doi:10.3772/j.issn.1002-0470.2016.05.005

基于协同采样主动学习的恶意代码检测①

张 凯②*** 王东安③** 李 超** 贾 冰****

(*中国科学院信息工程研究所 北京 100093)

(**国家计算机网络应急技术处理协调中心 北京 100029)

(***中国科学院大学 北京 100049)

(****河南省工人文化宫 郑州 450007)

摘 要 研究了基于机器学习分类算法的恶意代码检测,考虑到目前主要采用传统分类方法对恶意代码进行分类识别,这些方法需要通过学习大量标记样本来获得精准的分类器模型,然而样本标记工作只有少数专家才能完成,导致标记样本往往不足,致使分类结果准确率不高,提出了一种基于协同采样的主动学习方法。运用这种学习方法,仅需少量标记样本即可有效识别出恶意代码。实验证明,相对于传统的恶意代码分类方法,该方法能够显著提升分类准确率和泛化性能。

关键词 主动学习,支持向量机(SVM),概率性神经网络(PNN),协同采样

0 引 言

恶意代码特指对用户电脑和网络造成危害的恶意软件(malicious software),其英文写为 malware。近年来,由于互联网的飞速发展以及各种利益的驱使,恶意代码已经对网络空间造成重大威胁,而且其产生速度逐步加快,恶意代码家族或变种的数量急剧增长,人工分析手段已无法有效地应对。在此情况下,开始引入机器学习分类算法对恶意代码进行检测识别。传统机器学习分类方法通过学习标记样本训练模型,实现对未标记样本的自动分类,很大程度上减少了人工成本[1-10]。然而为了获得精准的分类模型,在训练时需要学习大量标记样本来建立模型,仍然需要较大的人工工作量,由于标记工作只有少数相关领域专家才能完成,标记样本数量往往不足,导致所建立模型的精确性差,致使分类结果准确率不高。此外,传统机器学习分类方法是利用标记样本来训练分类器,利用该分类器对新的数据集进行预测时,分类器并不发生任何变化,而该分类器并不一定适用于新的数据集,因而泛化能力不强。

针对上述问题,本文提出一种采用基于协同采样的主动学习的恶意代码检测方法,以有效降低对标记样本数量的需求。实验结果表明,在标记样本数目较少的情况下,仅使用少量的标记样本即可有效识别出恶意代码。相对于未使用主动学习的机器学习方法,在标记样本数量少的情况下,能够显著提升分类性能。同时该方法引入主动学习思想,与未知样本数据集进行交互,有效地提升了泛化能力。

1 相关工作

本节对算法设计中涉及的主要方法进行介绍,包括特征提取方法、主动学习方法以及选用的基准分类器——概率神经网络(probabilistic neural network, PNN)分类器和支持向量机(SVM)分类器。

① 国家自然科学基金(61202067, 61271275)和 863 计划(2012AA013001, 2013AA013205, 2013AA013204)资助项目。

② 男,1987 年生,博士生;研究方向:数据挖掘,信息安全;E-mail:zhangkai@iie.ac.cn

③ 通讯作者,E-mail:549036597@qq.com

(收稿日期:2015-12-16)

1.1　特征提取

本文旨在提取高级行为特征和威胁文本特征,并融合这两种特征以实现恶意代码分类。目前恶意代码高度模块化和功能多样化趋势越来越强,需要更为全面的特征要素描述恶意代码本质特性[11]。首先,恶意代码功能主要体现在其在系统中的行为;其次,一个正常的行为可能因配置的内容不同,其性质有所差别。因此本文以高级行为和威胁字符作基础分别提取高级行为特征和威胁文本特征,并构建组合特征空间模型。其中高级行为特征是由特定系统调用转换的高级行为及参数的集合,首先通过系统监控方法获取样本调用的原始 API 和详细参数信息,然后根据规则从中提取出高级行为。威胁文本特征是样本二进制文件包含的可显示字符串经过信息熵过滤后得到的集合。首先通过沙盒、虚拟机等监控系统分析执行过程中的二进制文件,输出其中可显示字符串,然后通过信息熵过滤提取作为威胁文本特征。

1.2　主动学习

主动学习是相对于被动学习提出的[12-15]。被动学习指的是随机从数据集中选取一定数目的样本进行标记,并利用标记样本训练分类器,然后对未标记样本进行分类。然而由于该方法获得的标记样本随机产生,因此获得的样本很可能存在信息冗余、噪声过多等问题,从而严重影响分类效果。主动学习则是将样本标记工作分为两步,首先标记少量样本作为初始训练集训练基准分类器,并利用该分类器对未标记样本分类,然后根据一定的规则,从分类结果中筛选一定数量的样本进行标记,并将标记后的样本与初始训练集结合形成新的训练集,以此训练新的分类器,最后利用新的分类器对未标记样本进行分类得到最终结果[16-19]。

根据上述说明,主动学习可以分为两个步骤,因此主动学习算法可以相应包含学习引擎和采样引擎两个部分。学习引擎指的是实现分类的机器学习方法,例如本文中应用的概率神经网络(PNN)算法和支持向量机(SVM)算法。采样引擎指的是主动学习第二步中筛选样本的策略。而主动学习的核心思想也在于采样引擎的设计。根据采样策略的不同,可以将主动学习算法分为基于成员查询综合(membership query synthesis,MQS)的主动学习、基于流的(stream-based)主动学习和基于池的(pool-based)主动学习[20]三种。其中,基于池的主动学习算法是目前研究最充分、使用最广泛的一类策略。按照选择未标记样本标准的不同,基于池的主动学习算法又可分为基于不确定性的抽样(uncertainty based sampling,UBS)策略、基于委员会投票的抽样(query by committee,QBC)策略和基于估计误差缩减(estimated error reduction,EER)的抽样策略等几种[21]。本文采用的协同采样方法隶属于委员会投票抽样(QBC)策略。

1.3　概率神经网络

概率神经网络(PNN)于 1989 年由 Specht 博士首先提出,是一种常用于模式分类的前馈型神经网络[22]。PNN 包含 4 层,分别为输入层、模式层、求和层、输出层,如图 1 所示。其中 X 为样本的特征属性,Y 对应的是样本的目标属性。本文中的 PNN 应用中,X 即为由高级行为特征和威胁文本特征组合而成的融合特征,Y 则对应 0 和 1 两个结果,分别指带是否为恶意代码。

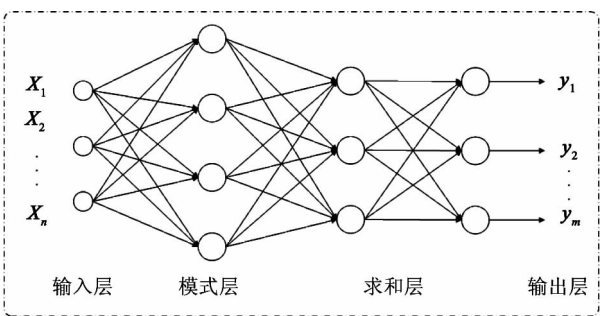
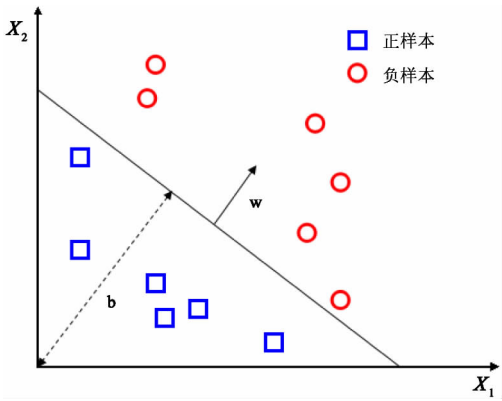


图 1　PNN 分类模型结构示意图

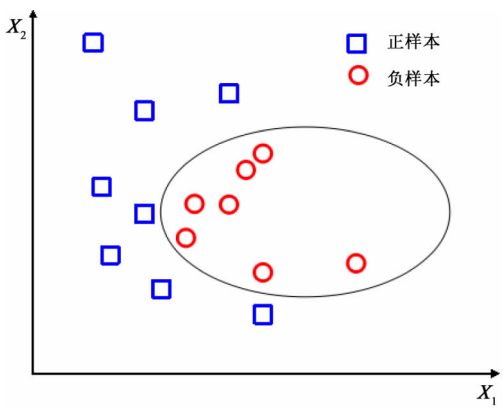
1.4　支持向量机

支持向量机(support vector machine,SVM)是基于监督学习的机器学习方法,可以分为两种处理情况:线性可分和非线性可分[23]。目标是获取最佳的分类面,即最大间隔分类器。在二维空间中,样本表现为平面上的点,可以使用支持向量机训练已标记样本,获得分界线(对于线性的情况,分界线为直线,而对于非线性的情况,则为曲线),将正负样本

分离开来。同样的，在 N 维空间中，我们可以使用支持向量机获取一个面，从而对正负样本进行分离（对于线性的情况，分界面为平面，而对于非线性的情况，则为曲面）。图 2(a)和图 2(b)分别为线性情况和非线性情况。



(a) 线性情况 SVM 分类器



(b) 非线性情况 SVM 分类器

图 2　SVM 分类器

2　算法设计

算法的核心思想是把对标记样本的标记工作分为两次完成。首先从未标记样本集中随机选取一定数量的样本进行第一次标记，并根据这些标记样本训练两个分类器，本文中选择使用 SVM 分类器和 PNN 分类器。利用两个分类器对未标记样本进行分类，比较两个分类结果并提取结果不一致的样本，然后再按照特定规则（本文中选用 SVM 分类器分类结果对应的置信度 d 做标准）筛选得到一定数量样本并做第二次标记工作。由于本次标记的样本很可能是分类器错误的分类结果，因此相对于其他的样本更有助于提升基准分类器性能。

算法执行的步骤如下：

步骤 1：利用训练集 Tr 学习获得 SVM 分类器 $C1$ 和 PNN 分类器 $C2$。

步骤 2：利用分类器 $C1$ 和 $C2$ 对测试集 Te 分类，分别得分类结果 $Ye1$ 和 $Ye2$。

步骤 3：比较分类结果 $Ye1$ 和 $Ye2$，选取结果不一致的样本形成集合 $Temp$，并根据参数 d 对样本进行升序排序。

步骤 4：选取前 $aNum$ 个样本形成主动学习样本集 Ta，并结合 Tr 形成新的测试集 Tr'。

步骤 5：利用 Tr' 再次对测试集 Te 分类，得最终分类结果 Ye'。

步骤 6：根据 Ye' 计算分类器评价标准参数，包括召回率、精度、F-Measure、准确率。

算法伪代码如下：

算法　基于协同采样主动学习的恶意代码检测

输入：训练集 Tr，测试集 Te，主动学习样本数目 $aNum$；

输出：分类评价参数指标（包括召回率、精度、F-Measure、准确率）；

```
Begin:
    [C1,C2] = train(SVM,PNN,Tr);
    [Ye1,Ye2] = text(C1,C2,Te);
    Temp = Compare(Ye1,Ye2);
    Temp = Qsort(d,Temp);
    Ta = top(aNum,Temp);
    Tr' = combine(Tr,Ta);
    C' = train(SVM,Tr');
    Ye' = text(C',Te);
    Evaluate(Ye');
```

由于主动学习把样本的标记工作分为了两步，而第二步中添加的样本是根据特定规则筛选出的信息丰富的样本，因此使得分类效果显著提升。而且由于在主动学习中会与未标记样本进行交互，因此泛化性能也得到显著加强。

3　实验结果与分析

3.1　实验数据集

实验共采用 840 个样本,其中黑样本(恶意代码)361 个,白样本(非恶意代码)3479 个。我们选取 100 个样本作为候选训练集,另外的 740 个样本做为测试集。经划分后,测试集样本数目为 740 个,黑白样本分布为:白样本 479 个,黑样本 361 个。实验中应用的组合特征由恶意代码高级行为特征和威胁文本特征两部分组成,其中高级行为特征维度为 131 维,威胁字符特征维度为 103734 维,经主成分分析(PCA)处理后为 131 维。组合特征即为高级行为特征和威胁字符特征的融合,维度为 262 维。

3.2　评价指标体系

实验中,我们利用召回率(recall)、精度(precision)、F1 值(F1 measure)和准确率(accuracy)四项性能参数对算法进行评价。四项性能评价指标参数的计算公式为

(1) 精度 $P = \frac{TP}{TP + FP}$;

(2) 召回率 $R = \frac{TP}{TP + TN}$;

(3) F1-Measure $F1 = 2\frac{P * R}{P + R}$;

(4) 准确率 $A = \frac{TP + FN}{TP + FP + TN + FN}$。

其中参数 TP、TN、FP、FN 的说明见表 1。

表 1　性能评价指标计算参数说明

参数	含义	说明
TP	True Positive	正样本被预测为正样本
TN	True Negative	正样本被预测为负样本
FP	False Positive	负样本被预测为正样本
FN	False Negative	负样本被预测为负样本

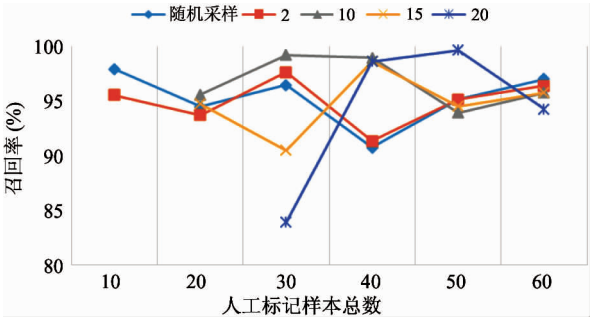
3.3　实验结果分析

实验的思路是,在标记样本总数相同的情况下,对比随机选取样本方法(即传统分类方法)和协同采样主动学习方法的分类效果。此外,又可以根据主动学习的参数不同分为多种情况。因此实验中,需要设定的参数有两个,一个是总共需要标记的样本数目 *rNum*,另一个是协同采样主动学习时需要标记的主动学习样本数目 *aNum*。本文在 *rNum* 分别为 10、20、30、40、50、60 六种情况下,对两种方法的性能进行测试。而其中采用协同采样主动学习方法时,又分为 *aNum* 取值 2、10、15、20 四种情况。

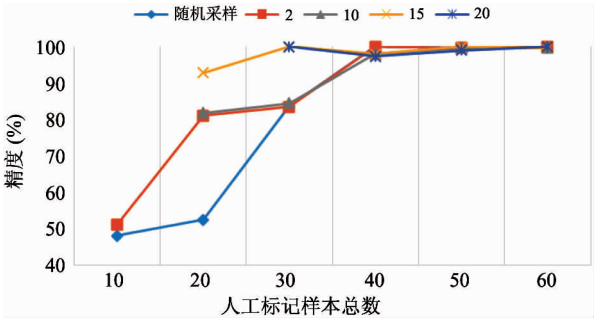
图 3(a) ~ (d)为实验结果,其中横坐标为标记样本数目的总数 *rNum*,纵坐标分别为召回率、精度、F1 值和准确率。图中标出了随机采样方法与四种情况下的协同采样主动学习方法,共 5 条曲线。由于精度和召回率两个指标相互制约,而 F1 值则是对二者的融合,因此重点采用 F1 值和准确率评价算法。从图 3(c)和 3(d)可以看出,采用协同采样主动学习方法其 F1 值和准确率均显著优于随机采样方法,尤其是在标记样本数目较少时,效果更为明显。例如在标记样本总数为 20 时,采用协同采样主动学习方法的 F1 值和准确率相对于随机采样方法,分别平均提高22.08% 和13.10%。

3.4　讨　论

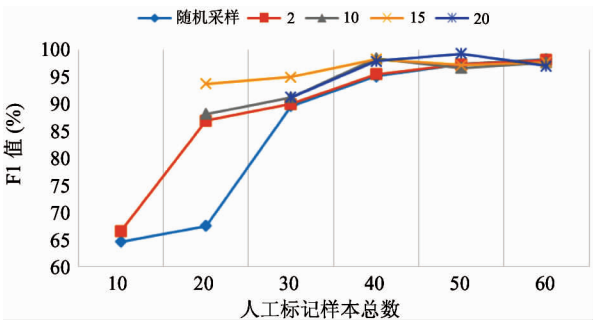
通过实验,证明了基于协同采样的主动学习方法的有效性。相比于随机采样的分类方法,采用基于协同采样的主动学习方法能够显著提高分类的各项性能指标。尤其是在标记样本总数相对较少情况下,提升更为明显。此外本文实验是将总体数据集划分为相互独立的训练集和测试集,因此也证明了协同采样主动学习方法优异的泛化性能。该方法有效的原因在于主动学习样本集中主要是基准分类器不能够正确分类的样本,这些样本相对于随机选取的样本,包含更多有益于提升基准分类器性能的信息。此外由于进行两轮训练,且第二轮训练中通过引用主动学习样本引入测试集的样本信息,从而提升了分类器的泛化性能。



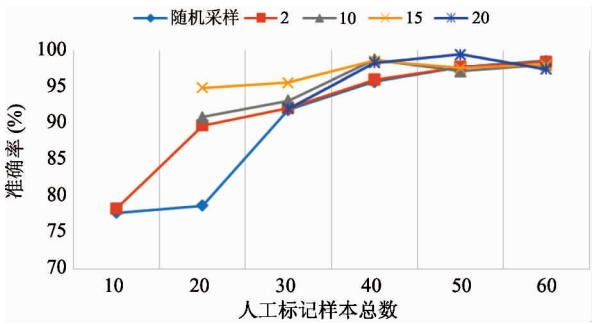
（a）召回率对比图



（b）精度对比图



（c）F1 值对比图



（d）准确率对比图

图 3　实验结果

4　结　论

面对每天捕获到的大量多样的恶意代码，尤其是针对一些特殊来源收集到的样本需要尽可能准确分析，而由于专家人数的限制使得人工标注样本成本昂贵。因此如何能够在标记样本数量相对不足情况下保证分类效果，成为亟待解决的问题。针对传统恶意代码分类器由于标记样本不足导致的分类效果不佳的问题，本文提出了协同采样主动学习方法，该方法在标记样本数量少的情况下，能够显著提升分类器的各项性能。

参考文献

[1] Kolter J Z, Maloof M A. Learning to detect malicious executables in the wild. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, USA, 2008. 470-478

[2] Balley M, Oberheide J, Ander J, et al. Automated classification and analysis of internet malware. In: Recent Advances in Intrusion Detection. Springer Berlin Heidelberg, 2007. 178-197

[3] Zhu X B, Jin X, Zhang X Y, et al. Context-aware local abnormality detection in crowded scene. *Science China Information Sciences* (*SCIS*), 2015, 58(5): 1-11

[4] Zhang X Y, Xu C, Cheng J, et al. Automatic semantic annotation for video blogs. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hannover, Germany, 2008. 121-124

[5] Zhu G, Wang J, Wu Y, et al. MC-HOG correlation tracking with saliency proposal. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2016. 1-7

[6] Wang S, Zhang X Y, Yun X, et al. Joint recovery and representation learning for robust correlation estimation based on partially observed data. In: Proceedings of the IEEE International Conference on Data Mining (ICDM) Workshop, Atlantic City, USA, 2015. 1-7

[7] Zhang X Y. Preference modeling for personalized retrieval based on browsing history analysis. *IEEE Transactions on Electrical and Electronic Engineering*, 2013, 8 (S1): 81-87

[8] Zhang X Y. Effective search with saliency-based matching and cluster-based browsing. *High Technology Letters*, 2013, 19(1): 105-109

[9] Zhang Y, Xu C, Zhang X, et al. Personalized retrieval of sports video based on multi-modal analysis and user preference acquisition. *Multimedia Tools and Applications* (*MTA*), 2009, 44(2): 305-330

[10] Zhang Y, Zhang X, Xu C, et al. Personalized retrieval of sports video. In: Proceedings of the ACM Multimedia (ACM MM) Workshop, New York, USA, 2007. 313-322

[11] Sathyanareayan V S, Kohlip, Bruhadeshwar B. Signature generation and detection of malware families. In: Information Security and Privacy. Springer Berlin Heidelberg, 2008. 336-349

[12] Zhang X Y, Wang S, Yun X. Bidirectional active learning: a two-way exploration into unlabeled and labeled dataset. *IEEE Transactions on Neural Networks and Learning Systems* (*TNNLS*), 2015, 26(12): 3034-3044

[13] Zhang X, Xu C, Cheng J, et al. Effective annotation and search for video blogs with integration of context and content analysis. *IEEE Transactions on Multimedia* (*TMM*), 2009, 11(2): 272-285

[14] Zhang X Y, Wang S, Zhu X, et al. Update vs. upgrade: modeling with indeterminate multi-class active learning. *Neurocomputing* (*NEUCOM*), 2015, 162: 163-170

[15] Zhang X. Interactive patent classification based on multi-classifier fusion and active learning. *Neurocomputing* (*NEUCOM*), 2014, 127: 200-205

[16] Zhang X Y, Cheng J, Xu C, et al. Multi-view multi-label active learning for image classification. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Cancun, Mexico, 2009. 258-261

[17] Zhang X Y, Cheng J, Lu H, et al. Selective sampling based on dynamic certainty propagation for image retrieval. In: Proceedings of the Advances in Multimedia Modeling (MMM), Kyoto, Japan, 2008. 425-435

[18] Zhang X Y, Cheng J, Lu H, et al. Weighted co-SVM for image retrieval with MVB strategy. In: Proceedings of the IEEE International Conference on Image Processing, San Antonio, USA, 2007. 517-520

[19] Zhang X Y. Dynamic batch selective sampling based on version space analysis. *High Technology Letters*, 2012, 18(2): 208-213

[20] Settles B. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009

[21] Dasgupta S, Langford J. A tutorial on active learning. Tutorial summary: Active learning. In: International Conference of Machine Learning, Montreal, Canada, 2009

[22] Specht D F. Probabilistic Neural Networks. 1990, 3: 109-118

[23] Khandoker A H, Palaniswami M, Karmakar C K. Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings. *IEEE Transaction on Information Technology in Biomedicine*, 2009, 13: 37-48

Malware detection using active learning based on collaborative sampling

Zhang Kai[*,***], Wang Dongan[**], Li Chao[**], Jia Bing[****]

([*] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093)

([**] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029)

([***] University of Chinese Academy of Sciences, Beijing 100049)

([****] Henan Worker's Cultural Palace, Zhengzhou 450007)

Abstract

The malware detection using classification algorithms based on machine learning was studied. In consideration of the fact that current malware recognition mainly uses traditional classification algorithms, thus leading to the application of machine learning models and low classification precision due to the unsufficiency of labelled samples, a new malware detection method using active learning based on collaborative sampling was proposed. The method can use less labelled samples to effectively recognize malware. The experiment showed that it had the higher classification precision and the better performance compared with traditional methods.

Key words: active learning, support vector machine (SVM), probabilistic neural network (PNN), collaborative sampling