厦门理工学院学术专著出版基金资助

基于数据挖掘的软件缺陷预测技术

马 樱 朱顺痣 ◎ 著

厦门大学出版社 国家一级出版社
XIAMEN UNIVERSITY PRESS 全国百佳图书出版单位

基于数据挖掘的软件缺陷预测技术

马　樱　朱顺痣　◎著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

图书在版编目(CIP)数据

基于数据挖掘的软件缺陷预测技术/马樱，朱顺痣著. —厦门：厦门大学出版社，2017.12
ISBN 978-7-5615-6283-3

Ⅰ.①基… Ⅱ.①马…②朱… Ⅲ.①软件-测试 Ⅳ.①TP311.5

中国版本图书馆 CIP 数据核字(2016)第 257380 号

出 版 人 郑文礼
责任编辑 陈进才
封面设计 蒋卓群
技术编辑 许克华

出版发行 厦门大学出版社
社　　址 厦门市软件园二期望海路 39 号
邮政编码 361008
总 编 办 0592-2182177　0592-2181406(传真)
营销中心 0592-2184458　0592-2181365
网　　址 http://www.xmupress.com
邮　　箱 xmupress@126.com
印　　刷 虎彩印艺股份有限公司

开本 787 mm×1 092 mm　1/16
印张 10.25
插页 2
字数 200 千字
版次 2017 年 12 月第 1 版
印次 2017 年 12 月第 1 次印刷
定价 35.00 元

厦门大学出版社
微信二维码

厦门大学出版社
微博二维码

本书由以下项目资助出版：

国家自然科学基金（61502404）

福建省自然科学基金（2015J05132）

福建省教育厅A类科技项目（JA14234）

厦门理工学院学术专著出版基金

前 言

软件缺陷预测是软件工程中的一个非常重要的研究课题。它基于软件历史数据中的模块缺陷记录(包括了软件模块的静态代码特征)来对新的软件模块进行缺陷预测,还能提供决策支持,指导软件项目的规划和过程管理。本书从机器学习的角度分析了软件缺陷预测的特点。同时分析了当前已有的预测方法在应用过程中存在的三个主要问题:(1)因为数据集是来自不同的项目或不同的领域,它们具有不同的数据分布,因此,基于传统方法建立的预测模型具有较弱的适应能力。(2)通过代码审查来手工标记有缺陷的模块是既费资金又费时间的活动,因此,数据集中的正样本数据是有限的。当前的方法是基于监督学习算法建立的预测模型,是仅基于有标记的数据的学习过程。这些方法不能满足要求,因为数量有限的标记数据不具备建立合适的预测模型的足够信息。(3)软件缺陷数据通常是类不平衡的,正样例数远远高于负样例数。该类不平衡问题已经很大程度地影响了缺陷预测模型的性能。本书研究了软件缺陷预测中的最先进的方法技术,包括这些方法的动机,进展,特点和劣势。在此基础上提出了创新的实用技术,并为解决软件缺陷预测中上述的三个问题提供了有效方法。

本书在以下几个方面取得了一些有价值的研究成果。

1. 基于迁移学习的软件缺陷预测研究

提出了新的缺陷预测算法,基于朴素贝叶斯的迁移学习算法(TNB)。不同于已有的预测模型选择与测试数据相类似的训练数据,该算法使用了训练数据中所有特征的实用信息。TNB首先计算出测试数据的分布情况,并将跨公司待预测的数据信息通过赋权的方式作用于训练数据。然后基于这些加权数据,建立缺陷预测模型。笔者还对已有的方法进行了理论分析,并在来自不同组织的数据集上进行了实验比较。结果表明通过TNB建立的预测模型不仅能得到更好的AUC性能,同时具有较低的运行

时间。

2. 基于半监督学习的软件缺陷预测研究

提出了改进的半监督学习方法，用于解决软件缺陷预测中类不平衡和标记数据有限的问题。基于协同训练风范的半监督学习，该方法采用随机抽样技术来对原始训练数据集和每轮更新后训练数据集进行抽样。这能解决半监督学习中的类不平衡问题，使得缺陷预测模型更实用。我们的方法与传统的数据挖掘方法相比，具有更好的预测性能。实验结果还表明，通过解决半监督学习中的类不平衡问题，有可能设计出更好的半监督分类器。

3. 基于主动学习的软件缺陷预测研究

引入主动学习策略，用于减少缺陷预测中标记缺陷模块的代价。本书提出了一种主动学习的方法——两阶段主动学习算法(TAL)，用于预测软件缺陷模块。该方法结合聚类方法和支持向量机技术，提高了预测性能并具有较少的标记代价，并通过实验验证了其有效性。

4. 基于核理论的软件缺陷预测研究

提出了基于核理论的非对称分类器用于建立软件缺陷预测模型。核方法能有效地解决线性不可分数据的分类问题。笔者从理论上分析了类不平衡问题对核主成分分析的影响。在此基础上，提出了非对称核主成分分类器(AKPCC)，试图解决核主成分分析中的类不平衡问题。由于在核主成分分析的基础上弥补了类不平衡问题带来的性能影响，这种方法提高了类不平衡数据集上的预测性能。因此这种方法较其他知名方法具有较优的 F-measure 性能。

作者

2017 年 10 月

ABSTRACT

Developing a good and stable software defect preditor has become a domestic and international scientific frontier, and attracts more and more attention of software industry. We make an analysis of software defect prediction problem from a machine learning perspective, where software characteristics are represented with static code features and defect predictors are learned from historical defect logs. We observe that the present existing defect predictors have reached a limit performance, due to three main problems in the application: (1) Data distributions are different among data sets, which come from different projects or different domains. Therefore, the predictor built with traditional method has weak adaptive ability. (2) Manual labeling defective modules is both costly and time consuming, so the positive data is limited. Recent approaches are based on supervised learning algorithms, which learn predictors with labeled data only. These methods cannot satisfy the requirements, since the limited labeled data are not enough to leaning predictors. (3) The software defect data are always class imbalanced data where the number of positive examples is much higher than that of others. Class imbalance problem has greatly influenced the performance of the defect predictor. In the thesis, we survey the state of the art in software defect prediction research, including motivations, progress, characteristics, and disadvantages. This thesis presents innovative and practical techniques for addressing the three problems mentioned above in software defect prediction as follows:

1. Research on predictive model on transfer learning method

Unlike the prior works selecting training data which are similar from the test data, we propose a novel algorithm called Transfer Naive Bays

(TNB), by using the information of all the proper features in training data. Our solution estimates the distribution of the test data, and transfers cross-company data information into the weights of the training data. On these weighted data, the defect prediction model is built. We also present a theoretical analysis for the comparative methods, and show the experiment results on the data sets from different organizations. It indicates that TNB is more accurate in terms of AUC, within less runtime than the state of the art methods.

2. Research on predictive model on semi-supervised learning method

We present an improved semi-supervised learning approach for defect prediction involving class imbalanced and limited labeled data problem. This approach employs random undersampling technique to resample the original training set and updating training set in each round for co-train style algorithm. It makes the defect predictor more practical for real applications, by combating these problems. In comparison with conventional machine learning approaches, our method has significant superior performance. Experimental results also show that with the proposed learning approach, it is possible to design better method to tackle the class imbalanced problem in semi-supervised learning.

3. Active learning for software defect prediction

We introduce active learning strategies into the defect prediction. An active learning method, called Two-stage Active Learning algorithm (TAL), is developed for software defect prediction. Combining the clustering and support vector machine techniques, this method improves the performance of the predictor with less labeling effort. The experiments validate its effectiveness.

4. Kernel based asymmetric learning for software defect prediction

A kernel based asymmetric learning method is developed for software defect prediction. Kernel method can deal with nonlinearly separable classification problem effectively. We also analyse the effect of class imbalance problem on kernel principal component analysis. The proposed meth-

od Asymmetric Kernel Principal Component Classification (AKPCC) improves the performance of the predictor on class imbalanced data, since it is retrieve the loss caused by class imbalance problem, based on kernel principal component analysis. This method has better F-measure performance than other well-known methods.

简略字表

缩写	完整英文	中文名称	页码
AKPCC	Asymmetric Kernel Principal Component Classifier	非对称主成分分类器	98
APLSC	Asymmetric Partial Least Squares Classifier	非对称偏最小二乘分类	103
AUC	the Area Under the receiver operating characteristic Curve	受试者工作特征曲线下面积	31
CA	Cluster Assumption	聚类假设	61
CC	Cross Company	跨公司模型	45
CCA	Canonical Correlation Analysis	典型相关分析	103
CD	Chebyshev Distance	切比雪夫距离	18
ED	Euclidean Distance	欧氏距离	18
EM	Expectation Maximization	最大期望	25
GM	Generative Model	生成式模型	59
KCCA	Kernel Canonical Correlation Analysis	核典型相关分析	98
KLDA	Kernel Linear Discriminant Analysis	核线性判别分析	98
KLPP	Kernel Locality Preserving Projections	核保局投影	98
KNN	K-Nearest Neighbor algorithm	K 近邻算法	17
KPCA	Kernel Principal Component Analysis	基于核的主成分分析	98
KPLS	Kernel Partial Least Squares regression	核偏最小二乘回归	98
LDA	Linear Discriminant Analysis	线性判别分析	103
MAD	MAnhattan Distance	曼哈顿距离	18
MA	Manifold Assumption	流形假设	61
MDP	Metrics Data Program	软件度量数据项目	2

续表

缩写	完整英文	中文名称	页码
MID	MInkowski Distance	闵可夫斯基距离	18
NASA	National Aeronautics & Space Administration	美国国家航空航天局器	1
NB	Naive Bays	朴素贝叶斯	40
OO	Object Oriented	面向对象	12
PCA	Principal Component Analysis	主成分分析	30
PLS	Partial Least Squares	偏最小二乘	103
RKHS	Reproducing Kernel Hilbert Spaces	再生核希尔伯特空间	96
ROC	Receiver Operating Characteristicl	受试者工作特征	33
Rus	Random Under Sampling	随机抽样	75
RusTri	Random undersampling Tri-training	随机抽样三分类器的协同训练方法	65
SC	Schwarz Criterion	施瓦茨信息准则	24
SMOTE	Synthetic Minority Over-sampling TEchnique	少数类生成过抽样技术	106
SSA	Semi-supervised Smoothness Assumption	半监督平滑假设	61
SVM	Support Vector Machine	支持向量机	21
TAL	Two-stage Active Learning	两阶段主动学习	84
TNB	Transfer Naive Bays	基于朴素贝叶斯的迁移学习	36
YATSI	Yet Another Two Stage Idea	两步分类法	27

主要符号表

符号	定义	符号	定义
x_i	模块 i 的度量属性向量	$H(C\|A)$	每个特征对该数据集划分时的信息量
y_i	模块标记	λ	拉格朗日算子
a_i	模块第 i 个属性	D_S	源域
$\mathfrak{R}^d$	原属性空间	T_S	源域学习任务
$V(G)$	圈复杂度属性	D_T	目标域
b	偏移量	T_T	目标域学习任务
θ'	预测模型	$P(c)$	先验概率
$\langle \cdot,\cdot \rangle$	内积	$P(a_j\|c)$	条件概率
$EV(G)$	基本圈复杂度属性	$P(c\|u)$	后验概率
$IV(G)$	设计复杂度属性	$\|\cdot\|$	集合的成员数目
$\delta(a,b)$	指示函数	$(\cdot)^T$	矩阵转置
$k(\cdot,\cdot)$	核函数	$E\{\cdot\}$	取期望值
μ	均值参数	$H(C)$	每类信息熵
$\sum_i$	协方差矩阵	$(\cdot)^{-1}$	求逆运算
D_l	带标记的数据集	s_i	相似属性个数
D_u	未标记的数据集	w_i	样本权重
$\max\{\cdot\}$	取最大值	$O(\cdot)$	算法复杂度
$\min\{\cdot\}$	取最小值	error_{rate}	整体误差度量
$P_S(X)$	源域数据边缘分布	$P_T(X)$	目标域数据边缘分布
E_T	输出结果方差	H	类间散布矩阵
$\text{argmax}\{\cdot\}$	取极大时值运算	E	类内散布矩阵

续表

符号	定义	符号	定义
$\mathrm{argmin}\{\cdot\}$	取极小时值运算	n_i	第 i 个类的样本数
Ω	模型成员组	$\overline{x}$	总样本集的平均向量
ε	泛化误差	x_{ij}	第 i 类中的第 j 个样本
$\overline{x}_i$	类平均向量	S_b^{Φ}	基于核的类间散度矩阵
C_j	第 j 个聚类质心	S_w^{Φ}	基于核的类内散度矩阵
$\Psi_R(D)$	代表性数据集	$\overline{u}_i$	基于核的各类平均向量
$\Psi_I(D)$	关键信息数据集	$\overline{u}$	基于核的所有样本的平均向量
e	最大错误率	$\Phi(x_j^i)$	基于核表示的第 i 类的第 j 个样本
H_i	第 i 个分类器	μ_{ji}^m	样本模糊数值
$\Phi(x)$	特征映射函数	$\{x_i\}_{i=n+1}^{n+n_t}$	测试数据集
K	Gram 矩阵(或者核矩阵)	K_t	测试数据集核矩阵
$\hat{C}$	对角化协方差矩阵	p_k	模型的后验概率
$\widetilde{u}^k$	第 k 个主成分	$\hat{l}_k(D)$	是当聚类数目为 k 时的模型对应的对数似然概率
$\{x_i\}_{i=1}^n$	训练点集	$J^{j,+}$	将样本 x_j 标记为正类的风险
β_k	第 k 个非线性主成分	$J^{j,-}$	将样本 x_j 标记为负类的风险
$\widetilde{\wedge}$	对角矩阵		

目 录