基于主动学习和 TCM-KNN 方法的有指导入侵检测技术

李 洋[1),2)] 方滨兴[1)] 郭 莉[1)] 田志宏[1)]

1)(中国科学院计算技术研究所 北京 100080)

2)(中国科学院研究生院 北京 100039)

摘 要 有指导网络入侵检测技术是网络安全领域研究的热点和难点内容,但目前仍然存在着对建立检测模型的数据要求过高、训练数据的标记需要依赖领域专家以及因此而导致的工作量及难度过大和实用性不强等问题,而当前的研究工作很少涉及到这些问题的解决办法.基于 TCM-KNN 数据挖掘算法,提出了一种有指导入侵检测的新方法,并且采用主动学习的方法,选择使用少量高质量的训练样本进行建模从而高效地完成入侵检测任务.实验结果表明,其相对于传统的有指导入侵检测方法,在保证较高检测率的前提下,有效地降低了误报率;在采用选择后的训练集以及进行特征选择等优化处理后,其性能没有明显的削减,因而更适用于现实的网络应用环境.

关键词 网络安全;入侵检测;TCM-KNN 算法;主动学习;数据挖掘

中图法分类号 TP309

Supervised Intrusion Detection Based on Active Learning and TCM-KNN Algorithm

LI Yang[1),2)] FANG Bin-Xing[1)] GUO Li[1)] TIAN Zhi-Hong[1)]

1)(*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080)

2)(*Graduate University of Chinese Academy of Sciences, Beijing* 100039)

Abstract Supervised network intrusion detection has been an active and difficult research topic in the field of intrusion detection for many years. However, there still exist some unresolved and scarcely addressed problems such as the difficulties in obtaining adequate qualified attack data for the supervised classifiers to model the attack patterns, the data acquisition task is always time-consuming and greatly relies on the domain experts, etc. Based on these, the authors first propose a novel supervised intrusion detection method based on TCM-KNN (Transductive Confidence Machines for K-Nearest Neighbors) data mining algorithm. Moreover, the authors introduce active learning method to select the most qualified data for training and thus assist TCM-KNN effectively in fulfilling the intrusion detection task. Experimental results demonstrate the proposed method has better results both in detection rate and false positives than the state-of-the-art intrusion detection methods. The method can also ensure good detection performance after optimizations by using instance selection and feature selection mechanisms. Therefore, it is more suitable for the real network applications than the traditional ones.

Keywords network security; intrusion detection; TCM-KNN (Transductive Confidence Machines for K-Nearest Neighbors) algorithm; active learning; data mining

收稿日期:2007-03-05;修改稿收到日期:2007-05-23.本课题得到国家自然科学基金(60573134)、国家信息安全计划项目基金(2005C39)资助.李 洋,男,1978 年生,博士研究生,主要研究方向为计算机网络信息安全、基于数据挖掘和机器学习方法的入侵检测技术等.E-mail:liyang@software.ict.ac.cn.方滨兴,男,1960 年生,中国工程院院士,主要研究领域为并行计算、网络信息安全.郭 莉,女,1969 年生,研究员级高级工程师,主要研究方向为网络与信息安全.田志宏,男,1978 年生,博士后,主要研究方向为计算机网络与信息安全.

1 引　言

入侵检测技术是网络安全防御体系的关键技术,它通过对网络和主机上某些关键信息进行收集分析,检测其中是否有违反安全策略的事件或攻击事件发生,并对检测到的事件发出警报.

目前常用的入侵检测技术主要有两种: 误用检测和异常检测[1]. 误用检测是建立在使用某种模式或者特征描述方法对任何已知攻击进行表达这一理论基础上的. 误用检测系统是将已知的攻击特征和系统弱点进行编码,存入知识库中,入侵检测系统(IDS)将所监视的事件与知识库中的攻击模式进行匹配,当发现有匹配时,认为有入侵发生,从而触发相应机制. 这种技术的优点是可以有针对性地建立高效的入侵检测系统,误报率低;缺点是对未知的入侵活动或已知入侵活动的变异无能为力,攻击特征提取困难,需要不断更新知识库. 异常检测基于已掌握了被保护对象的正常工作模式,并假定正常工作模式相对稳定,有入侵发生时,用户或系统的行为模式会发生一定程度的改变. 一般方法是建立一个对应"正常活动"的系统或用户的正常轮廓,检测入侵活动时,异常检测程序产生当前的活动轮廓并同正常轮廓比较,当活动轮廓与正常轮廓发生显著偏离时即认为是入侵,从而触发相应机制. 异常检测与系统相对无关,通用性较强. 它最大的优点是有可能检测出以前从未出现过的攻击方法,不像误用检测那样受已知脆弱性的限制,然而其误报率过高.

传统的入侵检测方法大多基于数据挖掘及机器学习方法,并且,他们都可以转化为相应的分类问题来解决. 如基于 SVM 算法的入侵检测模型、基于多分类支持向量机的网络入侵检测方法等,他们在一定程度上的检测效果比较好. 然而,他们均非常依赖于用于机器学习所需的训练数据集. 但在实际的入侵检测应用中,搜集网络攻击数据并对其进行标记用于训练,是一件非常困难且人力和物力耗费都相当大的工作,需要领域专家(domain expert)的参与. 因此,如何在训练数据稀缺的现实网络环境下保证入侵检测的效率(高检测率和低误报率),成为当前该方向的经典难题.

本文第 2 节介绍了国际上入侵检测方面的相关研究工作;第 3 节提出了一种新型的基于 TCM-KNN 数据挖掘算法的有指导入侵检测方法;第 4 节则针对该方法使用主动学习方法对其进行优化,旨在减少检测所需的已标记训练样本的数量及其工作量和提高算法的检测效果;第 5 节通过对实验结果详细的分析论证了本文所述方法的有效性;第 6 节总结了全文并展望了未来相关的工作.

2 国内外研究现状

在过去的几十年中,入侵检测一直是网络信息安全领域工作者关注的一个焦点和难点问题. 许多机器学习和数据挖掘相关的方法都被学者们用来完成入侵检测任务,并且出现了许多成功和著名的系统和方法.

MADAM(Mining Audit Data for Automated Models for Intrusion Detection)[2] 是数据挖掘算法在入侵检测方面应用最为著名和成功的系统之一. 该 IDS 系统采用离线检测方式,通过应用高效的相关性规则,它能够替代传统的硬编码以及配置文件来自动地产生相应的误用和异常检测模型来完成入侵检测任务. ADAM(Audit Data Analysis and Mining)是另一个普遍使用和非常著名的 IDS[3]. 它采用在线检测方式,使用相关规则以及分类方法来检测绝大多数已知的以及部分未知的攻击. 另外, IDDM(Intrusion Detection using Data Mining Techniques)是一个实时的网络入侵检测系统,它应用关联规则、元规则以及特征规则来进行误用和异常检测. 该 IDS 能够高效地采用数据挖掘方法来对网络数据进行描述建模并进行行为偏离分析,从而检测入侵[4].

除了上述几种基于数据挖掘方法的著名入侵检测系统外,许多其它的机器学习方法也均成功地应用于入侵检测领域. 比如,文献[5]使用决策树(decision tree)和模糊关联规则(fuzzy association rule)来进行入侵检测; Lippmann 等使用神经网络(neural network)来改进现有的入侵检测系统[6];支持向量机(support vector machines)成功地应用于有指导的入侵检测以及无指导的异常检测领域[7,8]. 信息安全领域的研究者们在过去的几十年中,通过使用这些经典的数据挖掘和机器学习方法,极大地推动了入侵检测领域研究的发展.

3 基于 TCM-KNN 算法的入侵检测

3.1 TCM-KNN 基本原理

在统计学习理论中,直推式方法(Transduction)通常是指对于一个样本的类别预测可以直接

通过训练数据中的所有样本来获得,而不是使用传统的归纳(induction)方法从训练数据中得出通用规则[9]. 该概念被广泛地应用于数据挖掘和机器学习领域,因为它只需要满足 iid 假设(即待归类的样本以及用于训练的数据集都是独立且同分布的),并且,它并不需要知道样本数据的分布类型以及分布参数.

直推信度机(Transductive Confidence Machines, TCM)[10]则使用 Kolmogorov 的算法随机性理论[9]建立了一种适应范围较广的机器学习置信度(confidence)机制. 它被用来衡量一个样本分别属于已经存在的几个类别的可信程度. TCM 中所采用的置信度机制基于随机性检测. 然而, Martin Lof 证明[10], 这种检测是不可计算的,因此,我们必须采用一种可计算且满足 Kolmogorov 的算法随机性理论的随机性检测函数来对该置信度进行估算. 这种检测函数的值称为 P 值. 我们通常将 P 值定义为待分类样本属于已存在的几类样本空间的概率. 其相对于某类样本空间的值越大,则表明它属于该类样本空间的可能性越大.

TCM-KNN(Transductive Confidence Machines for K-Nearest Neighbors)将经典的分类算法 K 近邻结合在 TCM 中,采用距离计算的方法根据已分类的数据集对观测点进行分类. 因此,在 TCM-KNN 中,为了计算待检测样本的 P 值,我们定义一种称为奇异值(strangeness)的指标.

定义 **1**. 待检测样本 i 相对于类别 y 的奇异值 α_{iy} 定义为

$$\alpha_{iy}=\frac{\sum_{j=1}^{k}D_{ij}^{y}}{\sum_{j=1}^{k}D_{ij}^{-y}} \tag{1}$$

其中, D_i^y 表示样本 i 与类别 y 中所有样本的距离(在本文中,样本之间的距离计算均通过表示他们的特征向量进行)的序列, D_{ij}^y 则表示该序列中第 j 个最短的距离; 同理, D_i^{-y} 则代表样本 i 与其他类别中(除类别 y 外)所有样本的距离序列, D_{ij}^{-y} 同样表示该序列中第 j 个最短的距离. 参数 k 则表示我们所要考虑的最近邻的数目. 通过该定义,我们不难看出:奇异值是基于样本特征向量在特征空间上的距离来设计的. 一般说来,同类别的样本由于具有相似性,它们的特征向量在特征空间上的分布具有聚集性,样本之间的距离比较小;不同类别的样本由于具有相异性,它们的特征向量在特征空间上的分布具有分散性,样本之间的距离比较大. 奇异值实际上是待检测样本 i 与待加入的类中其他样本最小的 k 个距离之和,与其他类别中样本的最小的 k 个距离之和的比率.

在定义 1 中,本文结合 K 近邻方法给出了奇异值的定义,并且采用 Euclidean 距离(欧氏距离)来计算样本之间的距离,其计算方式如下所示:

$$distance(Y_1,Y_2)=\sqrt{\sum_{i=1}^{|Y_1|}(Y_{1j}-Y_{2j})^2} \tag{2}$$

其中, Y_1 和 Y_2 分别指代两个样本(由该样本的特征向量表示), Y_{ij} 表示特征向量 $\boldsymbol{Y}_i$ 的第 j 维特征, $|\boldsymbol{Y}_i|$ 则表示特征向量 $\boldsymbol{Y}_i$ 的特征维数.

结合定义 1,我们可以给出 TCM-KNN 中, P 值的计算方法.

定义 **2**. 待检测样本 i 相对于类别 y 的 P 值计算为

$$P(\alpha_i)=\frac{\#\{j:\alpha_j\geq\alpha_i\}}{n+1} \tag{3}$$

其中, $\#$ 表示集合的"势",通常计算为有限集合的元素个数; α_i 为待检测样本的奇异值; n 为集合的个数; α_j 表示集合中任意样本的奇异值. 因此, P 值可以计算为 $\frac{j}{n+1}$ (j 为类别 y 中奇异值大于待检测样本 i 奇异值的样本个数). 并且,在计算过程中,通常一次处理一个样本. 不难看出, P 值取值区间为 $[0,1]$,并且其值越大,表明样本 i 归属于类别 y 的可能性越大.

3.2 基于 TCM-KNN 方法的入侵检测

以定义 1 和定义 2 为基础的 TCM-KNN 算法在本质上为分类算法. 在处理分类问题的应用中,它试图将样本归为已有分类中的某一类. 在计算过程当中,当训练集中的某类的任一样本与待分类样本的距离要小于用于计算奇异值的 k 个最短距离中的最大值时,则需要为该类中所有样本重新计算奇异值,从而为待分类样本重新计算 P 值(注意:对应于训练集中的每一类,待分类样本都有一个相应的 P 值需计算). 最后,我们待分类样本划分到最大的 P 值所对应的类. 并且,确定该种分类的置信度值为 1 – 第 2 个最大的 P 值.

在有指导入侵检测领域,我们通常将入侵和正常网络流量分成几大类. 比如:拒绝服务攻击类型、扫描攻击类型、权限提升攻击类型和正常网络流量等. 入侵检测领域经典的 KDD Cup 1999 数据集则将其分为 DoS、Probe、U2R、R2L 和 Normal 五大

类. 那么，通过采用上一节所讨论的基于奇异值度量和 P 值的 TCM-KNN 算法可以对这几种类型的网络流量进行分类，从而检测出网络入侵行为，图 1 给出了采用经典的 TCM-KNN 算法进行入侵检测的算法.

```
算法参数说明：k(选取的 K 近邻数目)、m(训练集样本数目)、
              c(已有攻击分类和正常流量)
输入：r(待检测样本)
输出：class_id(样本的类别编号)

for i=1 to m {
   根据定义 1 为训练集中的每个样本计算 D_i^y, D_i^{-y} 并存储;
   根据式(1)计算训练集中每个样本的奇异值 α 并存储;
}
for j=1 to c {
   对于类 j 中的每个样本 t, if (D_ik^j >= dist(t, r))
        将 r 加入类 j, 并根据式(1)重新为样本 t 计算奇异值 α;
   对于非类 j 中的每个样本 t, if (D_ik^{-j} >= dist(t, r))
        将 r 加入类 j, 并根据式(1)重新为样本 t 计算奇异值 α;
   为待检测样本 r 计算归属于类 j 的奇异值;
   为待检测样本 r 计算归属于类 j 的 P 值;
}
将待检测样本 r 归为 P 值最大之所对应的类, 该分类结果的置
信度为(1-第 2 最大值), return class_id;
```

图 1　基于 TCM-KNN 算法的入侵检测算法

通过图 1 不难看出，上述的 TCM-KNN 算法将根据计算出的最大 P 值来确定网络流量属于正常、DoS 攻击或者是某种特定的攻击类型. 它从本质上来说是一种基于置信度的分类方法，它只要求学习样本是独立同分布的，且不需要知道样本分布的具体类型和参数，因此适应性比较广泛. 这种弱前提条件也很有利于它与其他数据挖掘算法的融合. 它比其他数据挖掘及机器学习方法的优越之处还在于它并非从训练样本得到一个通用的判断规则后再依此对所有未知样本进行非此即彼的判断，这种学习算法不一定需要在某个模式类别的闭集上进行，只需根据不同假设类别情况下的置信度之间的相对大小来判断.

另外，我们可以对该算法进行简单的时间复杂度估算. 首先，为了确定正常训练集中各样本的奇异值，需要耗费 $O(m^2)$ 的时间开销. 其次，一旦有一个样本与训练集中任意一类样本中的某个样本的距离小于该类样本中最小的 k 个距离中的一个时，则需要将该样本加入到此类样本中并为此类中的所有样本重新计算奇异值. 这样的时间开销是非常大的，尤其是在训练集存在多类情况时将更加复杂. 可以看出，虽然第一个时间开销大的运算结果都可以在实际的入侵检测中通过一次离线计算方式得到并多次使用，不需要在检测的判定中临时计算，然而后一个所完成的计算需要在判定时计算而得，且其时间复杂度随着训练集规模的增大而增大. 因此我们认为，影响本算法时间开销的主要因素集中在数据集的规模以及样本所对应的特征向量的维数上，因而我们需要限制训练集的规模和减少特征维数，从而来降低时间开销. 本文第 4 节将介绍如何采用主动学习策略在保证 TCM-KNN 算法入侵检测性能的前提下，来限制训练集的规模；同时，在第 5 节本文也将用实验来证明采用特征选择方法对于 TCM-KNN 算法的有效性.

值得注意的是：在应用 TCM-KNN 算法进行实际的入侵检测工作时，由于该算法需要基于距离的度量方法来计算奇异值和 P 值指标，所以我们通常必须对实际的网络流量进行特征选取和抽象，形成相应的特征空间(feature space)，从而形成最终的特征向量(feature vector)来使用 TCM-KNN 算法来进行距离计算和入侵检测. 而这些相关的特征选取工作需要根据实际的应用场景来确定，从而达到较好的实践效果. 在本文的后述实验中，我们将采用入侵检测领域权威的 KDD Cup 1999 测试集来进行，该数据集采用了 41 个特征，并且数据集中的每条记录实际上已经完成了此处强调的特征选取以及特征向量的形成过程，因而可以方便地采用 TCM-KNN 算法进行距离计算和对几类入侵进行高效地检测.

4　面向 TCM-KNN 算法的主动学习方法

如同本文第 3 节所述，TCM-KNN 算法从实质上来说是一种有指导的分类方法，与其他数据挖掘和机器学习算法类似，为了用其进行入侵检测且达到较好的检测效果，同样需要对大量标记的训练样本进行训练. 在现实的网络环境中，准备足够的入侵样本进行分类学习的代价是非常昂贵的，标记一个正确的入侵样本往往要花费领域专家(domain expert)几小时甚至是几天的人力和相应的物力. 所以，如何通过少量的样本学习来训练得出较高质量的入侵检测分类器成为有指导入侵检测领域的经典难题. 并且，按照 3.2 节对 TCM-KNN 算法的时间复杂度分析，精简训练样本不但是有指导入侵检测需要解决的问题，也是降低 TCM-KNN 算法本身的时间复杂度、减少计算开销从而提升其实用性的必要步骤. 因而，本节将采用主动学习方法来解决这些

问题.

4.1 主动学习(Active Learning)方法原理

在数据挖掘领域,按照分类学习对训练样本的处理方式可将分类模型分为两类:被动分类模型和主动分类模型[11]. 被动学习也称为从样本中学习,它随机地选择训练样本,被动地接受这些样本的信息. 它对于具有严格顺序关系的训练样本来说是必要的,也是不可改变的. 然而,绝大部分分类学习中都认为训练样本是独立同分布的,并不具有必然的顺序关系,所以顺序地处理训练样本往往会使学习的分类器具有顺序相关性、对数据过分敏感(比如遇到噪声样本时,会使这种噪声一直传播下去,影响分类精度)、缺乏综合未带标注样本信息的能力. 在学习分类模型中,未带类别标注的样本往往包含有助于分类的信息. 在这种情况下,选择好的未带类别标注的样本,把它加入到当前的分类器中是相当重要的. 主动分类模型则是满足上述条件的一个最好选择,它对训练样本的选择是主动的,它首先选择最有利于分类器性能的样本来训练分类器. 它不但能够通过选择高质量的样本来训练分类器,并且大大地减少了用于训练的样本数量,减少了训练的时间开销以及"无用"或者"噪音"样本对于分类器的负面影响. 我们通常采用两种方法来评价主动学习方法的效果:(1)为达到某个性能指标相对于被动学习方法所需要的训练样本集的精简比例;(2)对于某个定量的训练集采用主动学习方法相对于被动学习所提升的性能比例.

经典的主动学习方法包含两个重要的部分:一个学习器(learner)和一个选择函数(query function). 通常情况下,用户训练的数据集可以分为已标记的样本集 TR 和未标记的样本集 U. 学习器则通过从 TR 和 U 的训练中得到. 在训练过程中,选择函数决定 U 样本集中的哪些样本将加入到样本集 TR 中进行训练从而得到学习器. 在常规的被动学习模型中,选择函数通常是随机选择样本进行学习,而主动学习模型则选择对于学习器最有意义的样本进行学习.

主动学习方法已经在各个领域得到成功的应用,它能够在保证分类器性能的前提下,极大地减少用于训练所需的数据量. 在国际上,Baldridge 和 Osborne 使用该方法在分析选择方面能够节约大概73%的用于标注样本的工作量[12];Tong 和 Koller 在文本分类方面采用主动学习方法也取得了非常积极的效果[13]. 在国内,宫秀军等将主动学习方法应用于贝叶斯网络分类模型[11],在有少量带标记的训练样本的情况下获得了较好的分类精度和召回率.

4.2 选择函数选取

在采用主动学习策略来为本文所述的 TCM-KNN 算法减少和精选训练数据集之前,我们需要为上一小节所述的选择函数选择合适的选择策略. 不确定采样(Uncertainty Based Sampling, UBS)以及投票选择(Query By Committee, QBC)方法是其中应用最为普遍和有效的两种[11]. UBS 选择策略假设分类器总是选择分类中最不确定的样本(通过赋予每个未标记样本一个基于概率的分类结果),然后交由领域专家进行标记,再进行学习. 这样的选择策略就避免了学习器对于重复的、无意义的样本的学习,既提高了学习效率,也减少了领域专家的手工标记工作量. QBC 选择策略则首先根据概念在搜索空间的分布获得概率假设,利用这个假设预测样本的标注. QBC 选择投票产生的标签与假设预测的标签不一致的样本作为候选样本,然后同样交由领域专家进行标记,再进行学习. 由于本文所述的 TCM-KNN 算法是一种基于置信度的数据挖掘方法,因而在本质上使用 UBS 选择策略较为恰当,本文将采用 UBS 选择策略来构建 TCM-KNN 算法的选择函数.

考虑本文所述的 TCM-KNN 算法,按照本文图 1 所示的伪代码,我们可以为每一个待判定的样本(网络流量,由其特征向量表示)得到一系列的 P 值,每个 P 值都对应了该样本属于每类样本的可能性大小. 将这些 P 值进行降序排列,可以得到最大的两个 P 值:P_j 和 P_k. 我们假设 $P_j > P_k$,那么该样本最终分类结果为类 j. 正如图 1 所示,P_j 表明了样本属于 j 类的可信程度(credibility),而 P_k 则表明了样本属于 j 类的置信程度(confidence). 理论上来说,P_j 越大且 P_k 越小则表明分类结果更为准确和合理. 因而,我们一般都希望 P_j 尽可能接近 1,而 P_k 尽可能接近 0. 那么,使用该 P 值对,我们可以得到如下 4 种可能,从而来构建主动学习方法的选择函数:

(1) P_j 高且 P_k 低:表明预测结果有较高的可信度和置信度;

(2) P_j 高且 P_k 高:表明预测结果有较高的可信度,但置信度较低;

(3) P_j 低且 P_k 低:表明预测结果有较低的可信度,但置信度较高;

(4) P_j 低且 P_k 高:表明预测结果可信度和置信度都较低.

不难看出，上述 4 种情况中第一种情况是最好的（P_j越大且 P_k越小时），对样本的预测结果最为确定和准确；而当 $P_j \approx P_k$时，预测效果最差. 那么，对于基于 UBS 选择策略的主动学习方法，我们根据 TCM-KNN 算法给出如下选择函数（query function）：

$$C(i)=|P_j - P_k| \tag{4}$$

式(4)中，$C(i)$表示 P 值对的绝对偏差程度，当 $C(i)$小于一个非常小的接近零的实数 ε 时，则表示分类器对样本的分类结果最不确定，从而该样本即为该选择函数选择出来的需要标记且学习的下一个样本. 实数 ε 为经验设定值，本文中依据多次反复实验设定其经验值为 0.1.

4.3　主动学习方法中止策略

本文所述的主动学习方法在执行过程中通常需要使用相应的策略来对其进行中止，一方面可以减少样本选择的执行时间，另一方面可以控制该方法对于本文所述数据挖掘算法 TCM-KNN 的检测率及误报率产生的负面影响，从而达到该算法中止后所产生的样本集合可以导致相对最佳的学习效果（高检测率和低误报率）的目的.

在本文中，我们使用最小均方预测误差（Mean Squared prediction Error，MSE）[14] 来作为本文所述主动学习方法的参考指标. 那么，针对训练集 D 在测试样本集 x 下的预测函数 f^* 的 MSE 可以定义如下：

$$MSE(f^*(x,D))=bias^2(f^*(x,D))+var(f^*(x,D)) \tag{5}$$

其中，$bias(f^*(x,D))=f(x,D)-E(f^*(x,D))$，$f$ 即为 TCM-KNN 检测算法. $bias$ 函数反映了训练集 D 对于预测函数 f^* 选取的敏感度，其取值为测试集 x 中所有样本的实际分类与预测分类差异的平方；而 var 函数则刻画了预测函数对于训练集 D 的敏感度，其取值为随着训练集 D 的选择增加，测试集 x 中所有样本的预测分类的差异幅度. 因此，为了获得较小的 MSE，则需要获得低偏差（bias）和变化值（variance）. 然而，在现实情况中，很难保证上述两个数值同时都很小，因而训练器只能在两者之间取得折衷，从而导致较小的 MSE 才能获得相对的理想结果. 并且，文献[14] 证明，在分类学习任务中，较低的 variance 指标相对 bias 作为评价标准更为适合和有效.

因此，根据以上所述，针对本文所述的 TCM-KNN 算法来说，在使用主动学习方法及其中止策略的过程中，我们期望通过选择较少的样本来获得最大的学习信息以及最佳的学习效果，也就是说，随着训练集中新的选择样本（通过主动学习方法）的不断增加，上述的期望偏差应该快速地下降，而变化值也应该慢慢地趋于稳定. 我们通过实验发现，这个时刻也往往是检测率达到最高的时刻，因而不但通过主动学习限制了训练集的规模，也保证了较高的学习效果. 所以，我们选取最低的变化值 $variance$ 作为主动学习方法的中止策略来控制其执行过程. 那么，我们可以进一步将测试集 x 中的预测分类与实际分类存在差异的比率来作为主动学习方法的可计算中止策略，它实际上也是一个经验阈值. 同理，本文中依据多次反复实验设定其经验值为 0.05.

4.4　面向 TCM-KNN 的主动学习方法

根据上面构造的选择函数以及中止策略，我们使用本文所论述的 TCM-KNN 算法作为学习器，给出了针对该算法的主动学习方法的伪代码，如图 2 所示. 在图 2 中，我们假定有一个少量的已标记好的训练集 TR 和一个未标记的样本池 U. 在算法中，我们循环地选取样本池中的样本，使用选择函数进行选择判定，一旦对应的 $C(i)$ 值小于一个预先设定的阈值 ε（本文取经验值 0.1），则请求标记并将其加入已标记好的训练集 TR 中进行训练，直到 $variance$ 小于主动学习方法中止阈值 r（本文取经验值 0.05）为止. 最后得到的训练集 TR 即为经过选择和精简后的训练集. 在本文的后续实验中将详细给出该方法的具体效果.

```
算法参数说明：l 为少量已标记训练集，U 为大量未标记样本
              池，x 为测试样本集合，ε 为不确定计算的阈值，
              r 为主动学习方法中止阈值

将训练集初始化为 l；
repeat
{
  从 U 中随机选择样本 i 并计算其 P 值；
  if ( C( i)< ε)
    {
      将样本 i 加入训练集 l 并从样本池 U 中删除该样本；
      根据测试样本集 x 计算 MSE 值；
    }
}
until ( variance< r)
```

图 2　面向 TCM-KNN 算法的主动学习方法

5　实验结果及分析

在本节，我们将对本文提出的有指导的入侵检测方法的有效性进行验证. 为了保证实验的说服力和方便性起见，本节采用入侵检测研究领域共同认

可及广泛使用的基准评测数据集 KDD Cup 1999 数据集进行测试.

在实验中,首先,我们将本文所述方法与入侵检测领域较为著名的人工神经网络(Artificial Neural Network, ANN)方法、SVM(Support Vector Machines, SVM)方法和 K 近邻(K-Nearest Neighbors, KNN)方法的检测效果进行了比较;然后,我们验证了面向 TCM-KNN 算法的主动学习方法在保证训练效果、精简训练集以及保证较高的入侵检测性能方面的有效性;并且,我们通过特征选择技术,对形成特征向量的特征进行冗余去除,得到最重要和必要的特征形成向量后再进行训练和检测,以测试 TCM-KNN 算法在进行基于特征空间约减的优化后的检测效果. 本文方法采用的评价指标为国际上通用的检测率(True Positive rate, TP)和误报率(False Positive rate, FP)指标. 其中,检测率定义为正确检测入侵样本的数量与测试集中入侵样本的数量之间的比率,而误报率则定义为错误判为入侵的正常样本数量与测试集中正常样本的数量之间的比率.

5.1 实验数据及预处理

本文采用的 KDD Cup 1999 数据集包括大约 4900000 条数据记录,每条都是从军方网络环境中模拟攻击所得的原始网络数据中根据设定的 41 个特征提取出来的,它们都是描述网络连接统计信息的特征向量. 它们包含五类数据:DoS, Probe, R2L, U2R 四类攻击数据(共包含 24 种攻击类型)以及正常(Normal)数据.

在该数据集所提取的 41 个特征中,主要有两类数据类型:数值型和名词型. 为了应用 TCM-KNN 算法进行实验,首先需要对其中的数值型数据进行归一化(normalization)处理,因为需要计算特征向量间的欧氏距离,而该距离容易出现由于取值范围的差异,而造成一个数值型数据影响另一个数值型数据的情况,所以需要对他们进行处理. 归一化处理方法的步骤如下.

首先分别计算出训练样本每个特征属性的均值和标准差:

$$mean[j]=\frac{1}{n}\sum_{i=1}^{n}instance_i[j] \tag{6}$$

$$standard[j]=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(instance_i[j]-mean[j])^2} \tag{7}$$

其中, $instance_i[j]$ 表示训练样本 i 中的第 j 个属性, n 表示样本的数目.

然后,我们将训练集中的样本按如下方式转换:

$$newinstance[j]=\frac{instance[j]-mean[j]}{standard[j]} \tag{8}$$

可见,式(8)实际上是将属性的取值转换为这个取值偏离均值时标准差的倍数,这样,我们就可以把样本的属性值从它自己的取值空间映射到标准的取值空间.

对于数据集中诸如协议类型、服务类型等名词型属性,我们则根据其每个取值在取值空间中出现的频率进行标准化,这样,这些属性的取值空间将被限定在 0~1 之间.

5.2 对比实验

在对比实验中,我们将 KDD Cup 1999 的数据集进行了提取. 我们在数据集中随机提取了 1843 条正常数据和 9783 条攻击数据,采用十折交叉验证(ten-fold cross validation)的方法,重复实验十次,取检测率和误报率的平均值对几个方法进行了对比,实验结果如表 1 所示.

表 1 对比实验结果

算法	TP/%	FP/%
SVM	99.5	0.14
ANN	99.8	0.22
KNN	99.2	0.33
TCM-KNN	99.7	0

在实验中,为了保证实验的公平性和可比性起见,我们对著名的机器学习算法软件 WEKA[15] 中的 SVM、ANN、KNN 三种算法挑选不同的参数进行了多次实验,各取其检测效果最好的结果作为比较. 其中,SVM 的参数为:类型为 C-SVC,核函数为 RBF(Radial Basis Function);ANN 算法采用三层结构,输入层、中间层和输出层各一个,*learning rate* 为 0.3, *momentum* 为 0.2, *training time* 为 1000, *validation threshold* 为 20;KNN 算法近邻数为 10,各属性不带权重. 对于本文所述的 TCM-KNN 方法,我们则使用一系列参数进行实验选择,最后选取了实验效果最为理想的参数:K 近邻数目为 50,置信参数为 0.95(在后面的实验中,也沿用这两个效果较为理想的参数). 从结果中我们不难看出,本文所述的 TCM-KNN 方法在检测率上要略高于其他三种著名的机器学习方法,而误报率则明显地低于它们,因而该方法在入侵检测领域较其他三者具有较好的应用优势.

5.3　采用主动学习方法后的实验

另外，为了验证主动学习方法对于 TCM-KNN 算法及其入侵检测的有效性，我们特别关注了 TCM-KNN 算法在达到同等较高的入侵检测率的条件下，采用主动学习方法所需样本的数量与随机采样方法所需样本数量的比较结果(如图 3 所示)．结果很显著：TCM-KNN 在达到 99.7%左右的高检测率的条件下，使用主动学习方法仅需要 40 个左右的样本进行训练，而随机采样则需要 2000 个左右，这就证明了在 TCM-KNN 算法中引入主动学习方法选择及限制样本规模对于保证检测性能的有效性，也为我们在实践中使用主动学习方法来提升基于 TCM-KNN 的入侵检测技术的性能和可用性方面提供了很好的参考依据．

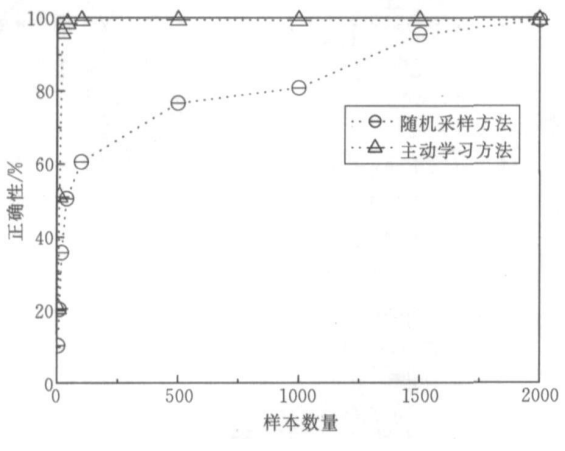

图 3　主动学习方法与随机采样方法在 TCM-KNN 算法应用中的实验对比

5.4　采用特征选择后的实验结果

由于测试数据集以及在实际应用过程中，冗余的特征极有可能对 TCM-KNN 的性能产生影响，特征选择可以优化其性能．为了对本文所述的 TCM-KNN 方法进行优化，减少其运算量，并通过特征约简来提升其在实践中的可用性．因此，最后，我们采用广泛使用的 Chi-Square 特征选取方法，对该训练集和测试集进行特征选择．从而对这些特征向量实行降维处理．

表 2 给出了特征选择后通过排序保留下来的 6 个重要特征，表 3 给出了特征选取前后 TCM-KNN 方法检测率和误报率的比较结果．不难看出，通过特征选择的降维处理后，TCM-KNN 方法的检测率和误报率都与未处理前基本一致(检测率有非常微弱的下降，但仍然是可以接受的)，因而充分说明了该方法用于入侵检测方法的可扩展性很强．在实践中，我们可以通过相应的优化工作来减少计算量，而保证性能没有减退．

表 2　特征选择后保留下来的重要特征

序号	Chi-Square 值	特征
1	17586.107	dst_host_rerror_rate
2	17368.831	src_bytes
3	17073.438	dst_bytes
4	17032.989	count
5	14357.396	num_compromised
6	16503.031	dst_host_srv_rerror_rate

表 3　特征选择前后的实验结果

	TP/%	*FP*/%
原始数据集	99.7	0
特征选择后的数据集	99.7	0.11

另外，为了说明上述主动学习方法和特征选择对 TCM-KNN 算法的优化效果，我们采用独立测试集(在这里我们使用训练集：9738 个样本和测试集：2460 个样本)，详细比较了性能优化前后 TCM-KNN 算法的分类器建立时间以及入侵检测时间(结果见表 4)．结果表明，采用优化方法后，建立时间和检测时间分别有 98.65%以及 66.45%的大幅度下降，从而说明了优化工作的合理性和有效性．

表 4　性能优化(特征选择＋基于主动学习的样本选择)前后的实验结果

	TP/%	*FP*/%	建立时间/s	检测时间/s
原始数据集	99.7	0	4173.43	56.21
特征选择后的数据集	99.6	0	1024.29	343.66

5.5　实验结果分析

本节上述大量实验结果都充分证明了本文所述方法的正确性和有效性：在对比实验中，TCM-KNN 算法的检测率要略高于目前最为有效的几种数据挖掘和机器学习算法，其误报率也相对较低；并且，采用了面向 TCM-KNN 算法的主动学习方法对其训练样本进行选择和规模限制后，其仍能保证较高的检测性能；另外，由于本方法需要计算大量特征向量之间的距离，如果数据量过大和表征数据的特征向量维数过多，将会引发“维灾难”(curse of dimensionality)和过大的运算量，而使得本方法的实用性能大打折扣．因此，我们也通过实验证明了本方法完全能够通过减少训练的数据量和特征向量的维数来提升实用性，而性能上没有明显的削减．

另外，特别值得注意的是：本节实验充分证明了本文所述主动学习方法不但能够极大地减少用于入侵检测训练集的规模，而且能够获得较高的入侵检

测性能(见图 3). 在应用的过程中,我们通常需要首先通过主动学习方法来对 TCM-KNN 算法的训练样本进行选取,然后使用该样本集进行训练从而完成入侵检测任务. 并且,随着网络流量的不断变化以及随之变化的训练样本质量的不同,我们需要周期性地使用主动学习方法来重新选择训练样本,从而保证样本质量和与之直接相关的 TCM-KNN 算法的入侵检测性能.

6 总结及展望

TCM-KNN 算法是一种非常有前途的数据挖掘方法,特别适应于分类学习领域,我们已经将其成功地应用于异常检测领域[16]. 本文将其应用于有指导的入侵检测领域,并引入主动学习方法对其训练数据进行选择和规模限制,不但保证了较好的入侵检测性能,并且极大地降低了由于标注样本所需的人力物力以及给算法带来的过大的计算开销,极大地提升了其在现实网络环境中进行入侵检测的实用性. 在经典 KDD Cup 1999 数据集上的大量实验证明:该方法行之有效,具有较高的检测率和较低的误报率,与国内外同领域的其他有指导入侵检测方法相比也具有相当的优势.

本文所提出的方法在实践中还需根据实际应用情况作进一步的改进,以提高其性能. 如何将本文所述方法应用于实际的网络环境中并加以优化以及将其有效地应用于检测 DDoS 攻击[17]以及检测 Web 服务器的异常情况将是我们下一步的工作重点.

参 考 文 献

[1] Bykova M, Ostermann S, Tjaden B. Detecting network intrusions via a statistical analysis of network packet characteristics//Proceedings of the 33rd Southeastern Symposium on System Theory. Ohio, Athens, 2001: 309-314

[2] Lee W, Stolfo S J. A framework for constructing features and models for intrusion detection systems. ACM Transactions on Information and System Security (TISSEC), 2000, 3(4): 227-261

[3] Barbarra D, Couto J, Jajodia S, Popyack L, Wu N. ADAM: Detecting intrusions by data mining//Proceedings of the 2001 IEEE, Workshop on Information Assurance and Security. West Point, NY, USA, 2001: 11-16

[4] Tamas A. IDDM: Intrusion detection using data mining techniques. Salisbury, Australia: DSTO Electronics and Surveillance Research Laboratory: Technical Report DSTO-GD-0286, 2001

[5] Luo J, Bridges S M. Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection. International Journal of Intelligent Systems, 2000, 15(8): 687-704

[6] Lippmann R P, Cunningham R K. Improving intrusion detection performance using keyword selection and neural networks. Computer Networks, 2000, 34(4): 597-603

[7] Eskin E, Arnold A, Prerau M, Portnoy L, Stolfo S P. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data//Barbara D, Jajodia S eds. Applications of Data Mining in Computer Security. Boston: Kluwer Academic Publishers, 2002: 78-99

[8] Mukkamala S, Janoski G, Sung A H. Intrusion detection: Support vector machines and neural networks//Proceedings of the IEEE International Joint Conference on Neural Networks. Honolulu, USA, 2002: 1702-1707

[9] Barbara D, Domeniconi C, Rogers J P. Detecting outliers using transduction and statistical testing//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2006: 55-64

[10] Proedru K, Nouretdinov I, Vovk V, Gammerman A. Transductive confidence machine for pattern recognition//Proceedings of the 13th European conference on Machine Learning. London, UK, 2002: 381-390

[11] Gong Xiu-Jun, Sun Jian-Ping, Shi Zhong-Zhi. An active bayesian network classifier. Journal of Computer Research and Development, 2002, 39(5): 574-579(in Chinese)
(宫秀军, 孙建平, 史忠植. 主动贝叶斯网络分类器. 计算机研究与发展, 2002, 39(5): 574-579)

[12] Baldridge J, Osborne M. Active learning for HPSG parse selection//Proceedings of the 7th Conference on Natural Language Learning. Edmonton, Canada, 2003: 100-108

[13] Tong S, Koller D. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 2001, 2(11): 45-66

[14] Friedman J. On bias, variance, 0/1-Loss, and the curse of dimensionality. Data Mining and Knowledge Discovery, 1997, 1(1): 55-77

[15] Witten I H, Frank F. Data Mining: Practical Machine Learning Tools and Techniques. Second edition. Netherlands: Elsevier, 2005

[16] Li Y, Fang B X, Guo L, Chen Y. Network anomaly detection based on TCM-KNN algorithm//Proceedings of the 2nd ACM Symposium on Information, Computer and Communications Security 2007 (ASIACCS 2007). Singapore, 2007: 13-19

[17] Tian Z H, Hu M Z, Li B, Liu B, Zhang H L. Defending against distributed denial of service attacks with an auction-based method. Web Intelligence and Agent Systems, 2006, 4(3): 341-351

LI Yang　born in 1978, Ph. D. candidate. His research interests include network security, intrusion detection techniques based on data mining and machine learning methods, etc.

FANG Bin Xing　born in 1960, member of Chinese Academy of Engineering. His research interests include parallel computing, network and information security.

GUO Li　born in 1969, professor. Her research interests include network and information security.

TIAN Zhi Hong　born in 1978, postdoctor. His research interests include network and security.

Background

The problem addressed in this paper is one of the most significant problems in network security and especially in intrusion detection field. Current intrusion detection methods are mostly based on data mining or machine learning schemes and they greatly depend on the quality of training dataset for building intrusion detection model. However, in the complex network environment training data is very scarce, difficult to acquire and the collection work is time-consuming, thus eventually result in that the detection performances of those methods are not ideal. Moreover, current researches usually ignored and hardly addressed the above problems in the recent years.

To solve these problems, this paper first presents a novel supervised intrusion detection methods based on TCM-KNN data mining scheme. Secondly, it introduces Active Learning method to fulfill instance selection task for TCM-KNN, which can effectively reduce the computational cost of TCM-KNN while keeping good intrusion detection performance, thus make TCM-KNN be a good candidate for intrusion detection in real network environment.

This work is supported in part by the National Natural Science Foundation of China under grant No. 60573134 and the National Information Security Project of China under grant No. 2005C39. These projects mainly are focused on how to secure the network security and information infrastructures by early detecting network intrusions and making corresponding responses as soon as possible. Because accurate and effective intrusion detection is the premise of intrusion response, the relevant work presented in this paper plays a vita role in the projects. To date, the authors have successfully developed supervised intrusion detection methods and unsupervised anomaly detection methods based on TCM-KNN scheme. They are embarking on applying them to large-scale network intrusion detection and response applications.