朴素贝叶斯增量学习在病毒上报分析中的应用

陈 亮[1] 郑 宁[1] 郭艳华[1] 徐 明[1] 胡永涛[2]
[1](杭州电子科技大学计算机学院 浙江 杭州 310018)
[2](公安部第三研究所 上海 201204)

摘 要 未知类别样本的增量学习中,合理的学习序列能够优化分类器性能,提高分类精度。从优先学习的样本被正确分类概率大小的角度,提出了一种基于样本分类结果可信度的朴素贝叶斯增量学习序列算法。该算法将学习样本分类结果中可信度大的样本优先进行增量学习。在此算法基础上,实现一个病毒上报分析系统用于可疑样本的自动化分析与检测。实验结果表明,经该算法增量学习后的分类器检测效果优于随机增量学习。

关键词 增量学习 学习序列 可信度

APPLYING NAÏVE BAYESIAN INCREMENTAL LEARNING IN VIRUS REPORTING AND ANALYSING

Chen Liang[1] Zheng Ning[1] Guo Yanhua[1] Xu Ming[1] Hu Yongtao[2]
[1](School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China)
[2](The Third Research Institute of the Ministry of Public Security, Shanghai 201204, China)

Abstract In incremental learning of class unknown samples, reasonable learning sequence can optimise the performance and precision of the classifier. In terms of the probability of right classification of preferentially learned samples, in this paper we propose a naive bayesian incremental learning algorithm based on the credibility of samples' classifed result. In the algorithm, those with higher credibility in learning samples' classified result have the priority of incremental learning. A system of virus reporting and analysing system is implemented based on the algorithm for automatic analysis and detection of suspicious samples. Experiment result indicates that the classifier of incrementally learning through the algorithm outperforms the classifer of randomly incremental learning in detection effect.

Keywords Incremental learning Learning sequence Credibility

0 引 言

Win32环境下,反病毒工程师借助逆向工程和调试技术,通过分析可疑程序 API及参数调用情况,依据经验判断上报的可疑样本是否为病毒。随着商业利益的驱动和病毒制作技术的普及,病毒种类和数量都在迅速增加,这给传统的人工分析手段带来了巨大的压力。鉴于此,本文实现一个病毒上报分析系统用于可疑样本的自动化分析和检测。在上报分析检测中,样本数据通常是分批获取,随着新样本的出现,需对分类模型进行增量学习,以对模型参数进行调整。贝叶斯方法由于其充分利用先验知识的特性,使之成为增量学习的自然选择。文献[1,2]从当前分类损失的角度探讨了增量贝叶斯模型中学习序列算法的问题。本文从优先学习的样本被正确分类概率大小的角度对学习序列算法的问题进行了探讨,利用分类器对已知类别样本的检测结果的先验知识,提出了一种基于样本分类结果可信度的朴素贝叶斯增量学习序列算法。

1 病毒上报分析系统模型

系统模块结构如图 1所示。系统由可疑样本上报模块、样本入库模块、自动化行为分析模块、增量学习模块和病毒检测模块组成。可疑样本上报模块为 Internet用户提供可疑样本上传的 Web服务,并将样本运行时的具体行为和检测结果通过邮件以 Xml形式反馈给用户。样本入库模块用于大规模样本的集中入库。自动化行为分析模块为系统核心模块,用于分析样本进程运行时所加载的系统 DLL,并根据事先定义的恶意行为来对特定 API进行监控。增量学习模块用于新样本入库时,检测

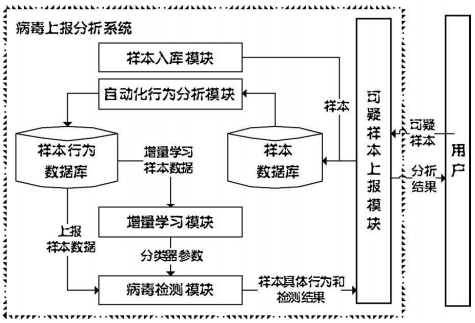

图 1 上报分析系统模块结构

收稿日期:2008-07-18。浙江省自然科学基金(Y106176);浙江省科技厅计划项目(2007C33058);上海市科委基金(06511502)。
陈亮,硕士生,主研领域:反计算机病毒,计算机取证。

模型的学习和分类器参数的调整, 其包括已知样本类型和未知样本类型的增量学习。病毒检测模块对上报样本进行检测并给出检测结果, 用户也可以根据系统返回的样本具体行为, 依据自身经验对样本性质进行判断。

2 病毒特征定义

API作为操作系统为应用程序提供的服务接口, 完成对网络、文件系统和其它重要系统资源的访问。通过跟踪病毒程序的 API调用, 可以掌握其具体的操作行为。文献[3 4]通过 API来描述病毒行为用于病毒的检测, 但仅使用 API并不能准确描述一个恶意行为。同一个 API调用, 其危险性随着其参数的不同而不同。拷贝一个可执行文件到系统敏感目录相对拷贝一个文本文件到用户目录的潜在危险性要高得多。因此本文采用 API与参数相结合的方式来描述恶意行为, 并依据专家经验定义了一个 35维的特征向量, 其每一维代表一种恶意行为事件, 该事件由相应的 API及其参数表示, 仅当程序调用的 API和参数均满足条件时, 我们才认为发生了该行为事件。表 1给出部分行为事件的定义。

表 1　常见病毒行为定义

行为类别	具体行为事件	对应的API调用	参数描述
文件相关类行为	MODIFY_FILEATTR 修改文件属性	SetFileAttributes	参数 dwFileAttributes为隐藏、系统或者只读
	COPY_FILE 复制 PE文件到关键目录	CopyFile CopyFileEx	源文件为 PE文件, 目标路径为系统路径或逻辑分区根目录
	WRITE_FILE向敏感目录下的 PE文件写入数据	WriteFile WriteFileEx	写入文件句柄对应为 PE文件, 目标路径为系统路径或逻辑分区根目录
	DELETE_FILE在关键目录删除文件	DeleteFile	删除的文件所在路径为系统路径或逻辑分区根目录
进程相关类行为	ISDEBUG检测是否处于被调试状态	IsDebuggerPresent CheckRemote DebuggerPresent	任意参数
	是否以 SHELL方式隐蔽启动外部程序	ShellExecute ShellExecuteEx	执行方式是否为 SW_HIDE方式 (隐藏窗口方式)
	CREATE_PROCSS 创建新进程	CreateRemote Thread	任意参数
	⋮	⋮	

3 朴素贝叶斯及其增量推理

朴素贝叶斯分类算法的基础是贝叶斯定理。在处理大规模数据时, 贝叶斯分类算法表现出较高的分类准确性和运算性能, 加之其充分利用先验信息的特性, 使之成为增量学习中的最佳选择模型。

为便于表述, 文中病毒统称为黑样本, 正常程序统称为白样本。对于样本空间的每一个样本可以看作是 n维空间的一个点 $X=(x_1, x_2, \cdots, x_n)$, 其中 $x_1, x_2, \cdots, x_n$ 分别为行为事件属性 $A_1, A_2, \cdots, A_n$ 的取值; $x_n \in \{0, 1\}$, 取值为 1时表示该事件发生, 0则表示未发生。D为样本类别属性, $D=\{d_0, d_1\}$, d_1 表示黑样本, d_0 表示白样本。依据朴素贝叶斯定理, 病毒检测判别式表示为:

$$C(X) = \arg\max_d \left(\frac{P(d)P(x|d)}{P(x)}\right)$$

根据朴素贝叶斯理论各条件属性相互独立的基本假设和 $P(x)$ 在判别式中对具体样本而言为常数, 判别式可简化为:

$$C(X) = \arg\max_d \left(P(d)\prod_{i=1}^{n} P(x_i|d)\right) \quad (1)$$

在无信息 Dirichle共轭先验假设条件下[5], $P(d_i)$ 的计算公式方法为:

$$P(d_i) = \frac{1+count(d_i)}{|D|+|T|} \quad (2)$$

$$P(x_i|d_i) = \frac{1+count(d_i \wedge x_i)}{|A_i|+count(d_i)} \quad (3)$$

其中, $count(d_i)$ 为训练样本中样本类别为 d_i的样本数; $|D|$ 为样本类别个数; $|T|$ 为训练样本总数; $count(d_i \wedge x_i)$ 表示样本类别为 d_i且属性 A_i取值为 x_i的样本数; $|A_i|$ 为属性 A_i的取值个数。

增量学习的任务之一是: 当新增学习样本集 $T'=\{X'_1, X'_2, \cdots, X'_n\}$ 时, 分类器参数的调整。

对于已知类别的学习样本, 根据共轭 Dirichle分布, 得到参数的重新估计为:

$$p'(d_i) = \frac{1+count(d_i)+count'(d_i)}{|D|+|T|+|T'|} \quad (4)$$

$$p'(x_i|d_i) = \frac{1+count(d_i \wedge x_i)+count'(d_i \wedge x_i)}{|A_i|+count(d_i)+count'(d_i)} \quad (5)$$

其中, $count'(d_i)$ 表示新增样本集中样本类别为 d_i的样本数。

对于未知类别的学习样本, 需用现有的分类器对其进行分类, 同时按照一定的学习序列对这些样本进行学习, 每加入一个样本 $X'=(x'_1, x'_2, \cdots, x'_n)$ 时, 其参数的重新估计为:

$$p'(d_i) = \begin{cases} \frac{\delta}{1+\delta} p'(d_i) & C(X') \neq d_i \\ \frac{\delta}{1+\delta} p'(d_i) + \frac{1}{1+\delta} & C(X') = d_i \end{cases} \quad (6)$$

其中 $\delta = |D|+|T|$。

$$p'(x_i|d_i) = \begin{cases} \frac{\zeta}{\zeta+1} p'(x_i|d_i) & C(X') = d_i \,\&\, x_i \neq x'_i \\ \frac{\zeta}{\zeta+1} p'(x_i|d_i) + \frac{1}{\zeta+1} & C(X') = d_i \,\&\, x_i = x'_i \\ p'(x_i|d_i) & C(X') \neq d_i \end{cases} \quad (7)$$

其中 $\zeta = |A_i| + count(d_i)$。

4 基于分类结果可信度的未知类别样本增量学习

基于分类结果可信度的未知类别样本增量学习的基本思想为: 充分利用分类器对已知类别样本检测结果的统计特性, 将分类结果正确概率大的样本优先加入原样本集合, 使得优先学习的样本尽可能为正确分类的样本。其实现过程主要分为两步。

1) 事件数排序

事件数为样本发生不同行为事件的次数。不同事件数下, 样本分类结果与其实际类别一致的概率往往不同, 即分类结果的可信度不同。事件数越大, 样本恶意行为越明显, 样本判定为黑样本且其实际为黑样本的概率也越大, 即样本分类结果为黑的

可信度越大。事件数排序就是根据不同事件数下分类结果可信度的大小，对事件数进行排序，将可信度大的事件数下的样本优先进行学习。文中分别使用 PB_i 和 PW_i 来衡量不同事件数下样本分类结果为黑和白的可信度：

$$PB_i=\frac{count(d_1 \wedge Evn=i)\times TPF_i}{count(d_1 \wedge Evn=i)\times TPF_i+count(d_0 \wedge Evn=i)\times FPF_i} \tag{8}$$

$$PW_i=\frac{count(d_0 \wedge Evn=i)\times TNF_i}{count(d_1 \wedge Evn=i)\times FNF_i+count(d_0 \wedge Evn=i)\times TNF_i} \tag{9}$$

其中 Evn 表示样本发生行为事件的个数；$count(d_1 \wedge Evn=i)$ 为某已知类型样本集合中事件数为 i 的黑样本数；TPF_i 为模型对该集合中事件数为 i 的样本的检出率，表示为该集合事件数为 i 的黑样本中正确检测为黑样本的样本比例；FPF_i 为误报率，表示该集合事件数为 i 的白样本中误报为黑样本的样本比例；TNF_i 为真阴性率，表示该集合事件数为 i 的白样本中正确检测为白样本的样本比例，$TNF_i=1-FPF_i$；FNF_i 为假阴性率，表示该集合事件数为 i 的黑样本中检测为白样本的样本比例，$FNF_i=1-TPF_i$。PB_i 值越大，事件数 $Evn=i$ 下的样本判定为黑的可信度越大；PW_i 值越大，事件数 $Evn=i$ 下的样本判定为白的可信度越大。

以下举例说明如何根据分类器对已知类别样本集分类结果的统计特性对事件数进行排序。样本集 L 按事件数划分为 $L=\{L_0, L_1, \cdots, L_n\}$，$n$ 为事件数，L_n 表示发生事件数为 n 的样本集合。分类器对该集合不同事件数下样本的检出率和误报率如表2所示，为便于计算，不同事件数下的黑白样本个数均取 100。

表 2 一个可信度计算表

事件数 Evn	TPF (%)	FPF (%)	PB_i (%)	PW_i (%)
1	100.00	30.00	76.92	100.00
2	80.00	10.00	88.89	81.82
3	90.00	20.00	81.82	88.89

事件数 $Evn=1$ 时的可信度计算：

$$PB_1=\frac{100\times 100\%}{100\times 100\%+100\times 30\%}\times 100\%=76.92\%$$

$$PW_1=\frac{100\times(1-30\%)}{100\times(1-100\%)+100\times(1-30\%)}\times 100\%=100\%$$

同理可得到其它事件数下样本分类结果的可信度。由表2中的 PB_i、PW_i 分别得到分类结果为黑和白的事件数序列 $<L_2, L_3, L_1>$ 和 $<L_1, L_3, L_2>$。在未知类别样本的增量学习过程中，对于判定结果为黑的样本，按序列 $<L_2, L_3, L_1>$ 进行学习，即事件数为 2的样本优先进行学习；对于判定结果为白的样本，按序列 $<L_1, L_3, L_2>$ 进行学习，即事件数为 1的样本优先学习。

2）同一事件数下的样本排序

同一事件数下的样本按照 PD值降序排序，PD值大的样本将优先进行学习。

$$PD(X)=\begin{cases}\dfrac{P(d_0)\prod_{i=1}^{n}P(x_i|d_0)}{P(d_1)\prod_{i=1}^{n}P(x_i|d_1)} & C(X)=d_0\\ \dfrac{P(d_1)\prod_{i=1}^{n}P(x_i|d_1)}{P(d_0)\prod_{i=1}^{n}P(x_i|d_0)} & C(X)=d_1\end{cases} \tag{10}$$

可信度增量学习算法描述如下：

输入：训练集 $T=\{<X_1, C_1>, <X_2, C_2>, \cdots, <X_n, C_n>\}$，其中 C_n 是 X_n 的类别；学习集 $L=\{X'_1, X'_2, \cdots, X'_n\}$。

输出：分类器 C。

（1）在训练集 T的基础上生成分类器 C；

（2）对训练集 T进行分类和统计，生成黑白样本事件数序列 $E\_blist=<Evn_b, Evn_g, \cdots, Evn_b>$ $E\_wlist=<Evn_w, Evn_t, \cdots, Evn_w>$；

（3）若 $L=\phi$，返回，算法结束，否则继续；

（4）利用公式（1）对测试集 L进行分类，获取类别标签；

（5）根据 E_blist、E_wlist 和 $PD(X_n)$ 对（4）中分类结果进行排序，生成黑白样本加入序列：

$$blist=<X'_b, X'_g\cdots, X'_b>$$

$$wlist=<X'_w, X'_t\cdots, X'_w>;$$

6）将学习样本 blist[1] 和 wlist[1] 加入训练集 T，$T=T+\{<blist[1], C(blist[1])>, <wlist[1], C(wlist[1])>\}$，$L=L-\{blist[1], wlist[1]\}$，并依据公式（6）和（7）更新分类器参数，转向（2）。

5 实验设计及结果

5.1 数据集描述

文中所有实验样本均为 PE结构的可执行程序。实验中黑白样本均由哈尔滨安天实验室提供。白样本选择上报可疑样本中最终判定为正常可执行程序的样本，这类样本大多因具有疑似病毒行为特征而被上报，但实际上仍然属于合法的正常程序。用这类白样本建模，更加符合上报分析的真实环境，使得检测模型更具实用性。

为保证数据建模的合理性，对原始黑白样本进行以下处理：

1）依据病毒命名规则去掉黑样本中病毒的近似变种，以避免黑样本中存在过多来自同一病毒的变种，使得训练后的分类器带有倾向性。

2）依据 MD5值去掉白样本中的重复样本。

另从初步的实验结果来看，对于发生事件数 1和 2的黑白样本，由于样本事件数较少且发生的事件相似，模型对其区分效果不佳。因此本文仅对发生事件数大于 2的样本进行实验。最终获取的试验数据集由 1854个黑样本（恶意可执行程序）和 1510个白样本（正常可执行程序）组成。

5.2 实验结果

黑白样本中分别随机选取 1460和 1189个样本作为初始训练集 T，余下 394和 321个样本看作无标签的增量学习集 L。表3为随机增量学习和可信度增量学习后的检测模型对初始测试集和初始训练集的检测效果。

表 3 检测效果

检测对象	随机增量学习			可信度增量学习		
	正确率 (%)	检出率 (%)	误报率 (%)	正确率 (%)	检出率 (%)	误报率 (%)
初始训练集 T	67.98	88.58	57.32	76.65	83.25	31.46
初始测试集 L	69.38	87.95	53.41	77.56	83.90	30.22

从表3中可以看出，基于可信度增量学习后的检测模型对

初始测试集和初始训练集的检测正确率均有所提高, 虽检出率有所下降, 但误报率有着较为明显的改善。

表 4为基于两种方法学习样本选入正确率的比较。学习集整体选入正确率为原始学习集 L中被正确标记选入的样本所占比例。从表中可以看出, 基于可信度增量学习的黑白样本选入正确率和总体选入正确率较随机增量学习均有所提高。

表 4　学习集 L选入正确率

学习算法类型	L中黑样本选入正确率 (%)	L中白样本选入正确率 (%)	整体选入正确率 (%)
随机增量学习	78.43	73.83	76.36
可信度增量学习	81.55	79.30	80.56

6 结　论

对于未知类别样本的增量学习, 其主要工作是寻找样本加入原有样本集合的学习序列。本文从优先学习的样本被正确分类并选入的角度对学习序列算法进行了探讨, 利用分类器对已知类别样本的检测结果的先验知识, 提出了一种基于样本分类结果可信度的朴素贝叶斯增量学习序列算法, 使得优先学习的样本尽可能为正确分类的样本。实验结果表明, 经该算法增量学习后的分类器检测效果优于随机增量学习, 另学习样本的选入正确率也优于随机增量学习。本文今后的工作将关注以下两个方面: 增加特征向量维数, 定义更多在病毒和正常程序中具有区分能力的行为特征, 使得程序在理论上能被捕获到更多的事件, 增加模型可检测样本范围; 另一方面, 未知样本增量学习中, 如何控制原有分类器误差的累积和传递也是今后研究的方向。

参 考 文 献

[1] 宫秀军, 刘少辉, 史忠植. 一种增量贝叶斯分类模型[J]. 计算机学报, 2002, 25(6): 645-650.

[2] 姜卯生, 王浩, 姚宏亮. 朴素贝叶斯分类器增量学习序列算法研究[J]. 计算机工程与应用, 2004(14): 57-59.

[3] Xu J Y, Sung A H, Chavez P, et al. Polymorphic Malicious Executable Scanner by API Sequence Analysis[C]//Proceedings of the Fourth International Conference on Hybrid Intelligent Systems. IEEE Computer Society, 2004: 378-383.

[4] 张波云, 殷建平, 蒿敬波, 等. 基于多重朴素贝叶斯算法的未知病毒检测[J]. 计算机工程, 2006, 32(10): 18-21.

[5] Langley P, Sage S. Induction of selective Bayesian classifiers[C]//Proceedings of the tenth Conference on Uncertainty in Artificial Intelligence. Seattle, WA: Morgan Kaufmann, 1994: 399-406.

(上接第 75页)

效益。大多数情况下 TDGA优于 TDMin-Min的平均信任效益。TDGA平均信任效益值大说明这种调度算法任务映射后执行越稳定, 执行结果更为可信。

从图 6中还可以看出当具有强信任关系的任务与资源增多时, 三种算法的平均信任效益都增大。随着任务数的增多和信任关系的增强, TDGA和 TDMin-Min的差距拉大, 说明随着任务数的增多和信任关系的增强, 基于信任机制的遗传算法在信任效益方面的优势更为显著。

(2) 任务完成率

任务完成率是指调度过程中能够成功完成的任务数的比率。任务完成率高的调度算法能够为用户提供更稳定的服务, 能够提高资源利用的效率。任务完成率高说明调度过程中成功匹配的任务数多, 信任关系无法满足而被迫丢弃的任务数目少。

从图 7中可以看出 TDGA在文中列出的情况下, 都能获得最高的任务完成率。可以看出随着任务数的增加和任务与资源之间信任关系的增强, 基于信任机制的遗传调度算法在任务完成率方面的优越性更明显。

由以上实验数据以及分析可以看出, 基于信任机制的遗传调度算法可以提高任务的平均信任效益值, 使得任务可靠性增强, 同时也使得更多的任务和资源合理匹配, 提高资源利用的效率。因此, 本文提出的这种算法是有效的。

6 总　结

本文提出了基于信任机制的渲染网格作业调度模型, 并对网格调度算法—遗传算法加以改进, 将信任机制考虑进去, 对渲染网格作业进行调度。与没有加入信任机制的遗传算法和基于信任机制的 Min-Min算法进行比较, 通过实验证明这种算法在提高信任效益和任务完成率等方面是有效的。实验显示改进后的基于信任机制的调度算法可以更加合理地分配资源给相应的任务, 提高平均信任效益和资源的利用率以及任务完成率, 更好地为用户服务。

本项研究进一步要做的工作是把身份信任也考虑进去, 寻找一种更好的方法来确定节点间的信任关系并对信任表进行维护和动态更新。同时找到一个合适的衰减函数来度量时间变化对信任值产生的影响。另外也可以考虑把经济学加入信任机制, 使资源和任务更有效的匹配。

参 考 文 献

[1] Min-You Wu, Wei Shu, Zhang H. Segmented min-min: a static mapping algorithm for meta-tasks on heterogeneous computing systems. Heterogeneous Computing Workshop, 2002 (HCW 2002) Proceedings. 9th, 2002: 375-385.

[2] 林剑柠, 吴慧中. 基于信任模型的数据密集型网格任务调度算法研究[J]. 信息与控制, 2006, 35(5): 667-672.

[3] 袁禄来, 曾国荪, 等. 网格环境下基于信任模型的动态级调度[J]. 计算机学报, 2006, 29(7): 1217-1224.

[4] 张伟哲, 刘欣然, 云晓春, 等. 信任驱动的网格作业调度算法[J]. 通信学报, 2006, 27(2): 73-79.

[5] 张伟哲, 方滨兴, 胡铭曾. 基于信任 QoS增强的网格服务调度算法[J]. 计算机学报, 2006, 29(7): 1157-1166.

[6] Aggarwal M, Kent R D, Ngom A. Genetic Algorithm Based Scheduler for Computational Grids[C]//Proc of the 19th Annual International Symposium on High Performance Computing Systems and Applications (HPCS'05): 209-215, Guelph, Ontario, Canada, May 2005.

[7] Farag Azzedin, Muthucumaru Maheswaran. Towards Trust-Aware Resource Management in Grid Computing System. Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, 2002.

[8] Zhang Weizhe, Liu Xinran, Yun Xiaochun, et al. Trust-driven job scheduling eurisitics for computing grid. Journal of Communications, 2006, 27(2): 73-79 (in Chinese).

[9] Huang Baobian, Zeng Wenhua. Trust mechanism-based dynamic task scheduling in grid Computing. Computer Applicaions, 2006.