·计算机工程与应用·

基于改进的TCM-KNN DoS检测算法

张凤荔，王 丹，赵永亮，冯 波，王 勇

(电子科技大学计算机科学与工程学院 成都 611731)

【摘要】由于实现方式简单、攻击形式多样、威胁范围广、不易防御和区分，拒绝服务(DoS)攻击已经成为网络的最主要安全威胁之一。该文提出了一种ITCM-KNN算法，在此基础上建立了DoS检测框架。使用标准数据集KDD Cup 1999进行算法验证和分析实验。采用基于信息增益算法选择了5个特征，在保证高检测效果的同时减少了特征的维数。该算法不需要对攻击进行学习和建模，使用少量的正常样本作为训练集，提高了检测性能。实验结果表明，改进的TCM-KNN算法检测率高于SVM等算法，达到99.99%。

关 键 词 拒绝服务攻击; 拒绝服务攻击检测; TCM-KNN算法

中图分类号 TP393.08 文献标志码 A doi:10.3969/j.issn.1001-0548.2014.01.013

Algorithm Based on ITCM-KNN for Denial of Service Detection

ZHANG Feng-li, WANG Dan, ZHAO Yong-liang, FENG Bo, and WANG Yong

(School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731)

Abstract Because of the simplicity of the implementation, various attacking forms, destructivity, and difficulty of filtering out, DoS has become one of the most serious security threats to the Internet. In this paper, we propose an improved transductive confidence machines for k-nearest neighbors (ITCM-KNN) algorithm and establish a framework for DoS detection. Evaluation and experiments of the algorithm are based on the standard dataset KDD Cup 1999 with 5 selected features using the information gain algorithm, which can ensure high detection rate while reducing the dimension of the features. The proposed algorithm does not need learning and modeling attacks. It only needs a small number of samples as training data set. The comparison results show that the true positive rate (TP) of the improved TCM-KNN algorithm is about 99.99%, which is higher than other detection algorithm such as SVM.

Key words DoS attack; DoS detection; improved transductive confidence machines for K-nearest neighbors

随着网络技术和网络应用的快速发展，网络所面临的威胁也越来越多，网络安全显得尤为重要。拒绝服务(denial of service, DoS)攻击是众多网络攻击中的一种。由于实现方式简单、攻击形式多样、威胁范围广、不易防御和区分，DoS攻击越来越成为网络最主要的威胁之一[1]。而且随着攻击技术和攻击工具的改进，DoS攻击也在不断地使用新技术和网络环境进行改进，分布式拒绝服务(distributed denial of service, DDoS)攻击[2]就是攻击者利用入侵过的或可间接利用的主机来发起的。

DoS攻击是一种破坏网络服务的黑客方式，具体的表现方式有：攻击者制造大流量的无用数据，使被攻击主机的网络拥塞；攻击者利用被攻击机提供服务或传输协议上处理重复连接或本身实现的缺陷，反复发出攻击性的重复服务请求或畸形的攻击数据，耗尽被攻击主机的资源(包括CPU、内存、网络带宽等)，使主机挂起甚至死机，从而无法处理合法用户的指令。

如果能够及时有效地检测出DoS攻击，就可以避免不必要的系统资源、网络带宽的浪费，从而更好地为系统或网络用户提供所需要的服务。

当前，根据不同的攻击检测手段，可以将DoS检测分为异常检测、特征检测和第三方检测。异常检测在统计的基础上建立网络数据的正常模式，如果待检测的数据偏离该模式即判定攻击发生，该检测机制可以发现未出现过的攻击。特征检测是建立

收稿日期：2012－01－05； 修回日期：2013－04－07

基金项目：国家自然科学基金(61133016)；国家自然科学基金、中物院联合基金(U1230106)；工信部科技重大专项(2011ZX03002-002-03)；国家信息安全计划(2010A14)；电子发展基金(信部运(2007)329)；四川省科技支撑计划(M110106012009FZ0148)

作者简介：张凤荔(1963－)，女，博士，教授，主要从事网络安全、移动数据管理及其应用等方面的研究.

一个所有已知的攻击方式的特征数据库，如果待检测数据与数据库中的某些特征参数匹配即判定攻击发生，该方法可有效而准确地检测已知的攻击，但却无法检测新的攻击。第三方检测则是利用外部信息来确定攻击，而不是通过自身来发现攻击。

针对DoS攻击的特征，本文提出了一种改进的TCM-KNN算法，并将其用于DoS检测。介绍了国内外关于DoS攻击检测的相关研究及传统的TCM-KNN算法，然后在此基础上根据DoS检测的特征对其进行改进，提出了一个基于改进的TCM-KNN算法的DoS攻击检测框架。采用标准数据集KDD Cup 1999进行算法验证和对比实验。

1　相关工作

文献[3]提出了用线性预测分析的方法来检测SYN Flooding攻击，利用TCP协议超时时使用的指数回退性质，对SYN包和SYN+ACK包的差别进行建模，可以成功地检测SYN Flooding攻击。由于只能检测SYN Flooding攻击，所以有一定的局限性。文献[4]运用支持向量机来对DoS攻击进行检测，并采用两种方法对DoS攻击特征的重要性进行排序。第一种是基于性能的方法，经过排序后选取19个重要特征进行检测，检测率为99.22%；第二种是SVM-Specific特征排序方法，选取了11个重要特征进行检测，检测率为99.16%。虽然经过特征选择后训练时间和测试时间有所减少，但是检测率也有所下降；同时，所选择的特征仍然过多。文献[5]将SVM方法和径向基函数神经网络方法(RBFNN)用于DoS检测。将这两种方法进行对比可知，SVM方法具有较高的检测率，但是在检测新的未知的攻击时，SVM方法需要的时间比RBFNN方法长。文献[6]建立了一个称为TOPS的系统，该系统通过启发式流量平衡来检测和过滤DoS带宽攻击。适合在路由器上使用，要将其用于不同的网络主机则还需要改进。文献[7]介绍了CIDS，利用选择的特征之间互相关性的改变来判定DoS攻击。CIDS是一种使用异常检测方法的入侵检测系统，对DoS攻击最佳的检测率为98.1%。文献[8]利用信息论的相对熵理论，针对网络流量中IP、端口等属性，分析在发生DoS攻击时造成这些属性分布特性的变化规律，计算出它们对相邻时刻网络流量序列之间相对熵值的影响，设计了基于相对熵网络DoS检测算法。从实验结果来看，该算法对DoS的检测率并不高，误报率和漏报率却比较高。相对于静态熵，文献[9]提出了一种基于活跃熵的DoS攻击检测模型。该检测模型是基于特征检测的，它利用活跃通信理论将信息熵与网络会话相关性结合起来，通过分析网络流量活跃熵值的变化实现对DoS攻击行为的检测。实验表明，该模型对DoS攻击的检测效果优于静态熵模型，但检测率也只有95.3%，误报率却高达4.7%。

本文采用了一种ITCM-KNN(improved transductive confidence machines for K-nearest neighbors)算法，该算法基于TCM-KNN [10-11]来对DoS攻击进行检测，是由Kolmogorov的算法随机性理论描述定义的具有置信判断能力的、基于置信度机制的机器学习方法，其应用领域包括模式识别、欺诈检测及“离群点”检测等。

2　算法介绍

2.1　算法基础—TCM-KNN算法

直推信度机TCM(transductive confidence machines)利用Kolmogorov算法的随机性模型，建立机器学习置信度机制，在实现测试样本类别判断的同时计算该次判断的置信度，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[10]，TCM本质上是一种在线学习算法。

TCM-KNN算法将TCM与经典的分类算法K-近邻结合起来，通过计算样本的特征向量之间的距离(欧式距离)，根据已分类的数据集对待检测样本进行分类。样本间的距离为：

$$D(i,j)=\sqrt{\sum_{n=1}^{m}(x_{in}-x_{jn})^2} \tag{1}$$

式中，$D(i,j)$表示样本i、j之间的距离；x为样本的某种特征属性；m为样本的特征维数。

在TCM-KNN算法中，为了根据已分类的类别对待检测样本进行分类，定义了奇异值来表征待检测样本与当前所有训练样本之间的差异程度。

定义 1　i表示待检测样本，y表示类别，定义i相对于y的奇异值α_{iy}为：

$$\alpha_{iy}=\frac{\sum_{j=1}^{k}D_{ij}^{y}}{\sum_{j=1}^{k}D_{ij}^{-y}} \tag{2}$$

式中，D_i^y是距离的序列，该序列是i与y中所有样本的距离；D_{ij}^y是序列D_i^y中第j个最短的距离；D_i^{-y}表示的序列是i与其他类别中(y除外)所有样本的距离；D_{ij}^{-y}是序列D_i^{-y}中第j个最短的距离；k是参考的最近

邻样本的数目。

TCM中采用的置信度机制是基于随机性检测的，对置信度的估算采用随机性检测函数来进行，定义*P*值为检测函数的值。

定义 2 待测样本*i*相对于类别*y*的*P*值：

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

式中，#是集合的“势”，表示有限集合的元素个数；α_i表示*i*的奇异值；α_j是集合中任意样本的奇异值；*j*是类别*y*中奇异值大于*i*的奇异值的样本个数；*n*是集合的元素个数；*P*值是待分类样本属于已存在的几类样本空间的概率。在待分类样本集中每个样本对应每类都有一个*P*值，*P*值可以计算为$\frac{j}{n+1}$，一次处理一个样本，*P*值的取值范围为[0,1]，*P*值越大表明*i*归属于*y*的可能性越大。

2.2 ITCM-KNN算法

将TCM-KNN算法用于网络DoS攻击检测需要进行改进， 原因在于检测DoS攻击不需要利用攻击数据建立分类，只需要从网络中提取具有正常行为的样本集，用特征向量来表示每个样本。然后判定待检测样本(也用相同的特征向量表示)相对于正常训练集，是属于正常类别还是DoS攻击类别。因此DoS攻击检测仅包含正常类别和DoS攻击类别两个类别。所以为了避免待检测样本相对于其他不必要类别的样本的奇异值的计算，需要对奇异值进行重新定义。此外，由于训练集中仅含有正常类别，这将导致在计算训练集中的训练样本的奇异值时，式(2)中的分母无法计算，从而导致*P*值无法计算，所以需要定义新的奇异值。

定义 3 *i*表示待检测样本，定义*i*相对于正常类别的奇异值α_i为：

$$\alpha_i=\sum_{j=1}^{k}D_{ij} \tag{4}$$

式中，D_{ij}是*i*与训练数据集中所有样本的距离中第*j*个最短的距离；*k*表示参考的最近邻样本数目。

这样只需要计算待检测样本相对于正常类别的奇异值即可。根据上述定义，不属于正常类别的样本的奇异值远远大于正常类别中的样本的奇异值，就可以区分DoS攻击数据与正常数据。*P*值的计算方法不变。

使用改进的TCM-KNN算法经过以下几个步骤，就可以对网络DoS攻击进行检测(假设*P*值的阈值为0.05，置信度为0.95)：

1) 给定一个由正常数据组成的训练集和一批待检测样本；

2) 对正常训练集中的每个样本，计算它们相对于其他训练样本的奇异值；

3) 对待检测样本，计算它相对于所有训练样本的奇异值；

4) 对待检测样本，计算它相对于正常训练集的*P*值；

5) 将该*P*值与阈值0.05比较，以置信度0.95进行判定待检测样本为异常或正常。

ITCM-KNN算法的伪码描述如下：

① 参数描述：*k*表示所选取的最邻近样本数，*m*表示训练集样本数，τ表示置信度阈值。

② 算法过程：

```
Input:  r ——待检测样本
for i=1 to m
{
   分别按式(1)和式(4)计算训练集中的每个样本i相对于其他训练样本的距离D_i和奇异值α_i；
 }
按式(4)计算r的奇异值α；
按式(3)计算r的P值；
if (P>τ)
   以(1–τ)的置信度判定待检测样本r为正常，
   Output “正常”;
else
   以(1–τ)的置信度判定待检测样本r为异常，
   Output “DoS”。
```

2.3 基于信息增益的特征选择方法

为了避免在计算样本间的距离时产生维灾难，减小特征空间，提高算法的检测效率，本文采用基于信息增益的方法进行特征选择，选出更有效的、更能代表DoS攻击的特征[12]。

特征选择有过滤器(Filter)和封装器(Wrapper)两种基本模式，封装器模式的计算复杂度高、速度慢、计算效率低。本文选择过滤器模式中的信息增益方法来对DoS攻击的特征进行选择。

熵表示任何一种能量在空间中分布的均匀程度，能量分布越均匀、越不确定，熵就越大。Shannon将熵运用于信息处理，提出信息熵的概念。信息熵是信息的量化度量，用来衡量一个随机变量取值的不确定程度[13-14]。

假设*S*是一个包含有*m*个类的样本集，类C_i中含

有s_i个样本。样本集S中总共有s个样本，则信息熵定义为：

$$H(C)=-\sum_{i=1}^{m}P(C_i)\log_2 P(C_i) \tag{5}$$

式中，$P(C_i)=\frac{s_i}{s}$表示任意样本属于类C_i的概率；$H(C)$表示将样本集C中的样本分成m个类的不确定程度。

如果给定样本集S中的样本的某一特征(或属性)F，再根据该特征对S进行划分。那么，在给定特征F的条件下对S进行分类的不确定程度可以表示为条件熵$H(C|F)$。设特征F有v个不同的值$\{F_1,F_2,\cdots,F_v\}$，则可以用特征F把样本集S划分为v个子集$\{S_1,S_2,\cdots,S_v\}$，其中S_j包含S中特征F取值为F_j的样本，有：

$$H(C\mid F)=-\sum_{i=1}^{m}\sum_{j=1}^{v}P(F_j)P(C_i\mid F_j)\log_2 P(C_i\mid F_j)=\sum_{j=1}^{v}P(F_j)H(C\mid F=F_j) \tag{6}$$

式中，$P(F_j)$为第j个子集S_j的权值，即特征F_j出现的概率。设s_{ij}表示S_j中包含的属于类C_i的样本个数，则$P(F_j)=\frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s}$。$H(C|F=F_j)$表示特征$F$取值为$F_j$的条件熵：

$$H(C\mid F=F_j)=-\sum_{i=1}^{m}P(C_i\mid F=F_j)\log_2 P(C_i\mid F=F_j)=-\sum_{i=1}^{m}P_{ij}\log_2 P_{ij} \tag{7}$$

式中，$P_{ij}=\frac{s_{ij}}{s_j}$表示子集$S_j$中的样本属于类$C_i$的概率。

将$P(F_j)$和$H(C|F=F_j)$代入式(6)，则有：

$$H(C\mid F)=\sum_{j=1}^{v}\frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s}\left(-\sum_{i=1}^{m}\frac{s_{ij}}{s_j}\log_2\frac{s_{ij}}{s_j}\right) \tag{8}$$

定义特征F的信息增益为：

$$G(F)=H(C)-H(C\mid F) \tag{9}$$

$G(F)$表示在给定特征F的条件下，对样本集S进行划分可以减少的不确定度，即特征F为划分S所提供的信息量。$G(F)$的值越大，则表示特征F对划分S的贡献越大，特征F就越具有区分度。

3　DoS检测框架

根据上述ITCM-KNN算法，图1给出了针对网络DoS攻击的检测框架，该检测框架主要包括两个部分，训练阶段主要包括以下几个模块：

1) 数据预处理：对训练数据进行预处理[15]，主要是进行数值化和归一化处理，以进行距离计算。

2) 特征选择：为了避免计算距离时出现维灾难，需要对训练样本的特征进行选择。

3) 训练模块：计算每个训练样本相对于其他训练样本的距离和奇异值。

检测阶段包括以下几个模块：

1) 数据预处理：对待检测样本进行与训练样本相同的预处理。

2) 特征选择：从表示待检测样本的特征向量中提取与训练样本相同的特征。

3) 基于TCM-KNN的DoS检测：计算待检测样本相对于各个训练样本的距离和奇异值，计算P值并进行判定。

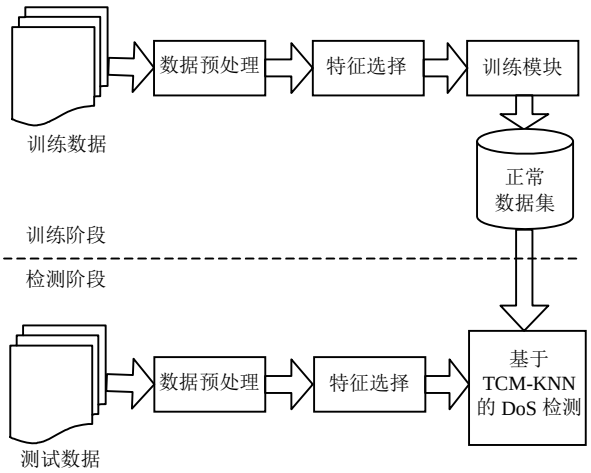

图1　基于TCM-KNN算法的DoS检测框架

4　实验分析

本文的实验基于以上检测框架对所提出的检测算法的有效性进行验证，并且将其与其他三种同类的检测方法(SVM、KNN和Naïve Bayes)进行比较，实验数据集采用基准评测数据集KDD Cup 1999[10]，评价指标为检测率 (true positive rate, TP)和误报率(false positive rate, FP)。

本文实验的硬件环境为：CPU，Pentium(R) Dual-Core E5200；RAM，2G，HardDisk，320 G。操作系统为Windows7旗舰版。此外还借助了Matlab和Weka工具。

KDD Cup 1999数据集[16-17]是MIT的林肯实验室对外公布的用于对入侵检测系统进行实验的官方数据集，是从军方网络环境中模拟攻击所得的原始网络数据中提取的，包含41个特征。该数据集包含大约4 898 431条记录和DoS、U2R、R2L和Probe四大类攻击数据和正常数据。

由于整个数据集的数据量太大，该数据集的创

造者们挑选了其中的10%进行了类别标注。这10%的数据中总共有494 021条记录，其中正常数据有97 278条，剩下的异常数据中含有DoS数据391 458条。本文仅将其中含有的正常数据和DoS数据提取出来作为训练集。另外，使用KDD Cup 1999提供的标准测试集来进行测试，该测试集中共有311 029条记录，正常数据有60 593条(占19.48%)，余下的异常数据中包含229 853条DoS数据。同样仅提取其中的正常数据和DoS数据作为测试集。样本的分布情况参照文献[16]。

本文的实验使用了正常类别中的1 000条作为TCM-KNN算法的训练集，另外，为其他三种算法(SVM、KNN、Naïve Bayes)提供了含有少量(占训练样本的1%)DoS攻击数据的训练集。测试集包含了70 000条数据，其中正常数据占80%、DoS攻击数据占20%。

4.1 数据预处理和特征选择

由于特征向量间的类型和取值范围的不同，计算特征向量间的欧氏距离会出现相互影响的情况，需要对特征向量中取值为数值型的数据进行归一化处理，取值为字符型的数据需要根据它们各自出现的概率进行了数值化处理，本文运用Matlab中的mapminmax( )函数来对样本(包括训练和测试样本)进行归一化处理，将样本的不同属性值范围映射到标准的取值空间[0,1]。

为了便于使用Weka平台中的信息增益方法进行特征选择，将10%数据集中的正常数据标记为1，DoS数据标记为0。经过特征选择，得到5个信息增益值较大的特征，如表1所示。

表1 特征选择结果

特征	描述
count	在过去的两秒内，和当前连接主机相同的连接数
src_bytes	从源地址到目的地址传输的字节数
dst_bytes	从目的地址到源地址传输的字节数
logged_in	1表示注册成功，0表示注册失败
dst_host_same_src_port_rate	目标主机相同的，具有相同源地址的连接百分比

4.2 结果分析

为了与其他三种同类的检测方法(SVM、KNN和Naïve Bayes)进行对比，为这三种方法提供了含有少量(占训练样本的1%)DoS攻击数据的训练集。通过选取不同的参数进行了大量的实验，选取效果较好的几组数据(如表2所示)，得到如图2所示的ROC曲线。

由实验可知，改进的TCM-KNN算法的误报率比SVM算法的误报率稍高一些(两种算法均在最佳情况下：2.23%相对2.1%)，检测率则有一定的提高(两种算法均在最佳情况下：99.99%相对95.3%)。改进的TCM-KNN算法的误报率比KNN算法的误报率稍高一些(两种算法均在最佳情况下：2.23%相对1.8%)，检测率则有一定提高(两种算法均在最佳情况下：99.99%相对96%)。此外，当k值很大时，KNN算法的效率会变得极低。改进的TCM-KNN算法的检测率和误报率比Naive Bayes算法的检测率和误报率都好(两种算法均在最佳情况下，检测率：99.99%相对97.1%，误报率：2.23%相对2.7%)。

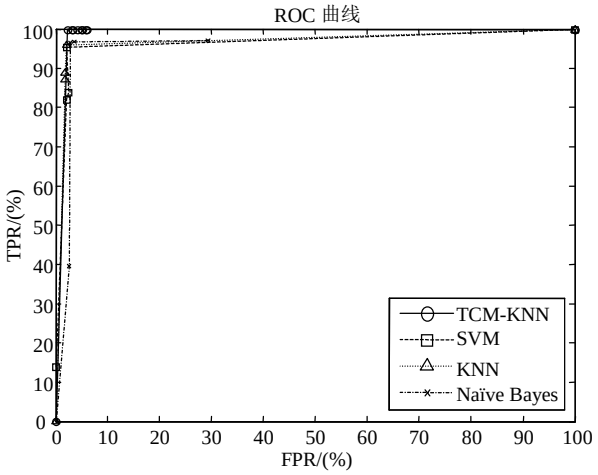

图2 对比实验的ROC曲线图

表2 图2中的部分测试数据结果

算法	检测率(TP)/(%)	误报率(FP)/(%)
	82.1	2.1
SVM	95.3	2.1
	83.7	2.4
	96	2.1
KNN	87.1	1.8
	88.9	1.8
	96.8	3.6
Naïve Bayes	96.6	2.8
	39.5	2.7
	99.99	2.23
改进的TCM-KNN	99.99	2.26
	99.99	2.3

5 结 论

本文根据DoS攻击的实际特点，对TCM-KNN算法进行了改进，提出了一个基于该改进算法的网络DoS检测框架。采用标准数据集KDD Cup 1999进行了算法验证和对比实验，证明将改进的TCM-KNN算法用于DoS检测的可行性和高效性。实验结果显示了改进的TCM-KNN算法对检测DoS攻击的有效性。对比分析结果显示，改进的TCM-KNN算法检测率高于SVM等算法，达到99.99%。SVM等算法需要大量的含有已标记的攻击数据的训练样本来对攻击进行学习和建模，而在真实的网络环境中很难得到

大量全面已标记的攻击数据。相比而言，改进的TCM-KNN算法不需要对攻击进行学习和建模，因而不需要大量的训练样本，在实践中更为实用。

本文给出的方法在不同的应用环境中需要根据具体需求进行不同改进。如何进一步降低误报率；如何运用其他更有效的方法进行特征选择，如何进行样本选择，选择出比较具有代表性的训练样本进行训练，如何将其运用于其他攻击的检测等，是未来需要进一步完成的工作。

参 考 文 献

[1] LIU Wen-tao. Research on DoS attack and detection programming[C]//2009 3rd International symposium on Intelligent Information Technology Application. Nanchang, China: IEEE, 2009: 207-210.

[2] HU Liang, BI Xiao-ming. Research of DDoS attack mechanism and its defense frame[C]//2011 3rd IEEE International Conference on Computer Reseach and Development. Shanghai, China: IEEE, 2011: 440-442.

[3] DIVAKARAN D M, MURTHY H A, GONSALVES T A. Detection of SYN flooding attacks using linear prediction analysis[C]//Proceedings of 2006 IEEE International Conference on Networks. Singapore: Saira Kuttan Publication Chair, 2006: 218-224.

[4] MUKKAMALA S, SUNG A H. Detecting denial of service attacks using support vector machines[C]//Proceedings of the 12th IEEE International Conference on Fuzzy Systems. Louis, Missouri, USA: IEEE, 2003: 1231-1236.

[5] TSANG G C Y, CHAN P P K, YEUNG D S, et al. Denial of service detection by support vector machines and radial-basis function neural nework[C]//Proceedings of 2004 International Conference on Machine Learning and Cybernetics. Shanghai, China: IEEE, 2004: 4263-4268.

[6] ABDELSAYED S, GLIMSHOLT D. An efficient filter for denial-of-service bandwidth attcks[C]//GLOBECOM'03, IEEE Global Telecommunications Conference. San Francisco, CA, USA: IEEE, 2003: 1353-1357.

[7] ZHANG Zheng, MANIKOPOULOS C N. Detecting denial-of-service attacks through feature cross-correlation[C]//2004 IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communication. Princeton, New Jersey: IEEE, 2004: 67-70.

[8] 李涵秋, 马艳, 雷磊. 基于相对熵理论的网络DoS攻击检测算法[J]. 电讯技术, 2011, 51(3): 89-92.
LI Han-qiu, MA Yan, LEI Lei. DoS attack detection based on relative entropy theory[J]. Telecommunication Engineering, 2011, 51(3): 89-92.

[9] 刘衍珩, 付枫, 朱建启, 等. 基于活跃熵的DoS攻击检测模型[J]. 吉林大学学报(工学版), 2011, 41(4): 1059-1063.
LIU Yan-heng , FU Feng , ZHU Jian-qi, et al. DoS detection model base on alive entropy[J]. Journal o f Jilin University (Engineering and Technology), 2011, 41(4): 1059-1063.

[10] 李洋, 方滨兴, 郭莉, 等. 基于直推式方法的网络异常检测方法[J]. 软件学报, 2007, 18(10): 2595-2604.
LI Yang, FANG Bin-xing, GUO Li, et al. A network anomaly detection method based on transduction scheme[J]. Journal of Software, 2007, 18(10): 2595-2604.

[11] 李洋, 郭莉, 陆天波, 等. TCM-KNN 网络异常检测算法优化研究[J]. 通信学报, 2009, 30(7): 13-19.
LI Yang, GUO Li, LU Tian-bo, et al. Research on performance optimizations for TCM-KNN network anomaly detection algorithm[J]. Journal on Communications, 2009, 30(7): 13-19.

[12] FARAOUN K M, RABHI A. Data dimensionality reduction based on genetic selection of feature subsets[J]. INFOCOM-Journal of Computer Science, 2007, 6(2): 9-19.

[13] 贾娴, 刘培玉, 公伟. 应用于入侵取证的改进信息增益算法[J]. 计算机应用, 2011, 31(8): 2156-2158.
JIA Xian, LIU Pei-yu, GONG Wei. Application of improved information gain algorithm in intrusion forensics[J]. Journal of Computer Applications, 2011, 31(8): 2156-2158.

[14] 刘庆和, 梁正友. 一种基于信息增益的特征优化选择方法[J]. 计算机工程与应用, 2011, 47(12): 130-132.
LIU Qing-he, LIANG Zheng-you. Optimized approach of feature selection based on information gain[J]. Computer Engineering and Applications, 2011, 47(12): 130-132.

[15] 徐震, 刘方爱, 郭胜召. 基于TCM-KNN 算法的数据预处理问题研究[J]. 山东科学, 2009, 22(1): 17-20.
XU Zhen, LIU Fang-ai, GUO Sheng-zhao. A study on the TCM-KNN algorithm based data preprocessing problem[J]. Shandong Science, 2009, 22(1): 17-20.

[16] 张新有, 曾华燊, 贾磊. 入侵检测数据集KDD CUP99研究[J]. 计算机工程与设计, 2010, 31(22): 4809-4812.
ZHANG Xin-you, ZENG Hua-shen, JIA Lei. Research of intrusion detection system dataset-KDD CUP99[J]. Computer Engineering and Design, 2010, 31(22): 4809-4812.

[17] The UCI KDD Archive, Information and Computer Science, University of California, Irvine. KDD cup 1999 data[DB/OL]. [2011–11–19]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

编 辑 黄 莘