

基于迁移学习的跨项目软件缺陷预测

张洋洋 荆晓远 吴 飞

(南京邮电大学 自动化学院 江苏 南京 210003)

摘 要: 软件缺陷预测在提高软件质量、控制与平衡软件成本方面起着举足轻重的作用,是软件工程的活跃领域。研究者们提出了许多预测技术,从不同层面解决了不同的问题。传统软件缺陷预测算法在面对跨项目软件缺陷预测中往往不能得到一个理想的结果,原因是训练数据样本(源数据)和测试数据样本(目标数据)之间的分布是不同的。为了解决这个问题,提出了一种基于迁移学习的跨项目软件缺陷预测算法。该算法首先采用了一种不同分布之间的距离度量方式,训练出一种模型来最小化训练数据和测试数据之间的分布差异以及条件分布差异,在映射过后的新的特征空间中两种数据集几乎拥有同样的分布。然后就可以采用传统的机器学习算法进行分类。实验结果表明,该算法具有较好的预测性能。

关键词: 软件缺陷预测; 迁移学习; 特征映射; 机器学习

中图分类号: TP181

文献标识码: A

文章编号: 1673-629X(2018)12-0083-03

doi: 10.3969/j.issn.1673-629X.2018.12.018

Cross-project Software Defect Prediction Based on Transfer Learning

ZHANG Yang-yang, JING Xiao-yuan, WU Fei

(School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Software defect prediction plays an important role in improving software quality, controlling and balancing software costs as an active area of software engineering. Researchers have proposed many prediction techniques to solve different problems at different levels. The traditional software defect prediction algorithm often cannot get an ideal result in the face of cross-project software defect prediction because the distribution between the training data sample (source data) and the test data sample (target data) is different. In order to solve this problem, we present a cross-project software defect prediction algorithm based on transfer learning. Firstly, the algorithm uses a distance measurement between different distributions to train a model to minimize the distribution difference and conditional distribution difference between training data and test data. After mapping the two data sets have almost the same distribution in the new feature space. Then the traditional machine learning algorithm can be used for classification. The experiment shows that the proposed algorithm has better predictive performance.

Key words: software defect prediction; transfer learning; feature map; machine learning

0 引言

跨公司软件缺陷预测问题不同于传统的机器学习问题,它的训练数据和测试数据属于不同的分布。为了解决这个问题,目前有许多不同的方法,Turhan等^[1]使用一种最近邻滤波器从源数据中选择与测试数据相似的数据作为训练数据。Zimmermann等^[2]使用决策树帮助项目管理者进行跨工程预测前对精确度、召回率和准确度进行估计。

1 相关工作

1.1 迁移学习技术

Jiang等^[3]认为,与样本类别高度相关的那些特征应该在训练得到的模型中被赋予更高的权重,因而提出了一种两阶段的特征选择框架。Dai等^[4]提出了一种基于联合聚类的预测领域外文档的分类方法CoCC,该方法通过对类别和特征进行同步聚类,实现知识与类别的迁移。还有一种方法^[5]通过最小化源数

收稿日期: 2018-01-10

修回日期: 2018-05-16

网络出版时间: 2018-07-04

基金项目: 国家自然科学基金(61702280); 江苏省自然科学基金(BK20170900); 江苏省高等学校自然研究项目(17KJB520025); 南京邮电大学引进人才科研启动基金(NY217009)

作者简介: 张洋洋(1992-),男,硕士研究生,研究方向为模式识别、软件缺陷预测等; 荆晓远,教授,研究方向为机器学习、软件缺陷预测、深度学习等。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20180703.1510.018.html>

据样本和目标数据样本在隐性语义的差距,从而求解降维后的特征空间,在该隐性空间上,不同的领域具有有限相同或者非常接近的均匀分布,因此就可以直接利用监督学习算法训练模型对目标领域数据进行预测。Jing 等^[6]提出一种基于字典学习的软件缺陷预测方法,能够高效地预测项目内缺陷分布。

1.2 软件缺陷预测

软件缺陷预测主要根据软件的基本属性(如程序代码行数、体积、软件复杂度等各种信息数据)和已经发现的缺陷等数据借助于机器学习、统计学习等技术进行挖掘和分析,进而预测出软件系统中可能遗留而且尚未被发现的情况^[7-9]。

与普通软件缺陷预测算法不同,跨项目软件缺陷预测中的训练数据和测试数据属于不同的分布。为了解决这个问题,Turhan 等提出了一种最近邻滤波器算法从源数据中选择相似的样本用于训练。Zimmermann 等考察了使用决策树时不同特征对缺陷预测质量的影响。Liu 等^[10]提出在多个数据库上基于搜索方法进行软件缺陷预测。与基于非搜索的模型比较,他们的方法具有较低的错误分类代价。Ma 等^[11]基于目标数据样本的特征信息对源数据特征设置权重,提出了一种基于朴素贝叶斯的跨公司软件缺陷预测模型。

2 基于迁移学习的软件缺陷预测

文中提出一种联合适配分布算法,通过一个特征映射 T 使特征 x 和标签 y 的期望在训练数据和测试数据之间相匹配:

$$\begin{aligned} \min_T \| E_{P(x,y)} [T(x) | y] - E_{P(t,y)} [T(x) | y] \|^2 \approx \\ \| E_{P(x,y)} [T(x) | y] - E_{P(t,y)} [T(t) | y] \|^2 + \\ \| E_{Q(x,y|x)} [y | T(x)] - E_{Q(t,y|t)} [y | T(t)] \|^2 \end{aligned} \quad (1)$$

因为没有带有标签的测试,所以条件概率 $Q(y | t)$ 不能明确地计算出来。最佳的估计就是假设 $Q(y | t) \approx Q_x(y | t)$,这样就可以通过使用一个分类器在有标签的训练集上训练出分类器 f 作用在无标签的测试集上。

通过最小化输入数据的重构误差,降维方法可以学习到一个迁移后的特征表示。为简便起见,选择主成分分析法(PCA)做数据重构。使用 $X = \{x_1, x_2, \dots, x_n\} \in \mathbf{R}^{n \times d}$ 作为输入数据的矩阵, $H = I - \frac{1}{a} \mathbf{1} \mathbf{1}^T$ 代表中心矩阵, $a = n + m$ 且 $\mathbf{1}$ 表示大小为 $a \times a$ 的全 1 矩阵,然后均方差矩阵为 XX^T 。PCA 学习的目标就是找到一个几何映射矩阵 $A \in \mathbf{R}^{d \times k}$ 以最大化以下问题:

$$\max_{A^T A = I} \text{tr}(A^T X H X^T A) \quad (2)$$

其中, $\text{tr}(\cdot)$ 表示矩阵的迹。并且这个优化问题可

以很好地使用特征分解: $XX^T A = A \Phi$, $\Phi = \text{diag}(\varphi_1, \varphi_2, \dots, \varphi_k) \in \mathbf{R}^{k \times k}$ 是前 k 个最大的特征值。然后得到最优的 k 维的特征表示: $Z = [z_1, z_2, \dots, z_k] = A^T X$ 。

然而,通过 PCA 方法将数据降维后,训练数据和测试数据之间分布的差异依然很大。因此主要的问题是通过一定的度量方式来降低两个分布 $P_x(x)$ 和 $P_t(t)$ 之间的距离。采用文献[12]提出的 Maximum Mean 作为距离度量方法,计算训练数据样本和测试数据样本之间的均值之差:

$$\left\| \frac{1}{n} \sum_{i=1}^n A^T x_i - \frac{1}{m} \sum_{j=n+1}^{n+m} A^T x_j \right\|^2 = \text{tr}(A^T X M_0 X^T A) \quad (3)$$

其中, M_0 是 MMD 矩阵,可以通过式 4 计算:

$$(M_0)_{ij} = \begin{cases} \frac{1}{n \times n} & x_i, x_j \in L \\ \frac{1}{m \times m} & t_i, t_j \in T \\ \frac{1}{n \times m} & \text{otherwise} \end{cases} \quad (4)$$

通过最小化式 3 即是最大化式 2,然后训练数据和测试数据经过 $Z = A^T X$ 的转换后,两种分布之间的距离变近了。这个方法其实就是 TCA^[13]。

更进一步,参照文献[14],最小化条件分布 $Q_x(y | x)$ 和 $Q_t(y | t)$ 也是一个很重要的方面。遗憾的是,由于在测试数据集中没有标签,所以就无法对 $Q_t(y | t)$ 进行直接建模。由于先验概率 $Q_x(y | x)$ 和 $Q_t(y | t)$ 之间联系紧密,转而去探索用类条件概率分布 $Q_x(x | y)$ 和 $Q_t(t | y)$ 来代替。现在由于具备了真实的训练数据标签和伪测试数据标签,可以匹配类条件概率分布 $Q_x(x | y=c)$ 和 $Q_t(t | y=c)$,其中 c 属于类标签的集合。通过修改 MMD 来度量两种条件分布之间的距离:

$$\left\| \frac{1}{n^{(c)}} \sum_{x_i \in L^{(c)}} A^T x_i - \frac{1}{m^{(c)}} \sum_{t_j \in T^{(c)}} A^T t_j \right\|^2 = \text{tr}(A^T X M_c X^T A) \quad (5)$$

其中, $L^{(c)} = \{x_i : x_i \in L \wedge y(x_i) = c\}$ 表示在训练数据中属于类别 c 的样本点, $y(x_i)$ 是它对应的标签。

测试数据中使用的伪标签不是正确的,我们仍然可以借用它们来匹配条件概率分布。理论参考是使用充分统计量代替密度估计。

整合前面两个优化的问题,可以得到最终的目标函数:

$$\min_{A^T X H X^T A = I} \sum_{c=0}^C \text{tr}(A^T X M_c X^T A) + \lambda \|A\|_F^2 \quad (6)$$

其中, λ 是一个正则化参数。

对于非线性问题,考虑核映射 $\psi: x \rightarrow \psi(x)$ 与核矩阵 $K = \psi(x)^T \psi(x) \in \mathbf{R}^{n \times n}$ 。

上述优化问题转化为:

$$\min_{A^T K H K^T A = I} \sum_{c=0}^C \text{tr}(A^T K M_c K^T A) + \lambda \|A\|_F^2 \quad (7)$$

通过约束优化理论,使用 $\Phi = \text{diag}(\varphi_1, \varphi_2, \dots, \varphi_k) \in \mathbf{R}^{k \times k}$ 代表拉格朗日乘子,然后推导出问题 7 的拉格朗日函数为:

$$\text{Lag} = \text{tr}[A^T (X \sum_{c=0}^C M_c X^T + \lambda I) A] + \text{tr}[(I - A^T X H X^T A) \Phi] \quad (8)$$

设置 $\frac{\partial \text{Lag}}{\partial A} = 0$, 最终得到特征分解问题:

$$(X \sum_{c=0}^C M_c X^T + \lambda I) A = X H X^T A \Phi \quad (9)$$

最终,求解最优适配矩阵 A 的问题转化为求等式 9 的最小 k 个特征向量。

整个算法的学习过程总结如下:

(1) 通过式 4 构造 MMD 矩阵。

(2) 重复: 求解式 9 所描述的特征分解问题,然后选择前 k 个最小的特征向量来构造适配矩阵 A ,在新的特征表示上训练出一个标准的分类器 f 来更新伪测试数据标签。根据式 6 的构造 MMD 矩阵 M 。直到收敛。

(3) 返回在新的特征表示上训练得到的分类器 f 。

3 实验

将提出的 JDBFM 算法与文献 [11] 提出的 TNB 算法、文献 [1] 提出的 NN-filter 算法进行对比,依据软件缺陷预测常用的性能指标召回率 (recall)、精确度 (precision)、假阳率 (pf) 以及 F -measure 对算法进行评价。

文中采用召回率、精确度和 F -measure 值评估模型的预测效果。由于高的召回率往往要以低精确度为代价,反之亦然。因此,可以使用 F -measure 将召回率和精确度综合起来进行评价。 F -measure 为召回率和查准率的调和平均数,值越高性能越好,其计算公式如下:

$$F\text{-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \quad (10)$$

对于 NN-filter 算法,每个测试数据都要从训练数据中选择 k 个最近邻的样本构成训练数据集来训练软件缺陷预测模型,与文献 [11] 保持一致,文中也选取 $k=10$ 。

文中提出的算法 JDBFM 中有两个参数需要设置:子空间基数 k 和正则项系数 λ ($\lambda=2.0$)。在下列的数据集上的实验中证实了在一个很大的参数的范围内可以得到一个相对稳定的结果。实验结果如图 1 和图 2 所示。

Train→test	NN-filter	TNB	JDBFM
ZXing→Safe	0.466 6	0.526 3	0.456 5
ZXing→Apache	0.511 1	0.544 9	0.475 6
Safe→ZXing	0.328 5	0.297 2	0.526 3
Safe→Apache	0.477 1	0.557 5	0.652 2
Apache→ZXing	0.360 1	0.409 0	0.573 0
Apache→Safe	0.669 2	0.692 3	0.718 8
Average	0.468 7	0.504 5	0.600 1

图 1 ReLink 数据库实验结果

Train→test	NN-filter	TNB	JDBFM
JDT→EQ	0.666 7	0.651 3	0.651 5
LC→PDE	0.230 8	0.229 1	0.220 6
EQ→ML	0.178 8	0.277 0	0.367 8
JDT→LC	0.166 7	0.322 5	0.371 4
EQ→LC	0.139 4	0.291 3	0.271 6
PDE→JDT	0.390 9	0.384 0	0.403 0

图 2 AEEEM 数据库实验结果

通过以上实验结果可以看出,NN-filter 算法所获得的实验结果 F -measure 值较 TNB 和 JDBFM 算法都要低一些。而提出的 JDBFM 算法考虑了目标数据和源数据之间的分布和条件分布,效果要比以上两种算法好。实验结果也证明了 JDBFM 算法的优越性。

4 结束语

在缺陷预测模型中,训练数据和测试数据分别来源于不同的工程,训练数据是有标签的,而测试数据是没有标签的,并且源数据和测试数据来源于不同的分布。提出的算法 JDBFM 能够很大限度地利用跨项目数据信息。通过在现有两份数据库上的实验,证明了该算法能显著提高跨项目软件缺陷预测的性能。

参考文献:

- [1] TURHAN B, MENZIES T, BENER A B, et al. On the relative value of cross-company and within-company data for defect prediction[J]. Empirical Software Engineering, 2009, 14(5): 540-578.
- [2] ZIMMERMANN T, NAGAPPAN N, GALL H, et al. Cross-project defect prediction: a large scale experiment on data vs. Domain vs. Process[C]//Proceedings of the 7th joint meeting of the European software engineering conference and the ACM SIG-SOFT symposium on the foundations of software engineering. Amsterdam, The Netherlands: ACM, 2009: 91-100.
- [3] JIANG Jing, ZHAI Chengxiang. A two-stage approach to domain adaptation for statistical classifiers[C]//Proceedings of the 16th ACM conference on information and knowledge management. Lisbon, Portugal: ACM, 2007: 401-410.

(下转第 90 页)

波算法,不仅可以 根据图像噪声点的数量自适应调整滤波窗口大小,还可以根据滤波窗口像素点相似度值对像素点进行自动分组,并赋予每组所有像素点权重值。该算法对滤波窗口中心像素点以及周围相似度接近的像素点赋予较大的权重值,所以较好地保护了图像的细节,解决了噪声去除和图像细节保留之间的矛盾,是一种性能良好、效率高的图像处理技术。

参考文献:

- [1] 方 政,胡晓辉,陈 永.基于多方向中值滤波的各向异性扩散滤波算法[J].计算机工程与应用,2017,53(4):195-199.
- [2] 沈德海,侯 建,鄂 旭,等.基于米字型窗口中值加权的滤波算法[J].计算机技术与发展,2017,27(9):78-81.
- [3] 苏育挺,张天娇,张 静,等.基于局部二值模式的中值滤波检测算法[J].计算机应用研究,2016,33(1):258-261.
- [4] 雷 芸.基于中值预滤波的非局部平均去噪算法[J].微电子学与计算机,2015,32(5):138-142.
- [5] 黄 燕,雷 涛,樊养余,等.基于自适应窗口的裁剪中值滤波方法[J].计算机科学,2015,42(1):303-307.
- [6] 刘嘉敏,彭 玲,袁佳成,等.基于二维变分模态分解和自适应中值滤波的图像去噪方法[J].计算机应用研究,2017,34(10):3149-3152.
- [7] 李 阳,庞永杰,盛明伟.结合空间信息的模糊聚类侧扫声纳图像分割[J].中国图像图形学报,2015,20(7):865-870.
- [8] 沈德海,张龙昌,鄂 旭,等.基于多子窗口的混合噪声滤波算法[J].计算机技术与发展,2015,25(6):69-72.
- [9] 钟 涛,张建国,左俊彦.一种改进的中值滤波算法及其应用[J].云南大学学报:自然科学版,2015,37(4):505-510.
- [10] 李志华,徐小力,王 宁,等.自适应中值滤波在东巴古籍图像去噪中的应用研究[J].北京信息科技大学学报:自然科学版,2015,30(5):36-39.
- [11] 贺东霞,李竹林,王 静.几种滤波算法在医学图像上的实现[J].计算机技术与发展,2014,24(8):165-167.
- [12] LI Zuoyong, LIU Guanghai, XU Yong, et al. Modified directional weighted filter for removal of salt & pepper noise[J]. Pattern Recognition Letters, 2014, 40: 113-120.
- [13] VIJAYKUMAR V R, SANTHANA G, EBENEZER D. Fast switching based median-mean filter for high density salt and pepper noise removal[J]. International Journal of Electronics and Communications, 2014, 68(12): 1145-1155.
- [14] TANWER G, CHAUDHURI S R B. A novel approach to remove random-valued impulse noise from digital image[C]//2016 twenty second national conference on communication. Guwahati, India: IEEE, 2016: 1-6.
- [15] FAN Aiai, WANG Guanglong. A mixed denoising method based on median filter and lifting wavelet technology for sewage sensing signal treatment[J]. Applied Mechanics and Materials, 2013, 330: 967-972.

(上接第 85 页)

- [4] DAI Weiyuan, XUE Guirong, YANG Qiang, et al. Co-clustering based classification for out-of-domain documents[C]//Proceedings of the 13th ACM conference on knowledge discovery and data mining. San Jose, California, USA: ACM, 2007: 210-219.
- [5] 戴文渊.基于实例和特征的迁移学习算法研究[D].上海:上海交通大学,2008.
- [6] JING Xiaoyuan, YING Shi, ZHANG Zhiwu, et al. Dictionary learning based software defect prediction[C]//Proceedings of the 36th international conference on software engineering. Hyderabad: ACM, 2014: 414-423.
- [7] 王 青,伍书剑,李明树.软件缺陷预测技术[J].软件学报,2008,19(7):1565-1580.
- [8] 罗云锋,袁可荣.基于 BBNs 的软件故障预测方法[J].电子学报,2006,34(12A):2380-2383.
- [9] 单锦辉,徐克俊,王 戟.一种软件故障诊断过程框架[J].计算机学报,2011,34(2):371-382.
- [10] LIU Yi, KHOSHGOFTAAAR T M, SELIYA N. Evolutionary optimization of software quality modeling with multiple repositories[J]. IEEE Transactions on Software Engineering, 2010, 36(6): 852-864.
- [11] MA Ying, LUO Guangchun, ZENG Xue, et al. Transfer learning for cross-company software defect prediction[J]. Information and Software Technology, 2012, 54(3): 248-256.
- [12] GRETTON A, BORGWARDT K, RASCH M J, et al. A kernel method for the two-sample problem[C]//Proceedings of NIPS. Minnesota [s. n.], 2006.
- [13] WU Rongxin, ZHANG Hongyu, KIM S, et al. Relink: re-moveing links between bugs and changes[C]//Proceedings of 19th ACM SIGSOFT symposium and the thirteen European conference on foundations of software engineering. Szeged, Hungary: ACM, 2011: 15-25.
- [14] LONG Mingsheng, WANG Jianmin, DING Guiguang, et al. Transfer feature learning with joint distribution adaptation[C]//IEEE international conference on computer vision. Sydney, NSW, Australia: IEEE, 2013: 2200-2207.