恶意软件检测中解决样本不平衡问题的策略①

李 瑞, 李希敏, 袁晓玲

(陕西财经职业技术学院 信息工程系, 咸阳 712000)

摘 要: 互联网技术已经使人们的生活和工作发生了巨大的改变. 然而, 人们在享受互联网提供的便利的同时, 也承受着恶意程序带来的威胁. 在数字化时代的今天, 与恶意程序的对抗已成为信息领域的焦点. 由于恶意软件检测中的恶意软件样本难于获取, 同时, 标记大量的样本也需要花费大量的人力和物力, 所获得的恶意软件样本远远少于正常软件样本, 因此各类的训练样本之间存在分布不平衡的分类问题. 为了解决该问题, 本文提出采用 SMOTE 过采样方法, 通过合理的增加少数类样本来解决样本不平衡问题.

关键词: 恶意软件; 不平衡问题; SMOTE

Malware Detection in the Strategy to Solve the Problem of Unbalanced Samples

LI Rui, LI Xi-Min, YUAN Xiao-Ling

(Information Engineering Department, Shanxi Vocational College of Finance and Economics, Xianyang 712000, China)

Abstract: Great changes have occurred in people's daily life and routine work due to the widely used internet technology. However, we have to face threaten from the malware. Due to this, detecting malware has received more and more attentions in recent years. Malware samples are hard to obtain. Meanwhile, it needs to cost a lot of resources. So there are less malware samples and malware detection is an imbalance problem. Imbalance problem means that the distributions of various types of training samples are imbalanced. To solve this problem, a suitable over-sampling method is employed via a reasonable increase in samples of a few samples to address imbalances.

Keywords: malware detection; imbalance problem; SMOTE

1 引言

1.1 论文研究背景及意义

恶意软件是一种利用计算机软、硬件所固有的脆弱性所编制的具有破坏功能的程序. 近年来, 互联网上研究恶意软件代码的网站不断增多, 人们可以直接从网上获得恶意软件代码的源码, 这些行为推动了恶意软件代码基本编写技术的普及; 另外, 伴随着互联网的广泛应用, 恶意软件问题越来越受到人们的关注, 有些恶意软件借助网络爆发流行, 极大地加快了病毒的传播速度, 加重了它的破坏强度. 而恶意软件的攻击可能造成敏感信息泄漏、主机系统崩溃和阻塞网络等各种各样严重的后果. 恶意软件巨大的破坏性和快速的传染性, 给人们的生产和生活造成了巨大的不便. 因而深入研究恶意软件的检测技术具有非常重要的现实意义.

恶意软件检测技术已成为当前网络安全技术领域内的一个研究热点, 目前已有的各种检测技术如启发式代码扫描、基于免疫原理的病毒检测技术和基于程序行为的病毒检测技术等各有特点, 但是应用起来仍然不够成熟, 且均有其局限性. 比如基于行为特征的检测方法无法对抗垃圾行为插入等行为混淆方法的干扰.

1.2 本文研究内容

针对传统的恶意软件检测方法存在的缺陷, 机器学习成为恶意软件检测的新方向. 基于机器学习的检测恶意软件技术主要通过学习恶意软件和正常程序的差异性发现有关的识别模式, 并利用这些模式进行相似性分类以发现含有类似模式的恶意软件.

① 收稿时间:2013-10-16 收到修改稿时间:2013-11-11

机器学习和模式分类算法应用于恶意软件检测, 主要利用了其分类的功能. 其基本思想是: 首先, 从样本数据集中选择一组数据作为训练集, 分析该训练数据样本属性建立一个分类模型, 描述预定的数据样本集. 假定每个样本属于一个预定义的类, 由类标号的属性确定. 分类模型可以用分类规则、判定树和其他数学公式的形式表示.

然后利用训练好的模型进行分类. 用该模型对测试集进行分类预测, 对于每个测试样本, 将已知的类标号与该样本的学习模型类预测结果进行比较, 如果结果一致, 表明分类预测是正确的. 如果模型的识别率达到要求, 就可以用它对类标号未知的样本数据进行分类. 利用数据挖掘和机器学习技术, 可以从大量的恶意程序和正常程序数据中学习到正确有效的识别信息, 形成区别恶意软件与正常代码的分类标准, 待检测程序通过训练过的分类器进行分类预测. 这种方法能够检测未知恶意程序, 最大程度防止系统受到破坏, 是目前重点的研究方法. 已有学者采用集成学习的方法实现了对恶意软件的检测[1].

随着互联网的发展, 恶意软件的数量越来越多. 尽管如此, 与正常软件的数量相比, 恶意软件仅仅占很小的比例, 这极大的增加了恶意软件检测的难度. 另外, 大量恶意软件样本也很难获取. 因此造成了样本的不平衡, 即正常软件数量远远多于恶意软件数量. 数据挖掘和机器学习已经广泛应用于恶意软件检测中, 然而, 传统的分类技术往往没有考虑样本不平衡问题. 这就造成了分类器可能将测试样本全部判别为多数类的情况, 而在恶意软件检测中, 我们更关心恶意软件也就是少数类的检测性能.

本论文提出通过采用 SMOTE 过采样方法对已经提取出来的恶意数据样本进行处理.

2 样本数据不平衡

2.1 不平衡分类问题概述

所谓不平衡分类问题, 是指各类的训练样本之间存在分布不平衡的分类问题. 以二分类问题为例, 一类样本的数量要远远少于另一类样本的数量. 在许多实际的应用领域中, 不平衡分类问题非常常见. 例如, 在欺骗信用卡检测、信息检索、医疗诊断以及基因序列中编码信息的挖掘等等问题均属于不平衡分类问题. 这些问题中对少数类的识别更为重要. 在医疗诊断中, 如果把正常人误诊为病人会给他带来精神上的负担, 而如果把一个病人误诊为正常, 则极有可能错过最佳治疗时间, 导致严重的后果.

不平衡问题早在 1998 年就提出, Kubat 等人根据获得的卫星图像, 通过分类的方法对石油井喷进行计算机自动检测[2]. 这其中数据的不平衡比例大概为 22 比 1. 此后, 又有学者研究了欺诈识别中的不平衡分类问题. Castillo 等人研究了文本分类问题[3][4]; Cohen 等人研究了医院传染病检测问题[5]; Yoon 等学者研究了生物信息学中不平衡问题的解决[6].

由于不平衡问题存在于许多实际问题中, 因此研究者对该问题已经做了大量的研究. 其相关研究主要围绕以下 3 个层次展开: (1)分类器的合理评价; (2)数据层面的处理方法; (3)算法层面的处理方法. 本文在此将从以上三个层面对不平衡分类问题做出综述[7].

2.1.1 分类器的合理评价[8]

作为一个分类器, 分类精度是主要的评价指标, 它反映了分类器对数据集的整体分类性能. 然而对不平衡分类问题而言, 用精度评价分类器的性能却并不合理. 以二分类为例, 假设正类样本占了 99%, 若将所有样本均归为正类, 那么即使获得了 99%的训练精度, 这样的分类器仍然是没有价值的. 在实际问题中, 我们往往更加关注少数类. 所提出的评价指标也更注重少数类对性能的影响. 常用的评价指标是 ROC 曲线, 以及该曲线下所覆盖的面积 AUC[9]. 由于 ROC 曲线和 AUC 可以客观地对待少数类和多数类, 因而可以在少数类识别率和大类识别率之间做出权衡. 为了定义 ROC 曲线, 在一个二分类问题中, 我们考虑表 1 所示的混淆矩阵. 其中: Pos 表示正类样本, Neg 表示负类样本; N=Pos+Neg 表示全体学习样本. TP(true positive)和 TN(true negative)分别表示被正确分类的正类和负类样本; FP(false positive)和 FN(false negative)分别表示被错分的正类和负类样本.

表 1 混淆矩阵

	预测正类	预测反类
预测正类	TP	FN
预测反类	FP	TN

利用混淆矩阵所定义的评价指标[7][8]如下所示:

平衡准确率 BA 为:

$$(TP/(TP+FN)+TN/(TN+FP))/2 \quad (1)$$

查全率 TPR 为:

TP/(TP+FN) (2)

查准率 P 为:

TP/(TP+FP) (3)

误警率 FPR 为:

FP/(FP+TN) (4)

计算 ROC 曲线时, X 轴表示 FPR, Y 轴表示 TPR. ROC 曲线上的一组点能够通过对分类器决策阈值调整得到, ROC 曲线越凸同时越靠近左上方, 表明分类器的泛化能力越强. AUC(Area Under Curve)是指 ROC 曲线下面包括的面积, 也就是 ROC 曲线的积分, AUC 能以定量的方式表示该 ROC 曲线对应的分类器的一般化能力. 值得注意的是, ROC 和 AUC 仅适合于两类问题, 对于多类问题, 不能直接应用.

2.1.2 数据层面的处理方法

数据层面的处理是对数据进行重采样, 包括过抽样和欠抽样两种. 其主要思想是通过合理地增加或者减少一些样本去平衡化数据, 以降低数据不平衡所带来的对分类器的影响. 具体描述如下:

(1)过抽样

过抽样是解决不平衡问题最常用的方法. 其基本思想是通过改变训练数据的分布来消除或减少少数类的不平衡. 过抽样通过增加少数类样本来提高少数类的分类性能[10], 其缺点是没有给少数类数据增加任何新的信息. 有学者提出了改进的过抽样方法, 该方法通过在少数类中加入随即高斯噪声或产生新的合成样本等, 如 SMOTE 算法[11].

(2)欠抽样

该方法通过减少多数类的样本数量来提高分类性能, 最直观的方法就是随机地去除一些多数类样本以减少多数类的样本规模, 而缺点则是容易丢失一些重要信息[12]. 有学者提出了基于遗传算法的方式将多数类训练集划分成几个簇[13], 每个簇都包含一定数目的多数类和少数类, 最后将从每个簇中取出的样本进行合并, 得到一个新的训练集, 再对其进行训练.

2.1.3 算法层面的处理方法

在算法层面, 针对不平衡分类问题, 主要通过改进原有算法(如 SVM, Adaboosting 等)或者设计更有效的新算法为主要的研究途径. 目前所提出的方法主要有以下几部分:

(1)代价敏感学习

在大部分不平衡分类问题中, 少数类是分类的重点. 在这种情况下, 将少数类正确的识别要比识别出大类更有价值. 代价敏感学习赋予各个类别不同的错分代价, 因而能够很好地解决不平衡分类问题. 以两类问题为例, 假定正类是少数类, 并且具有更高的错分代价, 那么在训练分类器时, 会对错分的正类样本做出更大的惩罚, 迫使训练后的分类器对正类样本有更高的识别率. 当前对代价敏感学习的研究主要集中在以下两方面: 一种是根据样本的不同错分代价重构训练集, 而不改变已有的分类算法. 第二种方式是在经典的分类算法中引入代价敏感因子, 设计相应的代价敏感分类算法. 在该方法中不同类的错分代价是不同的, 小样本赋予较高的代价, 大样本赋予较小的代价, 以此来平衡样本之间的数目差异.

代价敏感学习能够有效地提高少数类的识别率, 但也有其不足之处. 一方面, 在多数情况下, 真实的错分代价很难被准确地估计. 另一方面, 虽然许多分类器可以直接引入代价敏感学习, 如支撑向量机和决策树, 但是也有一些分类器不能直接使用代价敏感学习, 只能通过调整正负样本的比例或者决策阈值间接地实现代价敏感学习, 这样不能保证代价敏感学习的效果.

(2)集成学习方法

集成学习方法的主要思想在于将多个子分类器组合成一个集成分类器, 以提高分类性能. 在众多集成方法中, AdaBoost[12]方法是解决不平衡分类问题较好的一种方法. 但有实验也表明, AdaBoost 提高少数类样本的识别率的能力有限, 因为该方法是以整体分类精度为目标, 多数类样本由于数目多所以对精度的贡献大, 而少数类样本则由于数目少而贡献非常小, 因而分类决策是不利于正类的. 为了解决以上问题, 一些改进算法相继提出, 比如 AdaCost 和 RareBoost, 其主要策略是改变权值来更新规则, 使分类错误的少数类样本比多数类样本有更高的权值. 还有学者将 Boosting 算法与 SMOTE 算法结合成 SMOTEBoost 算法, 该算法每次迭代使用 SMOTE 生成新的样本, 取代原有 AdaBoost 算法中对样本权值的调整, 使得 Boosting 算法专注于少数类样本中的难分样本.

(3)单类分类器方法

在实际应用中, 有时要获取两类或者多类样本是非常困难的, 或者就是需要很高的成本, 只能获取单类样本. 此时对只含有单一类的数据进行训练是唯一

可能的解决办法. 单分类器是用来对只有一种类别的训练集进行分类的, 它能够有效地解决不平衡分类问题. 目前研究的比较多的还是基于 SVM 的方法[14].

(4)其他方法. 主动学习、子空间方法、特征选择方法、随机森林以及基于 SVM 的后验概率求解方法[15]等也是学习不平衡数据集的有效方法.

以上对不平衡分类问题做出了综述. 不平衡数据的存在是机器学习不能在工程实际中广泛应用的重要原因之一, 因而近年来引起了广泛的关注. 而恶意软件检测中, 由于恶意软件相对于正常软件, 始终属于少数类, 因而在恶意软件检测过程中也存在不平衡分类问题.

2.2 不平衡问题的特点

不平衡问题有三个显著的特点:

(1)难区分性: 这是不平衡问题的难点之一, 即目标类数据集中比例小, 从多数类中区分少数类难区分是个固有问题. 工程实际中有些问题是很容易区分的, 不需要复杂分类算法. 所以研究特定的少数类分类方法是现实需要.

(2)多态性: 一般情况下, 无论少数类还是多数类均可能包含多个属性, 并能划分为多个不同的子类, 而每个子类又有不同的特性. 因而该问题就转换成从多数类的许多个子类中区分出少数类及其子类, 这当中可能有些容易分开, 有些不容易分开. 这就形成了多个子类之间的多态性, 进而使问题变得复杂.

(3)稀有性: 稀有性是解决这种问题的关键. 如何快速、有效地识别少数类同时能够避免分类器对训练数据集的过拟合是目前需要尽快解决的问题.

3 数据预处理

当未知软件的特征提取出来之后, 由于恶意软件检测属于不平衡问题, 因而首先要对数据做预处理. 利用相应的不平衡策略对初始样本处理之后, 将产生的新样本与原始样本合并组成新的训练样本用于训练检测模型.

3.1 数据预处理需求

由于恶意软件检测问题属于机器学习中的不平衡问题, 因而首先需要通过相应的预处理, 将不平衡问题转化为机器学习可以解决的平衡分类问题. 其功能模型如图 1 所示.

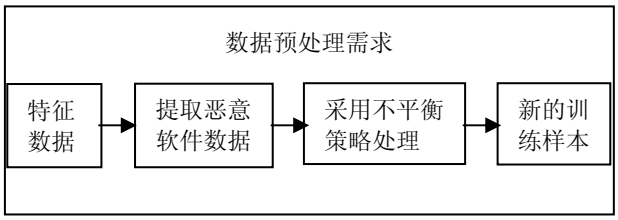

图 1 功能模型图

在利用相应的方法提取出特征数据之后, 由于恶意软件数据属于少数类样本, 因而只需对该类样本做出处理. 所以接下来将恶意软件数据提取出来, 然后采用相应的不平衡策略对恶意软件数据进行处理, 并且产生新的数据, 最后将原有数据与新产生的数据组合起来形成新的训练样本, 用于接下来的模型训练. 其中, 本论文主要通过采用 SMOTE 过采样方法对已经提取出来的恶意数据样本进行处理, 经过合理的增加恶意数据使之成为新数据样本, 然后将这些新数据样本和原有的特征数据组成新的数据样本, 作为下一步的训练样本.

3.2 数据预处理的主要策略

由于恶意软件相比于正常软件而言, 所能提取的样本数量少之又少, 因而两类样本之间存在不平衡. 如果直接采用这样的样本训练分类模型, 所得到的模型显然不能满足需求. 因此我们这里设置了数据预处理模块, 以解决数据样本之间的不平衡问题. 首先从已有的特征数据中将恶意软件样本提取出来, 然后利用不平衡策略合理地增加恶意软件样本. 最后将增加的样本与原有特征数据合并, 组成新的训练样本. 在数据预处理阶段, 本论文主要解决的问题是选择合适的方法对已经提取出来的恶意数据样本进行处理, 然后将经过处理的恶意数据样本和原有的特征数据组成新的数据样本, 作为下一步的训练样本, 用于训练分类模型.

3.3 解决不平衡问题的策略选择

在恶意软件检测中, 由于恶意软件相对于正常软件, 始终属于少数类, 该问题属于恶意软件检测中的不平衡分类问题, 因此本文将采用现有的研究成果对该问题提出解决方法, 即利用 SMOTE 方法来解决在恶意软件检测中存在的不平衡分类问题.

SMOTE(Synthetic Minority Over-sampling Technique)方法是一种新的过采样技术, 其通过产生少数类的合成样本, 并且控制新生成样本的数量与分布来平衡样本数据集. 该方法可以有效解决传统的过采样方法由于决策区间过小造成的分类器过拟合问

题. 同时该方法也是当前在工程实际中应用最为成功的一种方法. 因此, 本文将SMOTE方法用于解决样本的不平衡问题. 由于 SMOTE 算法是一种过采样方法, 因而这里主要通过对恶意软件过采样产生新的样本, 以降低不平衡样本带来的问题. 采用 SMOTE 对恶意软件过采样的步骤如下所示:

假设恶意软件的样本数量为 M, 正常软件的样本数量为 N, 这里 M 远远少于 N. 为了保证样本数量的平衡, 我们将产生 N-M 个恶意软件样本. 产生样本的过程如下所示:

(1)计算恶意软件样本集中每个样本 xi 的 K 个最近邻点;

(2)随即抽取少数类样本 xi 的最近邻点中的一个样本点 xj;

(3)计算样本点 xi 和 xj 之间的第 t 个属性的差距 d

$$a=xit-xjt \quad (3\text{-}1)$$

(4)产生一个介于 0 和 1 之间的随机数

$$r=rand(1) \quad (3\text{-}2)$$

(5)生成合成样本点, 其中第 t 属性值为:

$$Xht=xit+r*d \quad (3\text{-}3)$$

(6)重复(2)—(5)步, 直至产生 N-M 个样本.

以上为恶意软件样本产生的过程. 将新产生的样本和原有样本合起来组成新的样本集合, 可用于训练分类器以及恶意软件预测.

4 论文总结

近年来, 机器学习技术应用于恶意软件检测领域, 取得了一定的成果. 利用机器学习技术, 可以从大量的样本数据中学习到用以区分恶意程序和正常程序的分类规则, 可以利用这些分类规则构建分类器, 对未知程序的类别进行预测. 但由于恶意软件相对于正常软件, 始终属于少数类, 该问题属于恶意软件检测中的不平衡分类问题, 然而, 传统的分类技术往往没有考虑样本不平衡问题. 这就造成了分类器可能将测试样本全部判别为多数类的情况, 因此本文提出了采用 SMOTE 方法来解决在恶意软件检测中存在的不平衡分类问题.

参考文献

1 刘慧.数据相关选择性集成的恶意软件检测技术研究[学位论文].西安:西安电子科技大学,2011.

2 Kubat M, Holte RC, Matwin S. Machine learning for the detection of oil spills in satellite radar images. Machine Learning, 1998, 30 (2-3): 195–215.

3 Castillo MD, Serrano JI. A multistrategy approach for digital text categorization from imbalanced documents. SIGKDD Explorations, 2004, 6 (1): 70–79.

4 Zheng ZH, Wu X, Srihari RK. Feature selection for text categorization on imbalanced data. SIGKDD Explorations, 2004, 6(1): 80–89.

5 Cohen G, Hilario M, Sax H. Data imbalance in surveillance of nosocomial infections. Proc. of the 4th International Symposium on Medical Data Analysis. Berlin. 2003. 109–117.

6 Yoon K, Kwek S. An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in function genomics. Proc. of the 5th International Conference on Hybrid Intelligent Systems. 2005. 303–308.

7 叶志飞,文益民,吕宝粮.不平衡分类问题研究综述.智能系统学报, 2009,(4):148–156.

8 林志勇,郝志峰,杨晓伟.不平衡数据分类的研究现状.计算机应用研究, 2008(2):332–336.

9 Brefeld U, Scheffer T. AUC maximizing support vector learning. Proc. of ICML Workshop on ROC Analysis in Machine Learning. Bonn. 2005.

10 许丹丹,蔡立军,王勇.一种改进的少数类样本过抽样算法.计算机工程,2012,(4):67–69.

11 薛薇.非平衡数据集的改进 SMOTE 再抽样算法.统计研究,2012,(6):95–98.

12 孙晓燕,张化祥,计华.基于 AdaBoost 的欠抽样集成学习算法.山东大学学报(工学版),2011,(4):91–94.

13 李琳.基于粗糙集和遗传算法的聚类方法研究[学位论文].桂林:广西师范大学,2009.

14 赵玲,陈磊琛,余小陆,张盛意.SVM-KNN 分类算法研究.计算机与数字工程,2010,(6):29–31.

15 王鹏伟,李滔,吴秀清.一种基于 SVM 后验概率的 MRF 分割方法.遥感学报,2008,(2):208–214.