基于主动学习的计算机病毒检测方法研究*

张 勇[1] 张卫民[1] 欧庆于[2]
(61705 部队[1] 北京 100091)(海军工程大学信息安全系[2] 武汉 430033)

摘 要 针对传统病毒检测方法存在的更新速度慢、对未知病毒检测能力不足等问题,该文对主动学习理论在计算机病毒检测方面的应用进行了研究,提出了一种基于支持向量机主动学习的计算机病毒检测模型结构。此外,为了改进病毒检测的精度问题及主动学习过程的效率,利用相关 n-gram 方法实现了对样本文件的特征提取,并结合信任度测量理论实现了基于非确定抽样的询问功能。实验表明,该模型针对未知病毒具有较高的检测精度,并且能够极大地缩减训练时间及对训练数据的数量要求,提高系统的学习效率。

关键词 病毒检测;主动学习;支持向量机;非确定性抽样
中图分类号 TP309.7

Research on Computer Viruses Detection Approaches Based on Active Learning

Zhang Yong[1] Zhang Weimin[1] Ou Qingyu[2]
(No. 61705 Toops of PLA[1], Beijing 100091)
(Depart. of Information Security, Naval University of Engineering[2], Wuhan 430033)

Abstract Traditional computer viruses detection approaches update slowly and have poor ability in detecting unknown viruses. In this paper, the application of the active learning theory in computer viruses detection is studied, and a computer viruses detection model based on active learning of the support vector machine is proposed. Moreover, to improve the precision of the virus detection and the efficiency of the active learning process, feature extraction of the sampling files are realized by using the method of relevant n-gram, and combined with the trust measurement theory, query function based on the uncertainty based sampling is also realized. Experiments' results show that the model has very good detection precision against unknown computer viruses and can greatly shorten the training time and reduce the requirements of the training data and improve the learning efficiency of the system.

Key Words computer viruses detection, active learning, support vector machine, uncertainty based sampling
Class Number TP309.7

1 引言

自计算机病毒在 1986 年首次出现以来,其数量呈几何级数迅速增长,对计算机系统的安全构成严重威胁。Gartner[1]在其报告中将计算机病毒列为计算机系统的头号安全威胁,对于计算机病毒的检测已成为信息安全领域的热点研究方向。

当前,大多数反病毒工具利用病毒特征码匹配及启发式分类器(heuristic classifier)进行病毒检测。通过使用特征码作为病毒标签,并利用已知病毒的特征码产生检测模型,从而能够以非常低的错误概率对已知病毒进行有效的检测。然而,对于未知病毒而言,由于其特征码并未被加入到病毒特征库中,使得基于病毒特征码的方法不能够及时、有

* **收稿日期**:2011 年 5 月 11 日,**修回日期**:2011 年 6 月 13 日
基金项目:国家自然科学基金项目(编号:60774029)资助。
作者简介:张勇,男,工程师,研究方向:信息安全。张卫民,男,高级工程师,研究方向:信息安全。欧庆于,男,讲师,研究方向:信息安全。

效的对病毒进行检测,从而为系统带来极大的安全威胁。此外,很多病毒还能够通过变形、加密及控制流混淆等手段掩盖其特征码,进一步增加了病毒检测的难度[2]。

为了弥补传统反病毒工具所存在的不足,Witten[3]等人提出了基于机器学习手段的计算机病毒检测方法。在基于机器学习的病毒检测方法中,病毒及正常文件均由从其二进制可执行代码中所提取的特性向量所表示,并作为训练集对分类器进行训练。在病毒侦测过程中,利用分类器的分类能力,未知文件将被分类为病毒或正常文件。此外,为了保持分类的精确度,必须持续的利用新文件(病毒或正常文件)对分类器进行训练,并对训练集进行更新。然而,当一个文件被分类后,由于传统的基于机器学习的病毒检测方法采用的是随机学习方式(random learning),分类器本身并不能判断新文件是否能够作为一个新的样本被加入到训练集中。此外,为了将一个新样本加入至训练集中,需要人工的方式对其进行标签(病毒或正常)操作。而对一个未知样本进行标签操作,就必须依赖于人工方式对其进行深入分析,从而造成大量的时间损耗。因此,对训练集的自动更新及标签操作成为影响基于机器学习的病毒检测方法效率及可实施性的关键因素。

针对传统的基于机器学习的病毒检测方法中所存在的不足,本文对主动学习在计算机病毒检测方面的应用进行了研究,并在此基础之上提出了一种基于主动学习的计算机病毒检测模型。实验表明,通过主动学习方式的运用,能够在保持分类精度的同时显著降低需进行标签操作的训练实例的数量,并在很大程度上训练集能进行自动更新操作,从而极大地减少了病毒检测过程中人工参与的比例,有利于基于机器学习的病毒检测方法在实际应用中的推广。

2 主动学习

机器学习的主要目的是从有限的数据中提取通用模式[4]。目前,机器学习方法主要分为两大类:监督学习(supervised learning)及无监督学习(unsupervised learning)。其中,监督学习主要是让计算机从样例中学习输入到输出的函数对应关系,以预测输入目标的某些附加属性,如给出一个人的身高,预测其体重;无监督学习,其数据不包含输出值,学习的任务是理解数据产生的过程,这种类型的学习包括密度估计、分布类型的学习和聚类等。显然,针对病毒的检测应该属于监督学习的范畴。

无论是监督学习还是无监督学习,在传统的机器学习中,通常依据随机抽样的方式收集大量训练集样本,并根据这些样本推导分类器或模型,这种学习方式又被称为"被动学习(passive learning)"。

与被动学习方式不同,在主动学习中样本的收集不采用随机抽样方式,而是通过询问(querying)函数进行询问,并基于所接收的响应从未标签样本池中获取。在主动学习系统中,训练数据被分为被标签样本集 TR 及未标签样本池 U,并包含一个学习器 L 以及一个询问模块 q。其中,学习器 L 在被标签样例上进行训练,询问模块 q 决定未标签样本池 U 中的哪些样本将被选择进行标签操作,并被加入到被标签样本集 TR 中。被更新后的被标签样本集 TR 继续用来对学习器 L 进行训练。

3 模型结构

基于主动学习的计算机病毒检测模型如图1所示。模型由四部分构成,分别是:样本构造器、未标签样本池、训练集及主动学习器。模型对输入数据进行分类操作,以判断当前的文件是否为病毒文件。

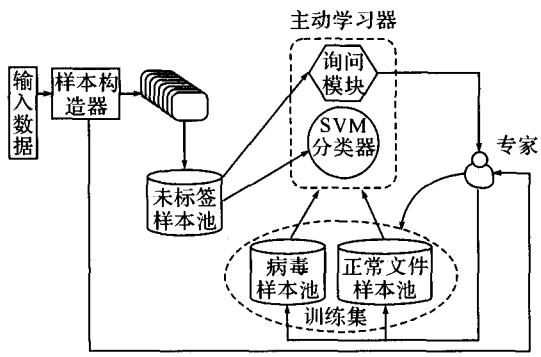

图1 基于主动学习的计算机病毒检测模型

其中,样本构造器负责对输入数据发生器所产生的输入数据进行特征提取操作,以生成分类器能够对其进行分类操作的向量数据,样本构造器生成的数据类型与训练集中的数据类型相同。总的来说,样本构造器所生成的特征数据应满足以下三点要求:

1) 能够反映输入数据的显性特征;

2) 能够反映输入数据的隐性特征;

3) 便于计算机处理。

训练集由病毒样本池及正常文件样本池组成,

分别包含了多个病毒文件及正常文件的特征数据。训练集中的原始数据由样本构造器生成，并通过专家进行人工标签操作最终构成病毒样本及正常文件样本。当训练集建立后，样本构造器产生的数据存储至未标签样本池中。

主动学习器由 SVM(支持向量机)分类器以及询问模块两部分组成，系统利用训练集病毒样本池与正常文件样本池中的训练样本对 SVM 分类器进行训练，并对为标签样本池中的样本进行分类操作。

询问模块通过特定的选取规则对未标签样本池中的样本进行选取，并将该样本发送给专家，通过人工方式对该样本进行分类，并最终将该样本更新至训练集中，实现模型的主动学习功能。询问模块对未标签样本的选取基于 pool-based[5] 方式进行设计。当训练集被更新后，询问模块根据选取规则从未标签样本池中选取样本，并将该样本送入 SVM 分类器中进行分类。SVM 对该样本分类完成后，重新更新训练集，询问模块再根据更新后的训练集及选取规则选取未标签样本。

4　样本的表征方式

对于二进制可执行文件而言，以字节序列的方式表征文件能够较好地保留其可执行的特性。基于字节序列的这种优势，Kephart[7] 首先提出了一种利用神经网络侦测引导段病毒的病毒检测方法。该方法以引导段中病毒代码的所有字节序列作为输入，并利用 ANN 分类器对其进行分类操作，从而实现了对未知病毒的检测功能。此后，Arnold[8] 及 Schultz[10] 等人也分别基于字节序列的表征方式提出了相应的病毒检测方案。然而，基于字节序列的表征方式只能够表示文件的显性特征，而对于文件的隐性特征，如：代码风格、文件编制所使用的开发工具及编译器等信息就显得无能为力了。而这些隐性特征对于未知病毒的成功侦测往往是至关重要的。针对这一问题，Abou-Assaleh[11] 提出了以对字节序列进行 n-gram 提取操作后获得的数据作为特征数据，并针对这些特征数据进行病毒检测的方法。

在我们的模型设计中，所有的样本，包括训练集中的病毒样本、正常文件样本，以及未标签样本池中的未标签样本均通过对输入数据进行关联 n-grams[12](relevant n-grams)提取操作后生成。关联 n-gram 的定义如下所示：

定义 1　D 为训练集，V、B 分别为训练集中的病毒子集及正常文件子集。

定义 2　对于 *n-gram* Ng，当 Ng 出现在程序 P 中时，与程序 P 相关的 Ng 文件频率 $\delta(Ng,P)$ 为 1；否则，为 0。而与类 C 相关的 Ng 文件频率 $\delta(Ng,C)=\sum_{P\in C}\delta(Ng,P)$，即与类 C 相关的 Ng 文件频率为类 C 中包含 Ng 的程序数量。

定义 3　设 Ng^t 为属于类 D 的程序 t 的 *n-gram* 总和，则与类 D 相关的 *n-gram* 总和 $Ng(D)$ 为 $\bigcup_{t\in D}Ng^2$。对于病毒子集 V 以及正常文件子集 B，假设 $Ng(T)$ 以与类相关的 Ng 文件频率的降序排列，则可得到两个按降序排列的序列，分别为 $\delta(Ng_i,V)$ 及 $\delta(Ng_i,B)$。设 V_r、B_r 分别为序列 $\delta(Ng_i,V)$ 及 $\delta(Ng_i,B)$ 中的前 $k=L/2$ 个元素，则定义 $Ng_k(D)=V_r\cup B_r$ 为类 $D=V\cup B$ 的相关 *n-grams*。

与普通 *n-gram*[11](CNG)不同，在相关 *n-gram* 中使用与类相关的文件频率，使得可以对每个类进行独立的分析。此外，由于在 *n-gram* 中针对单个文件进行 *n-gram* 操作，其产生的数据量远小于针对整个训练集进行 *n-gram* 操作时产生的数量，从而减小了系统的存储压力。

一旦完成相关 *n-gram* 的选取，则可利用信息检索的向量空间模型，以 $Ng_k(D)$ 表示训练集 D。此时，训练集 D 中的程序以向量 $t_1,t_2,\cdots,t_m$ 的形式表示。其中，$t_i(1\leq i\leq m)$ 为二进制 0 或 1，表示 $Ng_k(D)$ 中的第 i 个 *n-gram* 在该程序中一次或多次的出现。最终，训练集中的样本可利用来自于病毒子集和正常文件子集的标签向量集生成。

5　病毒的分类方法

对病毒的侦测过程，实际上就是对文件在病毒及正常文件两个类之间的分类问题。在我们的模型中，分类功能基于支持向量机(SVM)实现。支持向量机作为一种基于统计学习理论的机器学习方法，其处理能力可通过结构风险最小化原则进行改进。此外，由于支持向量机属于凸规划问题，其局部优化也就是对其的全局优化。

假定训练集 $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in(R^n\times Y)^l$，其中 $X_i\in R^n$，$y_i\in Y=\{1,-1\}$，$i=1,\cdots,l$，可以被一个超平面 $(w\cdot x)-b=0$ 分开。如果这个向量集合没有被超平面错误地分开，并且离超平面最近的向量与超平面之间的距离是最大的，则说这

个向量集合被最优超平面分开[13]。Vapnik[13]等人的研究表明,最优超平面具有以下特性:

1) 对最优超平面,系数 a_i^0 必须满足约束

$$\sum_{i=1}^{l} a_i^0 y_i = 0, a_i^0 \geq 0, i=1,\cdots,l;$$

2) 最优超平面是训练集中的向量的线性组合:$w_0 = \sum_{i=1}^{l} y_i a_i^0 x_i, a_i^0 \geq 0, i=1,\cdots,l$;

3) 只有所谓的支持向量可以在 w_0 的展开中具有非零的系数 a_i^0。

根据 Kunn-Tucker 条件可知,最优超平面的充分必要条件为,分类超平面满足条件 $a_i^0\{[(x_i \cdot w_0)-b_0]y_i-1\}=0, i=1,\cdots,l$。最终,最优超平面服从以下约束:

$$\max W(a) = \sum_{i=1}^{l} a - \frac{1}{2} \sum a_i y_i a_j y_j K(x_i, x_j)$$

$$\sum_{i=1}^{l} a_i y_i = 0, a_i \in [0, c], i=1,\cdots,l$$

基于最优超平面的分类规则由指示函数 $f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} a_i y_i K(x_i, x) + b\right)$ 确定。

6 基于非确定性抽样的询问模块

为了使分类器达到尽可能高的病毒检测精度,除训练集外还需要利用未标签样本池中的数据对系统进行持续的训练。而在主动学习系统中,利用分类预测不确定性高的未标签样本对训练集进行更新,往往能够极大地提高系统的训练效率[12]。为了能够充分反映被测样本的分类预测不确定性,在我们的模型中采用了基于信心指数及信任度测量的受控选取准则。

当询问模块需对未标签样本池中的样本进行选取时,首先分别对训练集 $D(D=V \cup B)$ 中的各个样本 x_i 与病毒样本池 V 中的各个样本 v_i 及正常文件样本池 B 中的各个样本 b_i 的欧几里得距离 d_E 进行计算,如式(1)所示:

$$d_E(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \tag{1}$$

其中,$x=\{x_1, x_2, \cdots, x_n\}$ 为被测签样本,$y=\{y_1, y_2, \cdots, y_n\}$ 为病毒样本池或正常文件样本池中的训练样本。当对训练集中各个样本的距离 d_E 计算完成后,选取与病毒样本池的最短距离 d_i^v 及与正常文件池的最短距离 d_i^b。然后,按照式(2)、(3),分别基于病毒样本池及正常文件样本池对训练集中的各个样本的陌生度[14]进行测量。在式(2)、(3)中,当被测样本与预期分类的距离(分子)增大或与非预期分类的距离(分母)减小时,其陌生度都将减小。因此,该方法是一种对单点陌生度的自然测量方法[15]。

$$a_{iv} = d_i^v / d_i^b \tag{2}$$

$$a_{ib} = d_i^b / d_i^v \tag{3}$$

当完成了对训练集中各样本的陌生度测量后,利用式(4)、(5),分别基于病毒样本池 V 及正常文件样本池 B,对未标签样本池中各个样本的 p 值[16]进行计算。其中,# 表示集的势,通过计算有限集中的元素数量获得。a_{new} 表示当前参与计算的未标签样本。

$$p_v(a_{new}) = \frac{\#\{1: a_{iv} \geq a_{new}\}}{n+1} \tag{4}$$

$$p_b(a_{new}) = \frac{\#\{1: a_{ib} \geq a_{new}\}}{n+1} \tag{5}$$

对于各个未标签样本的 p_v 及 p_b,可能存在以下4种情况:

1) p_v 非常大且 p_b 非常小。表明该被测样本属于病毒样本池具有较高的可信任度及信心指数;

2) p_v 及 p_b 都非常大。表明该被测样本属于病毒样本池具有较高的可信任度,但具有较低的信心指数;

3) p_v 及 p_b 都非常小。表明该被测样本属于病毒样本池具有较低的可信任度,但具有较高的信心指数;

4) p_v 非常小且 p_b 非常大。表明该被测样本属于正常文件样本池具有较高的信任度及较高的信心指数。

考虑以上四种情况,其中情况 a 及 d 明确地支持对被测样本的某种分类预测,而情况 b 及 c 则将造成分类预测的不确定性。此外,当 $p_v \approx p_b$ 时,同样也将造成分类预测的不确定性。为了对分类预测的不确定性进行测量,利用 p_v 与 p_b 之间逼近程度的概念来对进行其不确定性[17]。p_v 与 p_b 之间的逼近程度利用式(6)进行计算。

$$C(i) = |p_v - p_b| \tag{6}$$

$C(i)$ 越接近于0,表明对该被测样本进行的分类预测具有越大的不确定性。通过为 $C(i)$ 设置一阈值 &,当 $C(i)<\&$ 时将该被测样本提交给专家进行手工分类,并将分类后的样本更新至响应的分类集(病毒样本池或正常文件样本池)中,从而为训练集的结构提供新的信息。

7 实验与分析

为了验证模型对于病毒检测的有效性，我们利用从 VX Heavens 所获取的病毒样本及由 Windows 操作系统 system32 文件夹中所提取的正常文件样本作为实验基准数据集。所有的病毒样本及正常文件样本均为 Windows PE 格式，并被转换为 ASCII 格式的十六进制代码的形式作为模型训练集。此外，再从 VX Heavens 及 Windows 操作系统 system32 文件夹中抽取样本作为未标签数据集。

7.1 病毒检测能力测试

为了对模型的病毒检测能力进行评估，我们通过从病毒检出率（*TP*）及误报率（*FP*）两个方面与传统支持向量机方式、神经网络方式及 KNN 算法的病毒检测能力进行了比较。其中，*TP* 定义为被模型正确分类的未标签样本池中的病毒样本数量；*FP* 定义为被模型错误分类的未标签样本池中的正常文件样本数量。

对于传统支持向量机方式，我们选用 C-SVC 支持向量机算法，并选择半径基准函数（radius basis function）作为其核函数；对于神经网络方式，选用后向传播算法；对于 KNN 算法，设置 k 为 5，并使用线性最相邻搜索算法。测试分别基于正常训练集及小训练集进行，正常训练集包含 250 个病毒样本及 250 个正常文件样本，小训练集包含 56 个病毒样本及 56 个正常文件样本。针对不同的训练集，不同检测方式的运行结果如表 1、2 所示，从中可以看出采用主动学习方式的病毒检测模型的病毒检出率要高于其它的病毒检测方式，并且在训练样本数量较小的情况下，其误报率并没有显著增加，从而说明其病毒检测能力要优于其它的病毒检测方式。

表 1　基于正常训练集的病毒检测试验结果

病毒检测方法	*TP*(100%)	*FP*(100%)
传统 SVM 方式	99.3	0.8
神经网络方式	99.5	0.6
KNN 算法	99.1	1.3
基于主动学习的病毒检测模型	99.7	0.1

表 2　基于小训练集的病毒检测试验结果

病毒检测方法	*TP*(100%)	*FP*(100%)
传统 SVM 方式	98.4	2.7
神经网络方式	98.1	2.2
KNN 算法	97.3	4.8
基于主动学习的病毒检测模型	99.3	0.15

7.2 主动学习能力测试

为了对模型的主动学习能力进行测试，基于随机选取方式，分别选取 50 个病毒样本及 50 个正常文件样本组成初始训练集。同样，基于随机选取的方式，分别选取 250 个病毒样本及 250 个正常文件样本装入未标签样本池。在此基础之上，对模型中所采用的非确定性抽样询问方法及随机抽样方法对模型检测精度改进方面的影响进行对比。对比结果如图 2 所示，通过对比可以发现，当训练次数达到约 60 个时，采用非确定性抽样询问方式实现主动学习的病毒检测模型的病毒检测精度快速上升到 99.5%。而采用随机抽样选取方式时，当训练次数达到为标签样本池极限—即 500 时，其病毒检测精度只能达到约 70%，而要达到与采用非确定性抽样询问方式时相同的检测精度（99.5%），预计至少需要进行 2000 次训练。

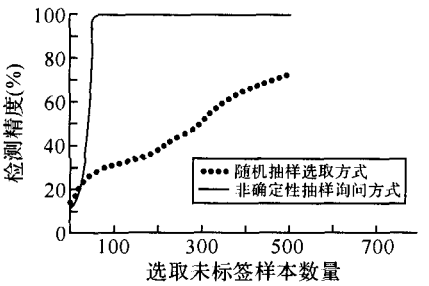

图 2　与随机抽样方式对比

以上结果表明，基于非确定性抽样询问的主动学习方式在缩短训练时间、提高训练效率等方面具有较大优势，能够基于较少的未标签训练数据，使得系统在短时间获得较高的病毒检测精度。

8 结语

本文针对目前病毒检测领域所存在的问题，提出了一种基于主动学习的计算机病毒检测模型。实验结果表明，通过相关 n-gram 特性提取方法及基于非确定性抽样询问的主动学习方式的运用，该模型在病毒检测精度、学习效率等方面与传统病毒检测方法相比具有明显的优势，在具体实现时能够较好地克服病毒演化速度与病毒检测系统更新速度之间的矛盾，从而提高对未知病毒的检测能力。

参 考 文 献

[1] Gartner Inc [EB/OL]. http://www.gartner.com/press_releases/asset_129199_11.html,2005

[2] Christodorescu, M., Jha, S. Static analysis of executables to detect malicious patterns[C]//Proceedings of the 12th USENIX Security Symposium (Security'03), USENIX Association, USENIX Association, 2003:169～186

（下转第 105 页）

多种匹配策略和算法；讨论了基于模糊综合评判的相似度评判的原理、方法、特点、适应性及其计算流程。实验结果表明，本文方法的性能比传统的多策略映射方法有所改进。

未来的工作包括：1）探索解决复杂匹配问题的方法，比如解决一对多和多对多匹配等复杂情况下的模式匹配问题；2）对多策略模式匹配模型作进一步的功能扩展。

参考文献

[1] 杨先娣，彭智勇，刘君强，等. 信息集成研究综述[J]. 计算机科学，2006，33(7)：55～59

[2] Melnik S, Molina-Garcia H, Rahm E. Similarity flooding: A versatile graph matching algorithm[C]//Proceedings of the 18th International Conference on Data Engineering, San Jose, Califomia, USA, 2002：117～128

[3] [EB/OL]. http://wordnet. princeton. edu/wordnet/

[4] Shvaiko P, Euzenat J. A survey of schema-based matching approaches[J]. Journal on Data Semantics Ⅳ, 2005, 3730：146～171

[5] Jaewoo kang, Jeffrey F. Naughton. On schema matching with opaque column names and data values[C]// Proceeding of the 2003 ACM SIGMOD International Conference on Management of Data 2003, San Diego, California, June 09-12, 2003：205～216

[6] Xuan Zhou, Julien Gaugaz, Wolf-Tilo Balke, et al. Query Relaxation Using Malleable Schemas[C]//Proceeding of SIGMOD, 2007：545～556

[7] 杨纶标，高英仪. 模糊数学[M]. 广州：华南理工大学出版社，2002：55～66

[8] 谢季坚，刘承平. 模糊数学方法及其应用[M]. 武汉：华中科技大学出版社，2005：31～36

[9] 王新富，林通，夏有峰. 基于模糊综合评判的作战决策方案满意度评价[J]. 舰船电子工程，2009，29(9)

[10] 郑昌，董文洪，牛庆功，等. 基于 AHP 和模糊综合评判的无人机效能评估[J]. 舰船电子工程，2009，29(6)

[11] 李欣浩，周春雷，李辉. 基于模糊综合评判的舰艇编队网络效能评估[J]. 舰船电子工程，2009，29(10)

[12] 李洪兴，汪培庄. 模糊数学[M]. 北京：国防工业出版社，1994：101～106

[13] [EB/OL]. http://metaquerier. cs. uiuc. edu/repository/datasets/bamm/

（上接第 93 页）

[3] Witten, I., Frank, E. Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco, 2000

[4] 徐琴珍，杨绿溪. 一种基于有监督局部决策分层支持向量机的异常检测方法[J]. 电子与信息学报，2010，32(10)

[5] Lewis, D., & Gale, W. A sequential algorithm for training text classifiers[C]//Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Springer-Verlag, 1994：3～12

[6] 宋玉珍，刘炼，曲付勇. 利用 Hopfield 神经网络解决 TSP 问题[J]. 舰船电子工程，2010，30(4)

[7] Kephart, J. O., Sorkin, G. B., Arnold, W. C., et al. Biologically inspired defenses against computer viruses [C]//Proceedings of the 14th IJCAI, Montreal, 1995,：985～996

[8] Arnold, W., Tesauro, G. Automatically generated win32 heuristic virus detection[C]//Proceedings of the 2000 International Virus Bulletin Conference, 2000

[9] 肖晶，吴学智. 一种基于神经网络的故障诊断新方法研究[J]. 舰船电子工程，2010，30(1)

[10] Schultz, M. G., Eskin, E., Zadok, E., et al. Data mining methods for detection of new malicious executables[C]//SP'01: Proceedings of the 2001 IEEE Symposium on Security and Privacy, IEEE Computer Society, Washington, 2001, 38

[11] Abou-Assaleh, T., Cercone, N., Keselj, V., et al. Detection new malicious code using n-grams signatures [C]//PST, 2004：193～196

[12] D Krishna Sandeep Reddy, Arun K Pujari: N-gram analysis for computer virus diction. Springer-Verlag France, 2006

[13] Vladimir N. Vapnik. 统计学习理论的本质[M]. 张学工，译. 北京：清华大学出版社，2000，91

[14] Tong S. Active learning: theory and applications. PhD thesis, Stanford University, California, 2001, 8

[15] Proedru K, Nouretdinov I, Vovk V, et al. Transductive confidence machine for pattern recognition[C]// Proceedings of the 13th European conference on machine learning, Heidelberg, Germany, 2002：381～90

[16] Barbara D, Carlotta D, James PR. Detecting outliers using transduction and statistical testing[C]//Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, New York, USA, 2006：55～64

[17] Ho SS, Wechsler H. Transductive confidence machine for active learning[C]//Proceedings of the IEEE joint conference on neural networks, Portland, USA, 2003：20～4