ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.10, October 2007, pp.2595−2604
DOI: 10.1360/jos182595

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于直推式方法的网络异常检测方法*

李 洋[1,2+], 方滨兴[1], 郭 莉[1], 陈 友[1,2]

[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 100049)

A Network Anomaly Detection Method Based on Transduction Scheme

LI Yang[1,2+], FANG Bin-Xing[1], GUO Li[1], CHEN You[1,2]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: Phn: +86-10-62600951, Fax: +86-10-62600905, E-mail: liyang@software.ict.ac.cn, http://www.ict.ac.cn

Li Y, Fang BX, Guo L, Chen Y. A network anomaly detection method based on transduction scheme. *Journal of Software*, 2007,18(10):2595−2604. http://www.jos.org.cn/1000-9825/18/2595.htm

Abstract: Network anomaly detection has been an active and difficult research topic in the field of intrusion detection for many years. Up to now, high false alarm rate, requirement of high quality data for modeling the normal patterns and the deterioration of detection rate because of some "noisy" data in the training set still make it not perform as well as expected in practice. This paper presents a novel network anomaly detection method based on improved TCM-KNN (transductive confidence machines for *K*-nearest neighbors) machine learning algorithm, which can effectively detect anomalies using normal data for training. A series of experiments on well known KDD Cup 1999 dataset demonstrate that it has lower false positive rate, especially higher confidence under the condition of ensuring high detection rate than the traditional anomaly detection methods. In addition, even provided with training dataset contaminated by "noisy" data, the proposed method still holds good detection performance. Furthermore, it can be optimized without obvious loss of detection performance by adopting small dataset for training and employing feature selection aiming at avoiding the "curse of dimensionality".

Key words: network security; anomaly detection; strangeness; TCM (transductive confidence machines); TCM-KNN (transductive confidence machines for *K*-nearest neighbors) algorithm

摘 要: 网络异常检测技术是入侵检测领域研究的热点和难点内容,目前仍然存在着误报率较高、对建立检测模型的数据要求过高、在复杂的网络环境中由于"噪音"的影响而导致检测率不高等问题.基于改进的TCM-KNN(transductive confidence machines for *K*-nearest neighbors)置信度机器学习算法,提出了一种网络异常检测的新方法,能够在高置信度的情况下,使用训练的正常样本有效地对异常进行检测.通过大量基于著名的 KDD Cup 1999 数据集的实验,表明其相对于传统的异常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.另外,在训练集有少量"噪音"数据干扰的情况下,其仍能保证较高的检测性能;并且在采用"小样本"训练集以及为了避免"维

* Supported by the National Natural Science Foundation of China under Grant No.60573134 (国家自然科学基金); the National Information Security 242 Project of China under Grant No.2005C39 (国家 242 信息安全计划项目)

Received 2006-10-10; Accepted 2007-01-23

灾难”而进行特征选取等优化处理后,其性能没有明显的削减.

关键词: 网络安全;异常检测;奇异值;直推式信度机;TCM-KNN 算法
中图法分类号: TP393 文献标识码: A

入侵检测系统是网络安全防御体系的一个重要组成部分,它通过对网络和主机上某些关键信息进行收集和分析,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(intrusion detection system,简称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,则认为有入侵发生,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.当有入侵发生时,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与正常轮廓比较,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.

异常检测的思想最早由Denning提出[2],即通过监视系统审计记录上系统使用的异常情况,可以检测出违反安全的事件.该思想很快被应用到网络异常检测.网络异常检测方法又分为两类[3]:有指导异常检测(supervised anomaly detection)和无指导异常检测(unsupervised anomaly detection).对于前者,系统会被给定一个全部为正常样本的数据集和一系列没有被标记的样本,任务就是找出这些未标记数据是否与正常的数据有偏离;概率统计分析方法[4]、人工免疫算法[5]、数据挖掘方法[6]等都属于该类范畴.然而,该类方法需要完全“干净”的数据集来生成模型,这在复杂的网络条件下常常是不可满足的,因而在实际中的应用并不是很普遍;对于后者,系统通常被给定一个未标记的训练集,并且不知道训练集中哪些数据是正常的,哪些数据是异常的,目标就是发现其中的异常样本.Columbia大学的Eskin[7]等人提出的基于聚类的估计算法、改进的*K*-近邻方法以及one-class SVM(support vector machines)方法都属于此类范畴.这些方法相对于有指导异常检测方法来说,在应用范围上更加广泛,它们并不需要完全“干净”的数据集建立模型,仅仅需要提供的训练集中正常数据相对于异常数据来说占绝大多数(通常情况下,正常数据的比率通常占 98.5%~99%左右,而异常数据占 1%~1.5%左右).并且,其中one-class SVM方法的检测率高达 98%,然而其误报率同样较高(高达 10%).

针对以上两类异常检测方法的优、缺点,本文提出了一种异常检测新方法,它基于TCM-KNN(transductive confidence machines for *K*-nearest neighbors)算法.该算法依据Kolmogorov的算法随机性理论,是一种有效的基于置信度机制的机器学习方法,已经广泛地应用于模式识别[8]、欺诈检测(fraud detection)及“离群点”检测(outlier detection)[9]等领域,并取得了较好的实践效果.本文首次将其应用于入侵检测的异常检测领域,并对其进行了改进.通过大量基于著名的KDD Cup 1999 数据集的实验测试,验证了其有效性.该方法与其他同类异常检测方法相比,可以在保证高检测率的前提下,极大地减少误报率.更为重要的是,在训练集存在“噪音”数据的干扰,以及在仅有“小样本”训练集的环境中,其均能保证较高的检测性能.

本文第 1 节介绍 TCM-KNN 算法的理论背景.第 2 节全面阐述基于该算法的网络异常检测方法.第 3 节给出算法的实验结果,并对实验结果进行分析和对比.第 4 节给出本文的结论.

1 TCM-KNN 算法理论背景

在统计学习理论领域中,直推式(transduction)方法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)方法采用从训练数据中得出的通用规则的方法来进行[9].这一概念被广泛应用于机器学习领域,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,简称TCM)[8]则使用Kolmogorov的算法随机性理论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[9],这种随机性检测是不可计算的.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 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(在本文中,样本之间的距离计算均通过表示它们的特征向量进行)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的奇异值定义α_{iy}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(除类别 y 外)中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:

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

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Y_i的第j维特征,$|Y_i|$则表示特征向量Y_i的特征维数.

结合定义 1,我们可以给出 TCM-KNN 中,*P* 值的计算方法如下所示.

定义 2. 待检测样本 i 相对于类别 y 的 *P* 值计算为

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

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.

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* 值需要计算).最后,我们将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 最大 *P* 值).经典的 TCM-KNN 算法伪代码如下所示.

算法 1. 经典的 TCM-KNN 算法.

算法参数说明:*k*(选取的最近邻数目)、*m*(训练集样本数目)、*c*(已有分类数)
输入:*r*(待检测样本);
输出:*class_id*(样本的类别编号).
/*算法开始*/
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_{tk}^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_{tk}^{-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* 值最大时所对应的类,该分类结果的置信度为 1(第 2 最大 *P* 值),return *class_id*;
/*算法结束*/

2 基于改进的 TCM-KNN 算法的网络异常检测模型

第 1 节所述的 TCM-KNN 算法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学习机器算法的融合.其不同之处还在于,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.

然而,将其应用于入侵检测的异常检测领域需要进一步地改进,主要是因为异常检测并不需要事先提供详尽的攻击数据,建立相应的分类.因此,不同于模式识别和误用检测,它不是一种简单的分类问题,而是根据已建立的"正常模式"对新来的数据作异常与否的一种判定.本节将详细阐述对其进行的改进以及基于此的一个网络异常检测框架的总体结构.

2.1 改进的TCM-KNN算法

网络异常检测的任务是根据正常训练集建立的模型,判定新来的数据正常或者异常.根据 TCM-KNN 算法的判定要求,我们可以将用于异常检测的训练集定义为从网络数据中抽取的具有正常行为模式的样本集,每个样本以其特征向量表示,其多维特征可定义为 IP 地址对、端口号、协议类型、TCP 连接的统计信息等.这非常类似于 KDD Cup 1999 数据集,其每条记录都包含提取的 41 个特征.那么,接下来的异常检测任务就是需要判定新来数据的特征向量相对于正常训练集是否异常来进行判定.并且在异常检测中,训练集中只有一类正常数据,不存在多类,所以正如第 1 节所述,我们需要对与该算法紧密相关的奇异值进行重新定义,其新定义如下:

定义 3. 待检测样本*i*相对于正常类别*y*的奇异值定义α_{iy}为

$$\alpha_{iy} = \sum_{j=1}^{k} D_{ij}^{y} \tag{4}$$

其中,各个符号的含义与式(1)中的完全相同,这里不再赘述.

该新定义使得不属于正常类样本的奇异值远远大于在该正常类中样本的奇异值,因而它充分地将非正常数据与正常数据进行"隔离(isolation)".该定义先后为Daniel[9]和Angiulli[10]使用,实践中取得了很好的区分效果.

因此,本文也借用了该定义.第 3 节的实验结果也证明了该定义在异常检测领域的有效性.

另外,基于奇异值的 P 值的计算方法与定义 2 相同,不需要作改变,算法 2 给出了本文所提出的改进的面向异常检测的 TCM-KNN 算法伪代码.使用改进的 TCM-KNN 算法进行异常检测的流程可以简单地描述为:给定正常训练集和一批待检测样本,通过其奇异值的计算以及事先计算好的正常训练集中所有样本的奇异值,我们可以得到待检测样本相对于正常训练集的 P 值,如果该值小于预定义的阈值τ(通常为 0.05),则我们可以以置信度为 $1-\tau$(通常情况下为 95%)来判定其为异常;否则认为其正常.

算法 2. 改进的用于异常检测的 TCM-KNN 算法.

算法参数说明:k(选取的最近邻数目)、m(训练集样本数目)、设定的置信度阈值τ

输入:r(待检测样本);

输出:*normal* 或者 *abnormal*.

/*算法开始*/

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 并存储;

    根据式(4)计算训练集中每个样本的奇异值α并存储;

}

根据式(4)计算待检测样本 r 的奇异值;

根据式(3)计算待检测样本 r 的 P 值;

if ($p\leq\tau$)

    以置信度($1-\tau$)判定样本r为异常,return *abnormal*;

else

    以置信度($1-\tau$)判定样本r为正常,return *normal*;

/*算法结束*/

下面我们简单分析该算法的时间复杂度.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,为了计算s个待检测样本的奇异值,则需要$O(sm)$的时间开销,而计算其相应的P值则只需要时间开销$O(m)$.不难看出,第一个时间开销大的运算结果都可以在实际的异常检测中通过一次离线计算方式得到并多次使用,不需要在异常检测的判定中临时计算,而只有后两个时间开销($O(sm)+O(m)$)所完成的计算需要在判定时计算而得.由于我们的检测模式是每次一个样本的判定方法.因此不难看出,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,在实践中我们对此可以加以控制以降低时间开销.本文第 3 节将会对此以实验说明对本方法采用“小样本”训练和降维处理的可行性.

2.2 基于TCM-KNN算法的异常检测框架

本节基于上述改进的 TCM-KNN 算法构建了一个网络异常检测框架,旨在说明如何在实际中使用该方法进行异常检测.如图 1 所示,在示意图的下半部分,训练阶段(training phase)为了建立实际应用中正常的训练集的工作主要包括如下几部分:

- 正常数据收集(normal data collection):从网络中采集能够反映应用正常行为的数据,不包括异常行为数据.用于构建检测阶段中用于异常检测的正常行为数据集;
- 正常数据选择(data selection):为了降低本文所述方法在距离计算中由于正常训练集规模过大而导致的计算量庞大的问题,需要根据实际应用对正常的网络数据进行有针对性的采样,比如:根据协议(Http,FTP,SMTP 等)的不同选取代表性的流量;根据一定的时间间隔来选择统计信息(例如:2 秒内网络中 SYN 和 ACK/SYN 包的比率)等方法,而不是对所有时刻的所有流量都进行相应的元信息和统计记录;
- 特征选择及向量化(feature selection and vectorlize):为了避免距离计算中有可能遇到的“维灾难”问题,需要对设定的特征进行特征选择.然后,将所有正常数据映射为表征数据的特征向量加入正常训练集,

作为异常检测的基准库(baseline).

图 1 上半部分给出了检测阶段(detection phase)的主要工作:

- 数据采集(data collection):网络数据采集模块主要负责从所监控的目标网络段中收集原始的网络数据.主要包括网络数据在链路层、网络层和传输层的数据元信息等;
- 数据预处理(data preprocess):由于改进的 TCM-KNN 算法处理的对象是由特征向量表征的独立点,并且用于异常检测的正常训练集也是由许多独立同分布的点所构成的,所以,该模块负责将从数据采集模块中收集来的原始数据按照实际应用中事先定义好的特征,处理成由这些特征组成的特征向量,交由后续模块处理;
- 异常检测(anomaly detection based on TCM-KNN):该模块则使用改进的 TCM-KNN 算法,根据正常训练集中的所有特征向量对新来的数据一个一个地进行判定.

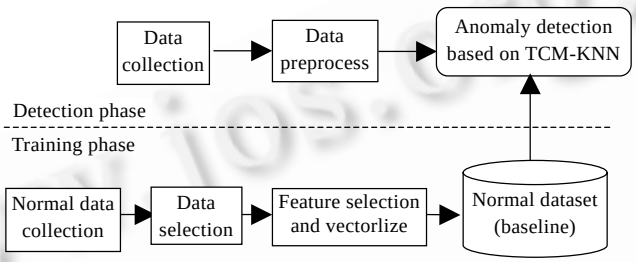


Fig.1 An anomaly detection framework based on TCM-KNN

图 1 基于 TCM-KNN 算法的异常检测框架

3 实验及其结果分析

本节我们将对所提出的异常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,本节采用研究领域共同认可及广泛使用的基准评测数据集 KDD Cup 1999 进行测试.由于该数据集的完备性,其实质上已经完成了图 1 中所述的训练阶段的大部分工作.我们只需要在实验中对数据进行相应的提取和特征选择工作,本节后面将会作详细介绍.

在实验中,首先,我们将本文所述方法与无指导异常检测领域较为著名的 Cluster 方法、*K*-近邻方法(*K*-nearest neighbors,简称 KNN)和 one-class SVM 方法以及常用的基于神经网络和基于超球面(quarter-sphere)空间划分的 SVM 方法的异常检测效果进行了比较;然后,我们评估了本文所述方法在训练集中的正常数据存在“噪音”数据(攻击数据)干扰下的性能;最后,我们测试了该方法在降低运算时间开销(采用“小样本”的正常训练集进行训练以及对训练样本实行降维处理)情况下的性能.在实验中,本文采用的评价指标为国际上通用的检测率(true positive rate,简称 TP)和误报率(false positive rate,简称 FP)指标.

3.1 实验数据集

本文采用的 KDD Cup 1999 数据集包括大约 4 900 000 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,包含有 5 类数据:DoS,Probe,R2L,U2R 这 4 类攻击数据(共包含 24 种攻击类型)以及正常数据.

为了进行上述几个实验,我们将KDD Cup 1999 的数据集进行了提取.对于本文所述方法与无指导异常检测方法的对比实验,我们从随机数据集中提取了 196 485 条正常数据和 2 050 条攻击数据(包括上述 4 类攻击),攻击数据占整个数据集的 1%,这主要是为了满足无指导异常检测方法的需求[7];对于后续对比本文所述方法在有“噪音”数据干扰下以及降低运算时间开销情况下和正常情况下的性能,我们对上述随机提取的数据集进行了再处理,具体数据组成将在相应实验中加以详述.

3.2 数据预处理

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,所以需要对它们进行处理.归一化处理的方法步骤为:

首先,分别计算出训练样本每个特征属性的均值和标准差:

$$mean[j]=\frac{1}{n}\sum_{i=1}^{n}instance_i[j] \tag{5}$$

$$standard[j]=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(instance_i[j]-mean[j])^2} \tag{6}$$

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:

$$newinstance[j]=\frac{instance[j]-mean[j]}{standard[j]} \tag{7}$$

可见,式(7)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.

3.3 与相关工作的对比实验

对比本文方法与Columbia大学的Eskin[7]等人提出的 3 种著名的无指导异常检测算法的性能,我们在实验中使用了随机提取出来的196 485条正常数据和2 050条攻击数据,采用十折交叉验证(ten fold cross-validation)的方法得到了如图2所示的ROC(receiver operating characteristic)曲线示意图(图中的每条ROC曲线通过调整相应算法的阈值得到).不难看出,本文方法具有很高的检测率,同时保证了相对较低的误报率,效果非常理想.

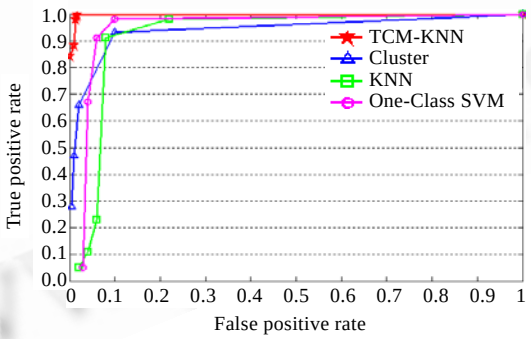


Fig.2 ROC curves for contrast experiment

图2 对比实验的ROC曲线示意图

同时,为了更有力地说明本文方法的有效性,我们采用上述的 2 050 条攻击数据作为独立测试集,将本文方法对几类具体攻击的检测效果与效果较好的基于神经网络(neural networks)[11,12]和基于超球面(quarter-sphere)空间划分方法的SVM检测方法[13]的检测效果进行了详细对比,结果见表 1.结果表明,本文所述方法在几类具体攻击的检测效果上均明显优于其他方法,充分说明了本文所述方法的有效性.然而,在U2R和R2L这两类攻击的检测效果上还有待提高.这主要是因为这类攻击在行为上与正常的行为有极大的相似性,非常难以分辨,因而检测率并不是十分理想,这项研究将成为我们下一步检测工作的重点.

Table 1 Comparison of true positive rate for each type of attack

表 1 各攻击类型的检测率比较

	TCM-KNN	Neural networks	Quarter-Sphere SVM
DoS	1230/1230(100%)	1123/1230(91.3%)	1187/1230(96.5%)
Probe	708/710(99.7%)	623/710(87.7%)	641/710(90.3%)
U2R	52/58(89.7%)	37/58(63.8%)	39/58(67.2%)
R2L	50/52(96.2%)	41/52(78.8%)	48/52(92.3%)

3.4 “噪音”干扰环境下的算法性能测试

由于在复杂的网络环境中很难保证异常检测所需要的正常训练集是完全“干净”的,因此,在这种环境下,算法的健壮性尤为重要,否则将会导致误报和漏报.因此接下来,我们测试了在训练集存在少量“噪音”干扰情况下与正常情况下的检测性能差异.由于上一节所述对比实验本文方法采用的是含有少量“噪音”数据的训练集(为了满足无指导异常检测方法的要求),因而在本实验中,我们将 196 485 条正常数据和 2 050 条攻击数据分为两部分:106 485 条正常数据作为训练集,90 000 条正常数据和 2 050 条攻击数据作为独立的测试集,以验证本算法的性能.如果测试结果与上一节中得到的结果相差不大,则能充分说明本文所述方法在有“噪音”干扰环境中的性能所受影响不大,体现了其健壮性.实验结果得到了如表 2 所示的实验数据.从表中可以看到,本文方法在正常训练数据不纯情况下的检测性能与正常情况下相差无几(检测率和误报率非常接近),因而表明了其较强的抗干扰能力.相比之下,神经网络方法在正常训练集中混杂有“噪音”数据(攻击数据)下的检测率则有大幅度的下降.这里需要注意的是:由于 Eskin 等人提出的 3 种著名的无指导异常检测算法以及 quarter-sphere SVM 方法本身需要训练集中有少量攻击数据进行训练,在本实验中不具可比性,所以此处并未对它们进行比较.

Table 2 Results for experiments using clean and unclean data

表 2 采用干净和有噪音数据的实验结果

	Clean dataset (%)		Unclean dataset (%)	
	TP	FP	TP	FP
TCM-KNN	TP=99.44	FP=1.74	TP=99.36	FP=1.72
Neural networks	TP=85.78	FP=8.73	TP=76.68	FP=5.89

3.5 采用“小样本”和降维处理后的算法性能测试

由于本文所述方法在距离计算中可能遇到由于训练集数量大、特征数目多而导致的“维灾难(curse of dimensionality)”及实用性不高的问题,在实践中,我们通常需要对训练集数量进行限制和提取少量重要特征.那么,我们就需要对本方法在采用“小样本”训练和对特征向量进行降维处理情况下的异常检测性能进行测试.

我们从事先提取的 196 485 条正常数据中再次采样,得到 18 369 条数据作为训练集,剩余的正常数据和攻击数据作为测试集.然后,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,选取了包括 dst_host_rerror_rate,count,src_bytes 等在内的 6 个特征,从而对这些特征向量实行降维处理.本文方法使用处理后的训练集进行独立训练和测试后的结果与未经过处理前的结果进行对比,见表 3 和表 4.可以看到,TCM-KNN 方法相对其他几种方法来说,采用“小样本”训练后性能影响程度非常小.而对于经过特征选择处理后的效果,各种方法的性能影响都非常小.这些都充分说明了优化工作对于本文所述方法的有效性和可行性.

Table 3 Experimental results after reducing the scale of training dataset (adopting small training dataset)

表 3 降低训练集规模后(采用“小样本”训练集)的实验结果

	Original dataset (%)		Dataset after scale reducing (%)	
	TP	FP	TP	FP
TCM-KNN	99.44	1.74	99.32	1.81
Neural networks	85.78	8.73	79.68	5.89
Quarter-Sphere SVM	90.28	7.63	84.77	4.73
Cluster	92.78	10.06	85.67	6.37
KNN	91.23	7.87	85.27	5.38
One-Class SVM	98.32	10.08	93.77	4.73

Table 4 Experimental results after feature selection

表 4 经过特征选择后的实验结果

	Original dataset (%)		Dataset after feature selection (%)	
	TP	FP	TP	FP
TCM-KNN	99.44	1.74	99.42	1.37
Neural networks	85.78	8.73	85.68	8.79
Quarter-Sphere SVM	90.28	7.63	90.77	7.73
Cluster	92.78	10.06	91.87	9.28
KNN	91.23	7.87	91.35	7.06
One-Class SVM	98.32	10.08	98.07	9.83

3.6 实验结果分析

上述大量实验结果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,其检测正确率要略高于one-class SVM算法(99.44%相对于98%),误报率则有显著降低(1.74%相对于10%).实验还表明,即便本方法的训练集有少量“噪音”数据(攻击数据),但它仍能保证较高的检测率和较低的误报率.当然,“噪音”数据的比例应当是相对较小的,一般只占整个训练集的 1%~1.5%左右,这在现实情况中也是合理和完全能够保证的;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”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.

除了本文比较了的国际上著名的异常检测方法以外,国内在入侵检测领域更多地关注有指导的检测方法,如基于多分类支持向量机的网络入侵检测方法[14]等.本文所述方法相对于它们来说最大的优点是:它并不需要对攻击方式的建模和学习(实践中也很难较全面地获得这些攻击数据),只需要“相对干净”的正常数据进行学习和检测,因而在实践中更为实用,本文的实验结果也证明了其高效性和可行性.另外,还与最近提出的采用模糊理论[15]和D-S证据理论进行异常检测[16]的方法进行了比较,根据本文对各个攻击类型的检测结果来看(见表 1),本文所述方法的检测率较之这几种方法要高.

从本质上来说,本文所述方法是通过正常训练集的所有特征向量来对新到来的数据进行具有较高置信度的评估,因而具有较高的区分度.只要正常集的数据足够可靠,则其检测率的正确性就有较高的保障,这也是本方法相对传统机器学习方法的一个较大优点.

4 结 论

TCM-KNN 算法在模式识别、离群点检测、欺诈检测领域取得了较好的成效,本文创造性地将其应用于入侵检测的异常检测领域,并根据异常检测的实际特点,对该算法进行了改进,提出了一种基于改进算法的异常检测新方法.在经典 KDD Cup 1999 数据集上的大量实验表明,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异常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.主要包括:如何有效地从网络数据中提取一定量的精简特征,避免其冗余而带来的“维灾难”;如何针对实际应用提炼少量用于正常建模的训练集,以降低算法运算的时间复杂度;如何对本文方法进行改进,结合其他机器学习算法,进一步提高对攻击行为和效果相对“模糊”的 U2R,R2L 攻击方式的检测率等.

References:

[1] Bykova M, Ostermann S, Tjaden B. Detecting network intrusions via a statistical analysis of network packet characteristics. In: Proc. of the 33rd Southeastern Symp. on System Theory. 2001. 309–314. http://masaka.cs.ohiou.edu/papers/ssst2001.pdf

[2] Denning DE. An intrusion-detection model. IEEE Trans. on Software Engineering, 1987,13(2):222–232.

[3] Lee W, Stolfo SJ. A framework for constructing features and models for intrusion detection systems. ACM Trans. on Information and System Security, 2000,3(4):227–261.

[4] Valdes A, Skinner K. Adaptive, model-based monitoring for cyber attack detection. In: Debar H, Mé L, Wu SF, eds. Proc. of the 3rd Int'l Workshop on the Recent Advances in Intrusion Detection (RAID 2000). LNCS 1907, Heidelberg: Springer-Verlag, 2000. 80−93.

[5] Aickelin U, Greensmith J, Twycross J. Immune system approaches to intrusion detection—A review. In: Nicosia G, *et al*., eds. Proc. of the 3rd Int'l Conf. on Artificial Immune Systems. LNCS 3239, Heidelberg: Springer-Verlag, 2004. 316−329.

[6] Lee W, Stolfo SJ. A Data mining framework for building intrusion detection models. In: Gong L, Reiter MK, eds. Proc. of the '99 IEEE Symp. on Security and Privacy. Oakland: IEEE Computer Society Press, 1999. 120−132.

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

[8] Proedru K, Nouretdinov I, Vovk V, Gammerman A. Transductive confidence machine for pattern recognition. In: Elomaa T, *et al*., eds. Proc. of the 13th European Conf. on Machine Learning. LNAI 2430, Heidelberg: Springer-Verlag, 2002. 381−390.

[9] Barbarà D, Domeniconi C, Rogers JP. Detecting outliers using transduction and statistical testing. In: Ungar L, Craven M, Gunopulos D, Eliassi-Rad T, eds. Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2006. 55−64.

[10] Angiulli F, Pizzuti C. Outlier mining in large high-dimensional data sets. IEEE Trans. on Knowledge and Data Engineering, 2005, 17(2):203−215.

[11] Ghosh AK, Schwartzbard A. A study in using neural networks for anomaly and misuse detection. In: Proc. of the 8th USENIX Security Symp. 1999. 141−151. http://www.usenix.org/events/sec99/full_papers/ghosh/ghosh.ps

[12] Manikopoulos C, Papavassiliou S. Network intrusion and fault detection: A statistical anomaly approach. IEEE Communications Magazine, 2002,40(10):76−82.

[13] Laskov P, Schafer C, Kotenko I. Intrusion detection in unlabeled data with quarter-sphere support vector machines. In: Proc. of the Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA 2004). 2004. 71−82. http://www2.informatik.hu-berlin.de/wm/journalclub/dimva2004.pdf

[14] Li KL, Huang HK, Tian SF, Liu ZP, Liu ZQ. Fuzzy multi-class support vector machine and application in intrusion detection. Chinese Journal of Computers, 2005,28(2):274−280 (in Chinese with English abstract).

[15] Zhang J, Gong J. An anomaly detection method based on fuzzy judgment. Journal of Computer Research and Development, 2003, 40(6):776−783 (in Chinese with English abstract).

[16] Zhuge JW, Wang DW, Chen Y, Ye ZY, Zou W. A network anomaly detector based on the D-S evidence theory. Journal of Software, 2006,17(3):463−471 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/463.htm

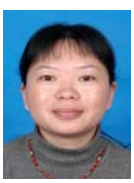
附中文参考文献:

[14] 李昆仑,黄厚宽,田盛丰,刘振鹏,刘志强.模糊多类支持向量机及其在入侵检测中的应用.计算机学报,2005,28(2):274−280.

[15] 张剑,龚俭.一种基于模糊综合评判的入侵异常检测方法.计算机研究与发展,2003,40(6):776−783.

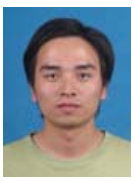
[16] 诸葛建伟,王大为,陈昱,叶志远,邹维.基于 D-S 证据理论的网络异常检测方法.软件学报,2006,17(3):463−471. http://www.jos.org.cn/1000-9825/17/463.htm

李洋(1978－),男,湖南湘潭人,博士生,主要研究领域为计算机网络与信息安全,基于数据挖掘和机器学习方法的网络异常检测技术.



郭莉(1969－),女,在读博士生,研究员,主要研究领域为计算机网络与信息安全.

方滨兴(1960－),男,博士,教授,博士生导师,中国工程院院士,CCF 高级会员,主要研究领域为网络信息安全,并行计算.



陈友(1981－),男,博士,主要研究领域为计算机网络安全,特征选择算法及其在入侵检测领域的应用.