,这些样例的标签通过弱分类器得到.然后,在扩展后的数据集上再次训练分类器.这个过程可以一直重复,直到得到合适的训练样本容量.

需要注意的是,为了避免数据集里类别不平衡问题的干扰,在这个方法里,我们选择的分类器同样是代价敏感的 ELM 模型.在实际工作中,另一个需要解决的问题是不同数据集的样本维度通常不一样,这个问题限制了这个方法的应用范围.为了解决这个问题,使得我们的方法更具一般性,我们将 RBM 引入到该方法中.

在第 1 节里已经介绍了 RBM.RBM 可以被视为一种样本编码工具,即可以通过非线性变换,将样本从一个维度空间映射到另一个维度空间.在编码的过程中,RBM 保证编码前和编码后的样本是等价的.在本节的工作里,我们采用 RBM 将不同数据集里的样本编码成统一的维度,这样就为分类器在不同数据集上进行迁移训练提供了可能性.我们用 $\boldsymbol{S}=\{\boldsymbol{X},\boldsymbol{T}\}$表示原数据集;同时,用 $\boldsymbol{S}_t$ 表示被迁移数据集 $\boldsymbol{S}_t=\{\boldsymbol{X}_t\}$.基于这些符号化的定义,该节

© 中国科学院软件研究所 http://www.jos.org.cn

提出来的方法可以用算法 4 来详细表述.

算法 4. 样本迁移算法.

(1) 输入:数据集 $\boldsymbol{S}$,其中,条件属性矩阵 $\boldsymbol{X}$,标签属性矩阵 $\boldsymbol{T}$,被迁移的数据集 $\boldsymbol{S}_t=\{\boldsymbol{X}_t\}$

(2) 初始化:ELM 的隐层节点个数 ncc,代价矩阵 $\boldsymbol{D}$,RBM 的隐层节点个数 n_h.
随机初始化:ELM 的输入层权重和隐层偏置

(3) 通过算法 1,利用数据 $\boldsymbol{X}$ 和 $\boldsymbol{X}_t$ 分别训练得到两个 RBM 模型

(4) 通过训练得到的 RBM 对 $\boldsymbol{X}$ 和 $\boldsymbol{X}_t$ 进行编码,得到新的数据表达形式 $\boldsymbol{X}_p$ 和 $\boldsymbol{X}_{tp}$

(5) 根据算法 2,利用数据 $\boldsymbol{X}_p$ 和 $\boldsymbol{T}$,训练得到一个代价敏感的 ELM:CELM

(6) 利用 CELM 对数据集 $\boldsymbol{X}_{tp}$ 里的样本进行分类

(7) 得到 $\boldsymbol{X}_{tp}$ 上的分类结果后,通过公式(11)计算每个样本的模糊度

(8) 对模糊度进行升序排列,选择前 k 个样例 tr_x 补充数据集 $\boldsymbol{X}_p$,
用 tr_x 被 CELM 分类得到的标签 tr_y 补充 $\boldsymbol{T}$

(9) 在补充后的数据集上重新根据算法 2 训练代价敏感的分类器

(10) 返回新得到的分类器的参数矩阵$\boldsymbol{\beta}$

3 本文实验

本文针对软件测试 Bug 报告数据集里存在的 3 个问题提出了 3 个解决方法,即针对训练数据集类别不平衡问题,提出了代价敏感的 ELM 分类方法;针对数据集被标记的样本较少的问题,提出了基于模糊度的半监督方法;针对训练数据集总体样本量较少的问题,提出了基于模糊度和 RBM 的样本迁移方法.为了验证我们提出的方法的有效性,我们开展了 3 组验证性的实验.

在本节中,实验所采用的数据集是真实的软件测试 Bug 数据集.数据集来自 3 个开源项目.本节实验里用到的数据集一共 21 个,其中,7 个 Eclipse[28]、7 个 Moizlla[29]、7 个 GNOME[30].数据集的名称以及数据集的不平衡度都在表 1~表 3 中被列出.其中,不平衡度 Imbalance Ratio(IR)的计算如下所示.

$$IR=num_maj/num_\mathrm{min}.$$

其中,num_min 是少数类的样本个数,num_maj 是多数类的样本个数.

Table 1 Datasets of Eclipse Bug reports

表 1 Eclipse Bug 报告数据集

项目	名称		IR
Eclipse	1	CDT_cdt-core	4.017 5
	2	JDT_Core	2.568 4
	3	JDT_Debug	2.426 1
	4	PDE_UI	2.663 3
	5	Platform_Debug	1.741 4
	6	Platform_SWT	4.923 2
	7	Platform_UI	2.542 2

Table 2 Datasets of GNOME Bug reports

表 2 GNOME Bug 报告数据集

项目	名称		IR
GNOME	1	ekiga_general	9.5
	2	Evolution_Calendar	4.626 2
	3	Evolution_Contacts	2.776 4
	4	Evolution_Mailer	2.902
	5	Evolution_Shell	2.444 4
	6	gnome-panel_Panel	3.942 4
	7	gnome-terminal_general	12.674 2

© 中国科学院软件研究所 http://www.jos.org.cn

Table 3 Datasets of Moizlla Bug reports

表 3 Moizlla Bug 报告数据集

项目	名称		IR
Moizlla	1	Core_Printing	7.018
	2	Core_XPCOM	2.020 1
	3	Core_XPConnect	5.3
	4	Core_XUL	4.090 2
	5	Thunderbird_General	4.051 1
	6	Core_Layout	2.861 5
	7	Core_Security_UI	6.267 2

在接下来的实验里,我们采用的评价指标分别是准确率和加权的 *F*-measure[31].其中,准确率是样本被正确分类的比率;加权的 *F*-measure 是两类样例在查全率和查准率上的一个综合性的指标,在本文实验里,该指标更能反映出分类器对严重 Bug 的误报程度.准确率的计算如下所示:

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN},$$

其中,*Accuracy* 是准确率,*TP*、*FP*、*TN* 和 *FN* 的关系可以用表 4 表示.

Table 4 Confusion matrix

表 4 混淆矩阵

Confusion matrix		Actual severity	
		non-severe	severe
Predicted severity	non-severe	TP: true positives	FP: false positives
	severe	FN: false negatives	TN: true negatives

F-measure 的计算如下:

$$F\text{-}meausre=\frac{(1+\gamma^2)\times Precision\times Recall}{\gamma^2\times Precision+Recall},$$

其中,在本文实验里,γ等于 1,*Precision* 代表查准率,*Recall* 代表查全率,这两个指标的计算可由下面的公式表达.

$$Precision=\frac{TP}{TP+FP},Recall=\frac{TP}{TP+FN}.$$

本文所用到的加权 *F*-measure 计算如下:

$$wF=W\cdot F\text{-}measure,$$

其中,*wF* 是加权的 *F*-measure;$W=(w_1,w_2)$,w_1 是第 1 类样例占的比例,w_2 是第 2 类样例占的比例.

本节的第 1 个实验是验证代价敏感的 ELM 在不平衡数据集上的分类效果.作为对比,我们用传统的 ELM 在相同的数据集上进行实验.除此之外,我们还比较了基本的分类器,即朴素贝叶斯模型(NB)、BP 神经网络模型(BPNN)和支持向量机(SVM)[32].数据集的划分是:把数据集平均分成 5 份,每次随机取 3 份作为训练集,剩下的 2 份作为测试集.为了保证客观性,我们对两个方法都进行了 100 次实验,最后的结果是取这 100 次实验的平均值.在每次实验里,为了保证公平性,代价敏感的 ELM 和传统的 ELM 的隐层节点数以及 BPNN 的隐层节点数都被设置成一样的个数,并且两个模型的输入层权重和隐层偏置也都被相同的随机数赋值.我们在实验里给权重和偏置随机赋权的方法是在区间[0,1]里随机取值.表 5~表 10 给出了这组实验的对比结果.

Table 5 Testing accuracies on Eclipse datasets

表 5 Eclipse 数据集上的测试准确率

ACC	NB	BPNN	SVM	ELM	C-ELM
Eclipse_1	0.590 4	0.579 4	0.454 9	0.676 8	**0.728 6**
Eclipse_2	0.517 8	0.549 2	0.551 9	0.581 7	**0.594 9**
Eclipse_3	0.557 5	0.57	0.562 5	0.591	**0.606 9**
Eclipse_4	0.583 5	0.588 1	0.526 3	0.633 4	**0.651 6**
Eclipse_5	0.503 9	0.593 0	0.554 3	0.605 4	**0.606 9**
Eclipse_6	0.609 7	0.617 8	0.573 3	0.658 4	**0.674 2**
Eclipse_7	0.4967	0.592 2	0.555 0	0.601 0	**0.611 8**

© 中国科学院软件研究所 http://www.jos.org.cn

Table 6 Weighted *F*-measures on Eclipse datasets

表 6 Eclipse 数据集上的加权 *F*-measures

WF	NB	BPNN	SVM	ELM	C-ELM
Eclipse_1	0.612 3	0.615 0	0.501 6	0.648 1	**0.676 4**
Eclipse_2	0.528 1	0.568 7	0.569 9	0.556 1	**0.570 7**
Eclipse_3	0.576 8	0.566 6	**0.578 9**	0.559 5	0.575 8
Eclipse_4	0.532 8	0.575 7	0.549 8	0.580 7	**0.596 1**
Eclipse_5	0.454 6	0.531 0	0.516 9	0.529 4	**0.531 3**
Eclipse_6	0.657 8	0.660 3	0.626 5	0.659 1	**0.673 2**
Eclipse_7	0.504 2	0.576 7	0.576 3	0.566 2	**0.578 6**

Table 7 Testing accuracies on GNOME datasets

表 7 GNOME 数据集上的测试准确率

ACC	NB	BPNN	SVM	ELM	C-ELM
GNOME_1	0.642 3	0.785 4	0.714 1	0.837 6	**0.858 3**
GNOME_2	0.468 8	0.686 5	0.523 4	0.781 0	**0.782 9**
GNOME_3	0.420 7	0.607 0	0.578 3	0.674 1	**0.679 1**
GNOME_4	0.561 3	0.606 3	0.549 0	**0.726 6**	0.725 2
GNOME_5	0.467 7	0.611 4	0.571 8	0.728 1	**0.730 3**
GNOME_6	0.459 4	0.657 0	0.610 1	0.761 6	**0.770 5**
GNOME_7	0.475 5	0.714 6	0.709 1	0.763 0	**0.780 0**

Table 8 Weighted *F*-measures on GNOME datasets

表 8 GNOME 数据集上的加权 *F*-measures

WF	NB	BPNN	SVM	ELM	C-ELM
GNOME_1	0.715 8	0.810 8	0.670 7	0.804 1	**0.817 9**
GNOME_2	0.521 0	0.701 8	0.574 0	0.697 0	**0.702 6**
GNOME_3	0.430 6	0.591 3	0.601 4	0.595 6	**0.602 8**
GNOME_4	0.584 3	0.607 1	0.574 5	0.610 5	**0.613 8**
GNOME_5	0.466 1	0.544 0	0.580 9	0.585 4	**0.593 7**
GNOME_6	0.499 0	0.674 5	0.675 5	0.669 7	**0.679 3**
GNOME_7	0.585 1	0.775 9	0.777 8	0.783 2	**0.795 1**

Table 9 Testing accuracies on Moizlla datasets

表 9 Moizlla 数据集上的测试准确率

ACC	NB	BPNN	SVM	ELM	C-ELM
Moizlla_1	0.730 5	0.678 4	0.561 3	0.751 2	**0.772 0**
Moizlla_2	0.468 3	0.595 0	0.575 8	0.592 2	**0.622 4**
Moizlla_3	0.666 7	0.666 7	0.647 1	0.714 0	**0.727 1**
Moizlla_4	0.682 5	0.714 3	0.682 5	0.740 6	**0.782 1**
Moizlla_5	0.558 2	0.551 9	0.544 0	0.580 2	**0.604 7**
Moizlla_6	0.589 6	0.580 9	0.613 2	0.635 6	**0.645 9**
Moizlla_7	0.563 4	0.625 4	0.607 7	0.696 6	**0.726 9**

Table 10 Weighted *F*-measures on Moizlla datasets

表 10 Moizlla 数据集上的加权 *F*-measures

WF	NB	BPNN	SVM	ELM	C-ELM
Moizlla_1	0.758 8	0.722 8	0.634 1	0.749 7	**0.762 3**
Moizlla_2	0.516 0	0.642 9	0.627 7	0.622 2	**0.644 9**
Moizlla_3	0.711 9	0.687 0	0.691 7	0.701 1	**0.713 3**
Moizlla_4	0.699 3	0.701 9	**0.709 0**	0.683 5	0.707 1
Moizlla_5	0.600 3	0.593 5	0.589 7	0.598 4	**0.616 3**
Moizlla_6	0.614 4	0.580 9	**0.636 9**	0.593 6	0.603 2
Moizlla_7	0.626 3	0.682 4	0.667 6	0.700 4	**0.721 8**

通过表 1~表 3 的最后一列可以看出,Bug 报告数据集都是有着不平衡度的数据集,而且有些数据集的不平衡度较大.在处理数据类别不平衡这个问题时,通过给少数类样例更大的错误分类代价,得到的分类准确率在绝大多数 Bug 报告数据集上有了很明显的提升.虽然在表 7 中,GNOME_Evolution_Mailer 的准确率结果要偏低,但是通过加权的 *F*-measure 值的比较可以看出,代价敏感的分类器在重要数据样本的分类上面表现仍旧良好,

© 中国科学院软件研究所 http://www.jos.org.cn

也就是对严重Bug的误报率较低.而且通过在所有数据集上比较加权的*F*-measure值可以看到,基于代价的分类模型都能取得很好的效果.与其他传统的分类器相比,ELM 方法需要设定的参数更少,尤其与 BPNN 的比较中,ELM 的训练时间要远远低于 BP 迭代调参所需要的时间,并且在绝大多数数据集上,ELM 方法可以取得较好的效果.在图 3 中,我们比较了欠采样方法(RUS-ELM)、过采样方法(ROS-ELM)、合成过采样方法(SMOTE)与代价敏感方法.欠采样方法是对多数类样例进行随机删减,可以看出,因为减少了训练样本的容量,并且在删减样本的过程中容易丢失重要数据,所以该方法效果不是很好;过采样的方法则是对少数类样例进行扩充,该方法可能会受到冗余数据的影响,造成分类器过拟合,对分类器泛化能力造成不好的影响;SMOTE 方法相对表现较好,但是该方法容易造成类间的 overlapping 现象,且在类别临界面生成样例时,容易将合成的多数类样例标记成少数类类别,从而不利于分类器对数据分布特点的学习.通过比较,基于代价敏感的平衡方法在绝大部分数据集上都有较好的结果,而且基于代价的方法表现得更加稳定,更适合被用于实际的工作中.从效率上来看,基于代价的方法不需要事先对数据集进行平衡化的预处理,而是可以进行端到端的训练,从而更加满足本文研究领域的需求.

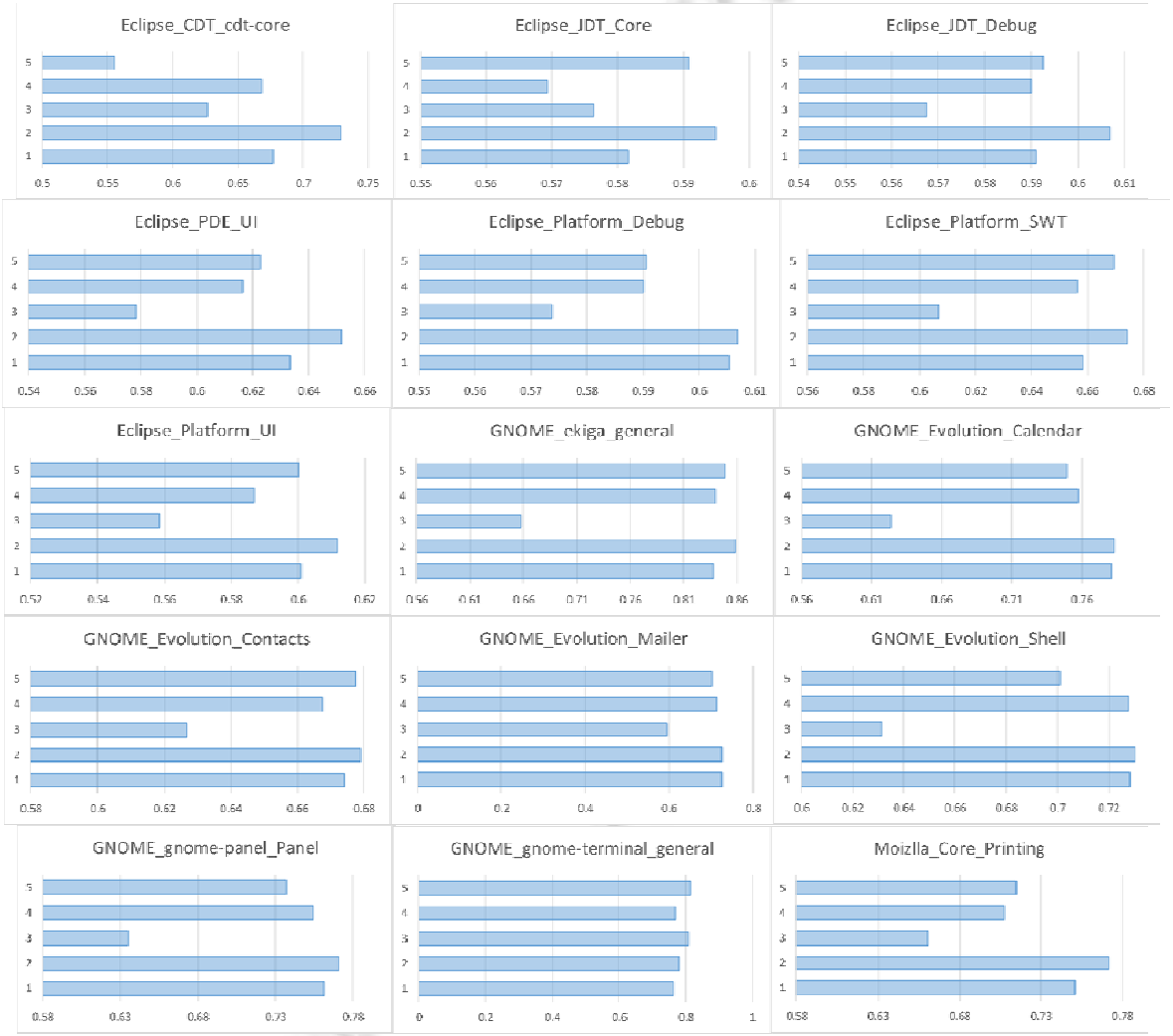


Fig.3 Comparison among RUS, ROS, SMOTE and cost-based method

图 3 欠采样、过采样、合成过采样和代价敏感方法的比较

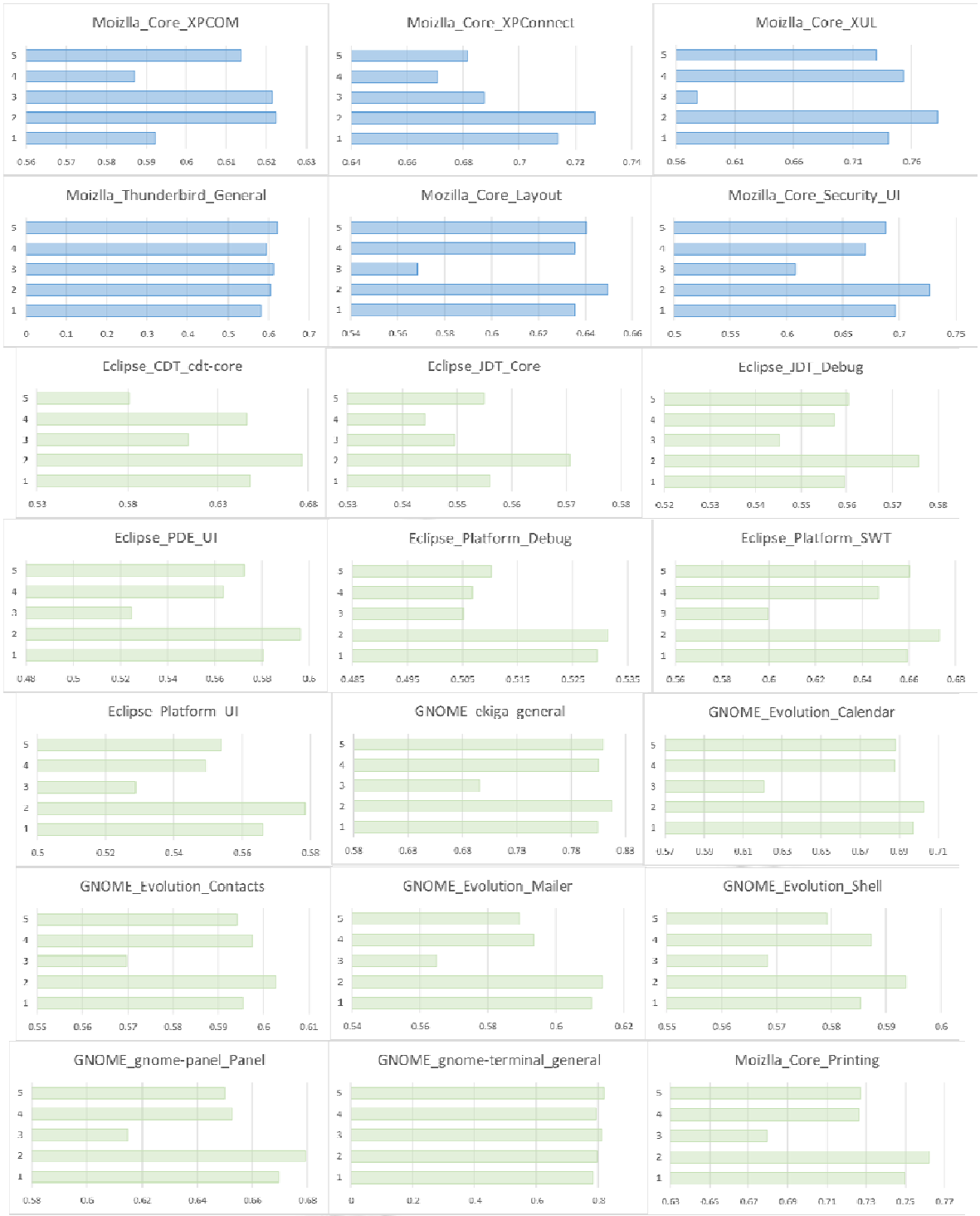


Fig.3 Comparison among RUS, ROS, SMOTE and cost based method (Continued 1)

图 3 欠采样、过采样、合成过采样和代价敏感方法的比较(续 1)

© 中国科学院软件研究所 http://www.jos.org.cn

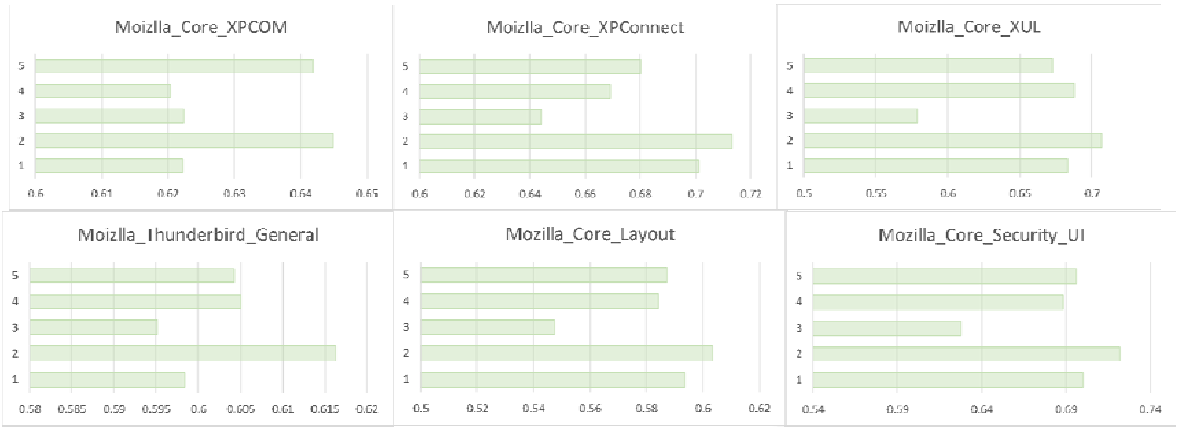


Fig.3 Comparison among RUS, ROS, SMOTE and cost based method (Continued 2)

图 3 欠采样、过采样、合成过采样和代价敏感方法的比较(续 2)

图 3 中的结果都是实验重复 100 次得到的平均值,蓝色图是准确度的结果,绿色图是加权 *F*-measure 值的结果.1~5 分别代表 ELM、CELM、RUS-ELM、ROS-ELM 和 SMOTE-ELM.

本节的第 2 个实验是验证基于模糊度的半监督方法在扩充训练集样本容量上的有效性.在这个实验里,我们采用的分类器是代价敏感的 ELM 模型,目的是为了避免数据集里不平衡问题的干扰.在该实验里,对原始数据集的划分是:将数据集划分成 6 份,每次随机选择其中的 3 份作为训练集,在剩下的数据中,随机选择 2 份作为测试集,最后留下的数据作为验证集.其中,验证集则是本文第 2.2 节里提到的未标注数据集.在实验过程中,我们先在原始的训练集上训练得到一个弱分类器,然后在验证集上对样例进行分类,之后通过模糊度的大小选择一定数量的验证集样例来扩充原始的训练集,然后用扩充得到的数据集再次训练一个分类器.为了保证客观性,前后两个分类器的各项参数的设置完全一样,具体操作方法如第 1 个实验所述.为了通过对比显示我们方法的有效性,我们用两次训练得到的分类器对测试集的样本进行分类,最后通过上述两个指标来评价分类结果,该结果同样是对 100 次的实验结果求平均值得到的最终结果.我们用 First 表示在原始数据集上训练得到的分类器的测试结果,用 Second 表示用扩充后的数据集训练的分类器的测试结果.这部分的实验见表 11.

Table 11 Experimental results of semi-supervised approach

表 11 半监督方法的实验结果

数据集名称	准确度		wflscore	
	First	Second	First	Second
Eclipse_CDT_cdt-core	0.682 4	**0.699 7**	0.694 1	**0.703 7**
Eclipse_JDT_Core	0.636 7	**0.641 7**	0.649 2	**0.653 5**
Eclipse_JDT_Debug	0.638 2	**0.639 2**	0.641 4	**0.642 2**
Eclipse_PDE_UI	0.641 7	**0.642 4**	0.650 8	**0.650 9**
Eclipse_Platform_Debug	0.596 4	**0.599 5**	0.597 1	**0.599 5**
Eclipse_Platform_SWT	0.718 3	**0.731 2**	0.736 1	**0.744 6**
Eclipse_Platform_UI	0.639 9	**0.640 1**	0.649 7	**0.649 9**
GNOME_ekiga_general	0.819 7	**0.849 2**	0.838 8	**0.857 7**
GNOME_Evolution_Calendar	0.759 3	**0.773 2**	0.768 2	**0.778 9**
GNOME_Evolution_Contacts	0.723 9	**0.727 4**	0.721 7	**0.725 0**
GNOME_Evolution_Mailer	0.722 2	**0.733 6**	0.725 3	**0.736 1**
GNOME_Evolution_Shell	0.753 7	**0.755 9**	0.753 1	**0.755 1**
GNOME_gnome-panel_Panel	0.794 9	**0.797 5**	0.793 9	**0.795 9**
GNOME_gnome-terminal_general	0.923 7	**0.926 4**	0.918 6	**0.920 9**
Moizlla_Core_Printing	0.730 1	**0.760 5**	0.761 9	**0.781 1**
Moizlla_Core_XPCOM	0.752 8	**0.755 7**	0.756 7	**0.758 6**
Moizlla_Core_XPConnect	0.706 1	**0.731 8**	0.732 2	**0.750 0**
Moizlla_Core_XUL	0.704 3	**0.719 9**	0.711 8	**0.721 4**
Moizlla_Thunderbird_General	0.698 8	**0.711 1**	0.709 5	**0.717 7**
Mozilla_Core_Layout	0.610 3	**0.626 9**	0.627 0	**0.640 5**
Mozilla_Core_Security_UI	0.778 5	**0.786 9**	0.784 1	**0.788 9**

© 中国科学院软件研究所 http://www.jos.org.cn

从表 11 可以看出,通过基于模糊度的半监督方法来扩充数据集,并且用新的数据集再次训练分类器,所得到的分类器的分类效果在每个数据集上都有了提升.这证明了我们提出来的这个方法对于改善分类器的泛化能力是有效的.

本节的第 3 个实验是验证基于模糊度和 RBM 的样本迁移方法的有效性.在这个实验里,我们采用的分类器同样是代价敏感的 ELM 模型.我们对原始数据集的划分方法和第 1 个实验里数据集划分方法一样.对于被迁移的数据集,我们没有采用任何的划分方法.在进行分类之前,我们首先对原始数据集的整体进行 RBM 编码,同时对被迁移的数据集的整体进行 RBM 编码.在这两次编码中,我们采用 3 个级联的 RBM 对输入数据进行维度空间的变换.在编码过程中,我们只需要保证两次编码的输出维度一样即可,中间隐层的维度则可以自主调节.分类实验的具体操作过程是,先在原始的训练集上训练得到一个分类器,然后对被迁移的样本进行分类,用分类结果计算每个样本的模糊度大小,根据模糊度大小来选择样本来扩充原始的训练集.之后,再利用扩充后的训练集重新训练一个分类器.在实验过程中,前后两个分类器的参数设置完全一样.为了体现实验的客观性,我们重复本实验的全部过程(编码过程和分类过程)100 次,最后取这 100 次的平均值作为实验结果.我们用 Before 表示在原始数据集上训练得到的分类器的测试结果,用 After 表示在被扩充后的数据集上训练得到的分类器的测试结果.这些结果展示在表 12 中.

Table 12 Experimental results of transferring approach

表 12 迁移方法的实验结果

数据集名称	准确度		wflscore	
	Before	After	Before	After
Eclipse_CDT_cdt-core	0.728 8	**0.737 3**	0.718 6	**0.722 8**
Eclipse_JDT_Core	0.617 2	**0.618 7**	0.628 3	**0.629 5**
Eclipse_JDT_Debug	0.614 6	**0.616 5**	0.622 3	**0.624 4**
Eclipse_PDE_UI	0.650 8	**0.653 0**	0.657 3	**0.659 5**
Eclipse_Platform_Debug	0.590 9	**0.594 3**	0.594 0	**0.597 5**
Eclipse_Platform_SWT	0.776 3	**0.779 2**	0.755 8	**0.756 8**
Eclipse_Platform_UI	0.590 0	**0.591 1**	0.603 5	**0.604 0**
GNOME_ekiga_general	0.908 1	**0.911 7**	0.904 1	**0.909 2**
GNOME_Evolution_Calendar	0.803 8	**0.804 3**	0.805 7	**0.806 4**
GNOME_Evolution_Contacts	0.728 9	**0.729 5**	0.732 4	**0.733 3**
GNOME_Evolution_Mailer	0.675 7	**0.676 0**	0.693 6	**0.693 8**
GNOME_Evolution_Shell	0.755 8	**0.756 8**	0.756 2	**0.757 5**
GNOME_gnome-panel_Panel	0.779 4	**0.780 6**	0.784 4	**0.785 7**
GNOME_gnome-terminal_general	0.926 4	**0.927 4**	0.919 9	**0.920 2**
Moizlla_Core_Printing	0.847 7	**0.852 6**	0.821 9	**0.823 3**
Moizlla_Core_XPCOM	0.782 2	**0.788 6**	0.767 7	**0.769 8**
Moizlla_Core_XPConnect	0.776 9	**0.783 6**	0.773 4	**0.775 9**
Moizlla_Core_XUL	0.742 9	**0.746 7**	0.740 4	**0.742 8**
Moizlla_Thunderbird_General	0.755 6	**0.760 9**	0.729 8	**0.730 6**
Mozilla_Core_Layout	0.660 9	**0.665 8**	0.662 6	**0.664 2**
Mozilla_Core_Security_UI	0.823 5	**0.833 4**	0.800 3	**0.803 3**

从表 12 中我们可以看出,无论是准确率还是加权的 *F*-measure 值,再次训练的分类器都取得了较好的效果.这个结果显示,基于 RBM 和模糊度的样本迁移方法在扩充样本容量并且提升分类器的泛化能力这些方面是有效的.

本节的实验基本验证了我们提出方法的可行性和有效性.其中,如第 3 节所述,在利用半监督方法扩充训练集时,理论上可以进行多次迭代,逐渐扩充训练集的样本容量.在之前的实验里,我们设置的迭代次数为两次,即对分类器进行再训练的优化策略.再训练的结果已经表明,我们的方法是有效的.为了进一步验证我们的方法在数据集样本量继续扩充的条件下仍旧是有效的,我们将迭代次数增加,图 4 显示了随着迭代次数增加,我们的方法在每个数据集上的表现.

在上面的实验中,为了保证客观性,我们将实验重复了 100 次,然后取平均值.在每一次的实验中,我们提前将数据集按照 3:2:1 的比例分别分成训练集、测试集和验证集.其中，验证集是用来扩充训练集样本容量的数据集合.然后,我们利用半监督方法逐步地对训练数据集扩充了 5 次,每一次扩充后都在测试集上进行测试,并给

© 中国科学院软件研究所 http://www.jos.org.cn

出上图所示的测试结果.值得注意的是,为了克服随机赋权的影响,5 次迭代中,分类器的随机参数都是相同的值,实验结果就只会受到数据集样本容量变化的影响.从图 4 可以看出,在利用本文的半监督方法的前提下,随着数据集样本容量的增加,分类器的性能也在逐步提升,这一现象在 21 个数据集上都有体现.通过这组实验,验证了我们提出来的方法的有效性,即在人工标记的数据不充足的条件下,可以利用有效的半监督机制[33,34]来逐步扩充训练样本集,最终可以达到提升分类性能的目的.

Fig.4 Results of five tests on 21 datasets

图 4 21 个数据集 5 次测试的结果

本节实验里,网络结构参数、代价矩阵的设置以及其他参数的具体情况,都详细地列在了本文的附录里.

4 总 结

本文主要解决软件测试 Bug 报告中的样例分类问题.在实际工作中,我们发现 Bug 报告数据集普遍存在着 3 个问题:第 1 个是样本类别不平衡问题,第 2 个是数据集里被标注的样本个数不充足的问题,第 3 个是数据集的样本总容量不充足的问题.这 3 个问题都会对分类工作产生不良的影响.为了解决这 3 个问题,我们提出了 3 种方法:第 1 种是基于代价敏感的 ELM 模型来增加少数类样例被错误分类的代价,从而解决类别不平衡问题带来的影响;第 2 种是基于模糊度的半监督方法来自动标注没有标签的数据,从而扩充有标签的数据集;第 3 种是基于 RBM 和模糊度的样本迁移方法来扩增原有数据集的方法.本文通过几个实验验证了我们提出的方法是有效的、可行的.最后,可以注意到,本文提出来的方法具有很好的扩展性,可以与不同的模型[35−38]相结合.本文未来的研究工作将与集成学习[39,40]的方法相结合,以进一步提升目前的方法在 Bug 报告分类问题上的效果,同时将结合后的方法应用于更多的相关领域[41−43].

© 中国科学院软件研究所 http://www.jos.org.cn

References:

[1] Xia X, Lo D, Wang XY, Zhou B. Accurate developer recommendation for bug resolution. In: Proc. of the WCRE Congress. Koblenz: IEEE, 2013. 72−81. [doi: 10.1109/WCRE.2013.6671282]

[2] Guo SK, Chen R, Li H. Using knowledge transfer and roughset to predict the severity of Android test reports via text mining. Symmetry, 2017,9(8):161−179. [doi: 10.3390/sym9080161]

[3] Antoniol G, Ayari K, Penta MD, Khomh F, Gueheneuc YG. Is it a bug or an enhancement? A text based approach to classify change requests. In: Proc. of the CASCON Congress. 2008. 304−318. [doi: 10.1145/1463788.1463819]

[4] Menzies T, Marcus A. Automated severity assessment of software defect reports. In: Proc. of the ICSM Congress. 2008. 346−355. [doi: 10.1109/ICSM.2008.4658083]

[5] Tian Y, Lo D, Xia X, Sun CN. Automated prediction of bug report priority using multi-factor analysis. Empirical Software Engineering, 2015,20(5):1354−1383.

[6] Feng Y, Chen ZY, Jones JA, Fang CR, Xu BW. Test report prioritization to assist crowdsourced testing. In: Proc. of the FSE Congress. New York: ACM Press, 2015. 225−236.

[7] Runeson P, Alexandersson M, Nyolm O. Detection of duplicate defect reports using natural language processing. In: Proc. of the ICSE Congress. Minneapolis: IEEE, 2007. 499−510.

[8] Sun C, Lo D, Khoo SC. Towards more accurate retrieval of duplicare bug reports. In: Proc. of the ASE Congress. Lawrence: IEEE, 2011. 253−262.

[9] Lamkanfi A, Demeyer S, Giger E, Goethals B. Predicting the severity of a reported bug. In: Proc. of the MSR Congress. 2010. 1−10.

[10] Yang XL, Lo D, Xia X, Huang Q, Sun JL. High-impact bug report identification with imbalanced learning strategies. Journal of Computer Science and Technology, 2017,32(1):181−198. [doi: 10.1007/s11390-017-1713-3]

[11] Huang GB, Zhu QY, Siew CK. Extreme learning machine: Theory and applications. Neuro-Computing, 2006,70(1):489−501.

[12] Huang GB, Zhou H, Ding X. Extreme learning machine for regression and multicalss classification. IEEE Trans. on Systems, Man, and Cybernetics. Part B, Cybernetics, 2012,42(2):513−529.

[13] Wang XZ, Xing HJ, Li Y, *et al*. A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning. IEEE Trans. on Fuzzy Systems, 2015,23(5):1638−1654.

[14] Pan JL, Tsang IW, Kwok JT, Yang Q. Domain adaptation via transfer component analysis. IEEE Trans. on Neural Networks, 2011, 22(2):199−210.

[15] Hinton GE. A practical guide to training restricted Boltzmann machines. Momentum, 2010,9(1):926−947.

[16] Dai WY, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. In: Proc. of the ICML Congess. 2007. 193−200.

[17] Tang JX, Deng CW, Huang GB. Extrem learning machine for multilayer perceptron. IEEE Trans. on Neural Network and Learning Systems, 2016,27(4):809−821.

[18] Han K, Yu D. Speech emotion recognition using deep neural network and extreme learning machine. In: Proc. of the INTERSPEECH. 2014. 223−227.

[19] Zeng YJ, Xu X, Fang YQ, Zhao K. Traffic Sign Recognition Using Extreme Learning Classifier with Deep Convolutional Features. Springer Int'l Publishing, 2015. 223−228.

[20] Kim Y, Jernite Y, Sontag D, Rush AM. Character-aware neural language models. In: Proc. of the AAAI Congess. 2016. 2741−2749.

[21] Hinton GE. A fast learning algorithm for deep belief nets. Neural Computation, 2006,18(7):1527−1554.

[22] Wang XZ, Zhang TL, Wang R. Noniterative deep learning: Incorporating restricted Boltzmann machine into multilayer random weight neural networks. IEEE Trans. on Systems, Man, and Cybernetics: Systems, 2017, Early Access. [doi: 10.1109/TSMC.2017.2701419]

[23] Huang GB. An insight into extreme learning machines: Random neurons, random features and kernels. Cognitive Computation, 2014,6(3):376−390.

[24] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.

[25] Hinton GE. Training products of experts by minimizing contrastive divergence. Neural Computation, 2002,14(8):1771−1800.

[26] Zhu HY, Wang XZ. A cost-sensitive semi-supervised learning model based on uncertainty. Neurocomputing, 2017,251:106−114.

[27] Zadeh LA. Fuzzy sets. Information & Control, 1965,8(3):338−353.

[28] Eclipse. 2018. http://bugs.eclipse.org/bugs

[29] Mozilla. 2018. http://bugzilla.mozilla.org

© 中国科学院软件研究所 http://www.jos.org.cn

[30] GNOME. 2018. http://bugzilla.gnome.org

[31] Zhou Y, Tong Y, Gu R, Gall H. Combing text mining and data mining for bug report classification. In: Proc. of the ICSME Congess. 2014. 311−320.

[32] Shi XF, Xu GQ, Shen FR, Zhao JX. Solving the data imbalance problem of P300 detection via random under-sampling bagging SVMs. In: Proc. of the IJCNN Congress. 2015. 12−17.

[33] Shen FR, Yu H, Sakurai K, Hasegawa O. An incremental online semi-supervised active learning algorithm based on a self-organizing incremental neural network. Neural Computing & Applications, 2011,20(7):1061−1074.

[34] Huang GB, Song S, Gupta JND. Semi-supervised and unsupervised extreme learning machines. IEEE Trans on Cybern, 2014, 44(12):2405−2417.

[35] Qiu TY, Shen FR, Zhao JX. Review of self-organizing incremental neural network. Ruan Jian Xue Bao/Journal of Software, 2016, 27(9):2230−2247 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5068.htm [doi: 10.13328/j.cnki.jos.005068]

[36] Huang GB, Li MB, Chen L. Incremental extreme learning machine with fully complex hidden nodes. Neurocomputing, 2008, 71(4-6):576−583.

[37] Huang GB, Chen L, Siew CK. Universal approximation using incremental constructive feedforward with random hidden nodes. IEEE Trans. on Neural Network, 2006,17(4):879−892.

[38] Huang GB, Chen L. Enhanced random search based incremental extreme learning machine. Neurocomputing, 2008,71(16-18): 3460−3468.

[39] Zhai JH. Fusion of extreme learning machine with fuzzy integal. Fuzziness and Knowledge-Based Systems, 2013,21:23−24.

[40] Zhai JH, Xu HY, Wang XZ. Dynamic ensemble extreme learning machine based on sample entropy. Soft Computing, 2012,16(9): 1493−1502.

[41] Wang LJ, Li M, Cai SB, Li G, Xie B, Yang FQ. Internet information search based approach to enriching textual descriptions for public Web services. Ruan Jian Xue Bao/Journal of Software, 2012,23(6):1335−1349 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4088.htm [doi: 10.3724/SP.J.1001.2012.04088]

[42] Mou LL, Li G, Zhang L, Wang T, Jin Z. Convolutional neural networks over tree structures for programming language. In: Proc. of the AAAI Congress. 2016. 1287−1293.

[43] Meng Z, Mou LL, Li G, Jin Z. Context-aware tree-based convolutional neural networks for natural language inference. In: Proc. of the KSEM Congress. 2016. 515−526.

附中文参考文献:

[35] 邱天宇,申富饶,赵金熙.自组织增量学习神经网络综述.软件学报,2016,27(9):2230−2247. http://www.jos.org.cn/1000-9825/5068.htm [doi: 10.13328/j.cnki.jos.005068]

[41] 王立杰,李萌,蔡斯博,李戈,谢冰,杨芙清.基于网络信息搜索的 Web Service 文本描述信息扩充方法.软件学报,2012,23(6): 1335−1349. http://www.jos.org.cn/1000-9825/4088.htm [doi: 10.3724/SP.J.1001.2012.04088]

附录

证明:用 softmax 的输出来表示后验概率的合理性.

定义 $x(1),\ldots,x(m)$为 m 个输入向量,且 $x(i)\in\mathbb{R}^{1\times ncc}$,$ncc$ 为输入向量的维度,$x(i)_j$ 是向量 $x(i)$第 j 维属性值,$i=1,2,\ldots,m$,$j=1,2,\ldots,ncc$.对 ELM 的输出层来讲,其输入向量$x(i)$是隐层的输出向量,因此在 ELM 中,ncc 为隐层节点数.为了简化表达,下文出现的 x 表示 $x(i),\forall i=1,2,\ldots,m$,$y(1),\ldots,y(m)$为输入向量对应的真实类别标签,其中,$y(i)\in\{1,2,\ldots,ns\}$,$i=1,2,\ldots,m$.$ns$ 为类别个数,在 ELM 中,ns 表示输出节点数.

定义一个指示函数 $A(u,v)=\begin{cases}1, & u=v\\ 0, & \text{其他}\end{cases}$,其中,

$$u,v\in\{1,2,\ldots,ns\} \tag{A.1}$$

定义一个概率函数$\pi()$:$\mathbb{R}^{1\times ncc}\to[0,1]^{ns}$,该函数将输入向量 x 映射为一个 ns 维向量,其中,第 u 维$\pi(x)_u$代表输入向量属于第 u 类的概率大小.基于以上定义,下面证明$\pi()$具有 softmax 函数形式.

根据$\pi()$的定义,可知$\pi()$应满足以下性质:

$$\pi(x)_u\geqslant 0 \tag{A.2}$$

© 中国科学院软件研究所 http://www.jos.org.cn

$$\sum_{u=1}^{ns}\pi(x)_u=1 \tag{A.3}$$

$$\sum_{i=1}^{m}\pi(x(i))_u x(i)_j=\sum_{i=1}^{m}A(u,y(i))x(i)_j \tag{A.4}$$

根据信息论,$\pi()$的熵可以定义为

$$-\sum_{v=1}^{ns}\sum_{i=1}^{m}\pi(x(i))_v\log(\pi(x(i))_v) \tag{A.5}$$

根据信息论最大熵原理,即通常选取熵值最大的概率分布作为随机变量的分布,所以在公式(A.2)~公式(A.4)的约束下,最大化公式(A.5),该过程是一个求条件极值的问题,通常使用拉格朗日乘数法求解该问题.这里,拉格朗日方程可以写成:

$$L=\sum_{j=1}^{ncc}\sum_{v=1}^{ns}\beta_{j,v}\left(\sum_{i=1}^{m}\pi(x(i))_v x(i)_j-A(v,y(i))x(i)_j\right)+\sum_{v=1}^{ns}\sum_{i=1}^{m}\lambda_i(\pi(x(i))_v-1)-\sum_{v=1}^{ns}\sum_{i=1}^{m}\pi(x(i))_v\log(\pi(x(i))_v) \tag{A.6}$$

其中,$\beta_{j,v}$与λ_i为拉格朗日乘数,且$\beta_{j,v}$为矩阵$\beta\in\mathbb{R}^{ncc\times ns}$中的元素.注意到条件(A.2)为不等式,所以无法被包含进方程(A.6),但是可以在方程(A.6)的解中,选择满足条件(A.2)的值为最终解.在求解过程中,L 对$\pi(x(i))_v$求偏导,且极值在偏导数等于 0 处取得,即

$$\frac{\partial L}{\partial\pi(x(i))_v}=0=x(i)\beta_v+\lambda_i-\log(\pi(x(i))_v)-1 \tag{A.7}$$

其中,$\beta_v\in\mathbb{R}^{ncc\times 1}$为$\beta$中第 v 列列向量.对公式(A.7)等价变形有:

$$\pi(x(i))_v=\mathrm{e}^{x(i)\cdot\beta_v+\lambda_i-1} \tag{A.8}$$

根据公式(A.3),有:

$$\sum_{u=1}^{ns}\mathrm{e}^{x(i)\cdot\beta_u+\lambda_i-1}=1 \tag{A.9}$$

即

$$\mathrm{e}^{\lambda_i}=\frac{1}{\sum_{u=1}^{ns}\mathrm{e}^{x(i)\cdot\beta_u-1}} \tag{A.10}$$

根据公式(A.8),有 $\mathrm{e}^{\lambda_i}=\frac{\pi(x(i))_v}{\mathrm{e}^{x(i)\cdot\beta_v-1}}$,代入公式(A.10),有:

$$\pi(x(i))_v=\frac{\mathrm{e}^{x(i)\cdot\beta_v-1}}{\sum_{u=1}^{ns}\mathrm{e}^{x(i)\cdot\beta_u-1}} \tag{A.11}$$

公式(A.11)右侧分式上下同乘 e,有:

$$\pi(x(i))_v=\frac{\mathrm{e}^{x(i)\cdot\beta_v}}{\sum_{u=1}^{ns}\mathrm{e}^{x(i)\cdot\beta_u}} \tag{A.12}$$

公式(A.12)可以简化成:

$$\pi(x)_v=\frac{\mathrm{e}^{x\cdot\beta_v}}{\sum_{u=1}^{ns}\mathrm{e}^{x\cdot\beta_u}} \tag{A.13}$$

在 ELM 中,β_v可被看作是连接隐层与输出层第 v 个节点的权重向量,β是输出层权重,则 $x(i)\beta_v$是网络输出 y_{iv};同理,$x(i)\beta_u=y_{iu}$.综上,证明满足概率性质的概率函数$\pi()$具有 softmax 的函数形式,从而证明用 softmax 函数作用在 ELM 的输出层输出后验概率的合理性. □

实验 1 的参数设置.

© 中国科学院软件研究所 http://www.jos.org.cn

数据集	代价矩阵	隐层节点数
Eclipse_CDT_cdt-core	[0,5;2,0]	500
Eclipse_JDT_Core	[0,3;1,0]	1 000
Eclipse_JDT_Debug	[0,3,1,0]	500
Eclipse_PDE_UI	[0,3;1,0]	500
Eclipse_Platform_Debug	[0,5;2,0]	200
Eclipse_Platform_SWT	[0,5;2,0]	1 500
Eclipse_Platform_UI	[0,3;1,0]	2 000
GNOME_ekiga_general	[0,5,2,0]	500
GNOME_Evolution_Calendar	[0,5;2,0]	1 000
GNOME_Evolution_Contacts	[0,5;2,0]	1 000
GNOME_Evolution_Mailer	[0,2;1,0]	3 000
GNOME_Evolution_Shell	[0,3;1,0]	500
GNOME_gnome-panel_Panel	[0,5;2,0]	500
GNOME_gnome-terminal_general	[0,5;2,0]	1 500
Moizlla_Core_Printing	[0,5;1,0]	500
Moizlla_Core_XPCOM	[0,5;2,0]	500
Moizlla_Core_XPConnect	[0,2;1,0]	100
Moizlla_Core_XUL	[0,5;2,0]	1 000
Moizlla_Thunderbird_General	[0,5;2,0]	1 000
Mozilla_Core_Layout	[0,5;2,0]	1 500
Mozilla_Core_Security_UI	[0,5;2,0]	400

实验 2 的参数设置.

数据集	代价矩阵	隐层节点数	*k* 值
Eclipse_CDT_cdt-core	[0,5;2,0]	200	50
Eclipse_JDT_Core	[0,6;1,0]	200	50
Eclipse_JDT_Debug	[0,3;1,0]	200	10
Eclipse_PDE_UI	[0,3;1,0]	100	10
Eclipse_Platform_Debug	[0,5;2,0]	200	10
Eclipse_Platform_SWT	[0,3;2,0]	1 000	50
Eclipse_Platform_UI	[0,5;2,0]	500	5
GNOME_ekiga_general	[0,5;2,0]	500	50
GNOME_Evolution_Calendar	[0,3;1,0]	1 000	50
GNOME_Evolution_Contacts	[0,3;1,0]	500	20
GNOME_Evolution_Mailer	[0,5;2,0]	2 500	200
GNOME_Evolution_Shell	[0,3;1,0]	200	50
GNOME_gnome-panel_Panel	[0,3;1,0]	500	50
GNOME_gnome-terminal_general	[0,3;1,0]	500	50
Moizlla_Core_Printing	[0,3;1,0]	500	50
Moizlla_Core_XPCOM	[0,3;1,0]	200	10
Moizlla_Core_XPConnect	[0,2;1,0]	100	5
Moizlla_Core_XUL	[0,5;2,0]	200	50
Moizlla_Thunderbird_General	[0,5;2,0]	500	50
Mozilla_Core_Layout	[0,5;2,0]	1 500	50
Mozilla_Core_Security_UI	[0,5;2,0]	200	50

实验 3 的参数设置.

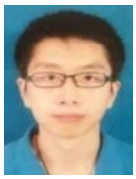
数据集	被迁移数据集	代价矩阵	隐层节点数	*k* 值
Eclipse_CDT_cdt-core	Eclipse_JDT_Core	[0,5;2,0]	20	10
Eclipse_JDT_Core	Eclipse_JDT_Debug	[0,5;2,0]	20	10
Eclipse_JDT_Debug	Eclipse_Platform_Debug	[0,5;2,0]	200	10
Eclipse_PDE_UI	Eclipse_Platform_Debug	[0,5;2,0]	50	10
Eclipse_Platform_Debug	Eclipse_Platform_SWT	[0,5;2,0]	200	10
Eclipse_Platform_SWT	Eclipse_CDT_cdt-core	[0,5;2,0]	100	10
Eclipse_Platform_UI	Eclipse_PDE_UI	[0,5;2,0]	20	10
GNOME_ekiga_general	GNOME_Evolution_Calendar	[0,5;2,0]	100	50
GNOME_Evolution_Calendar	GNOME_Evolution_Shell	[0,5;2,0]	200	20
GNOME_Evolution_Contacts	GNOME_Evolution_Shell	[0,5;2,0]	200	10
GNOME_Evolution_Mailer	GNOME_gnome-terminal_general	[0,3;1,0]	20	10
GNOME_Evolution_Shell	Moizlla_Core_Printing	[0,5;2,0]	200	10
GNOME_gnome-panel_Panel	GNOME_ekiga_general	[0,3;1,0]	50	10
GNOME_gnome-terminal_general	GNOME_gnome-panel_Panel	[0,5;2,0]	100	50
Moizlla_Core_Printing	Moizlla_Core_XPCOM	[0,5;2,0]	20	10
Moizlla_Core_XPCOM	Moizlla_Core_Printing	[0,5;2,0]	20	10
Moizlla_Core_XPConnect	Moizlla_Core_XUL	[0,5;2,0]	20	10
Moizlla_Core_XUL	Moizlla_Core_XPConnect	[0,5;2,0]	20	10
Moizlla_Thunderbird_General	Moizlla_Core_XUL	[0,5;2,0]	30	10
Mozilla_Core_Layout	Moizlla_Core_XPCOM	[0,5;2,0]	20	10
Mozilla_Core_Security_UI	Moizlla_Core_XPCOM	[0,5;2,0]	20	10

© 中国科学院软件研究所 http://www.jos.org.cn

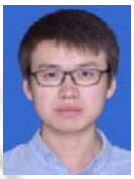
序号	RBM 节点数(原数据集)	RBM 节点数(被迁移数据集)
1	100-500-100	100-200-100
2	100-200-100	100-200-100
3	100-500-100	100-500-100
4	100-500-200	100-1000-200
5	100-500-100	100-500-100
6	500-500-500	500-500-500
7	100-500-200	100-200-200
8	100-500-200	100-200-200
9	500-500-100	500-500-100
10	500-500-100	500-500-100
11	300-300-200	300-300-200
12	100-200-100	100-200-100
13	100-200-100	500-500-100
14	100-200-100	100-200-100
15	100-500-200	100-200-200
16	100-500-200	100-200-200
17	100-500-200	100-200-200
18	100-500-200	100-200-200
19	100-500-200	100-200-200
20	100-500-200	100-200-200
21	100-500-200	100-200-200

实验 4 的参数设置.

数据集	代价矩阵	隐层节点数	k 值
Eclipse_CDT_cdt-core	[0,5;2,0]	200	10
Eclipse_JDT_Core	[0,6;1,0]	200	10
Eclipse_JDT_Debug	[0,3;1,0]	200	20
Eclipse_PDE_UI	[0,3;1,0]	100	10
Eclipse_Platform_Debug	[0,5;2,0]	200	10
Eclipse_Platform_SWT	[0,3;2,0]	1 000	50
Eclipse_Platform_UI	[0,5;2,0]	300	20
GNOME_ekiga_general	[0,5;2,0]	500	20
GNOME_Evolution_Calendar	[0,3;1,0]	1 000	20
GNOME_Evolution_Contacts	[0,3;1,0]	500	20
GNOME_Evolution_Mailer	[0,5;2,0]	2 500	100
GNOME_Evolution_Shell	[0,5;2,0]	200	10
GNOME_gnome-panel_Panel	[0,3;1,0]	500	50
GNOME_gnome-terminal_general	[0,5;2,0]	500	10
Moizlla_Core_Printing	[0,3;1,0]	500	10
Moizlla_Core_XPCOM	[0,2;1,0]	200	5
Moizlla_Core_XPConnect	[0,2;1,0]	200	5
Moizlla_Core_XUL	[0,5;2,0]	200	20
Moizlla_Thunderbird_General	[0,5;2,0]	500	20
Mozilla_Core_Layout	[0,5;2,0]	1 500	50
Mozilla_Core_Security_UI	[0,5;2,0]	200	20



张天伦(1991－),男,河北保定人,博士生,主要研究领域为机器学习,软件工程,计算视觉.



杨溪(1993－),男,硕士生,主要研究领域为机器学习,计算视觉,模糊集.

陈荣(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,软件故障诊断,行为识别,运筹学.

祝宏玉(1994－),女,硕士,主要研究领域为神经网络,不平衡数据处理,大数据算法.

© 中国科学院软件研究所 http://www.jos.org.cn