文章编号:1001 - 9081(2010)01 - 0178 - 03

基于质心 Voronoi 图的网络异常检测算法

王　雷, 侯瀚雨
(湖南大学 软件学院,长沙 410082)
(wanglei@ hnu. cn; coolmagic1980@ 163. com)

摘　要:网络异常检测技术是入侵检测领域研究的热点内容,但由于存在着误报率较高等问题,并未在实际环境中得以大规模应用。基于质心 Voronoi 图,提出一种新的异常检测算法。在该算法中,首先利用质心 Voronoi 图来对样本数据进行聚类,然后基于聚类结果,计算出各个样本点的点密度,并以此来判断样本数据是否异常。最后,通过基于 KDD Cup 1999 数据集的实验测试,仿真结果表明,新算法在具有较低的误报率的同时,也具有良好的检测率。

关键词:聚类; 入侵检测; 误检率; 检测率; ROC
中图分类号: TP393. 08　　**文献标志码**:A

Algorithm of anomaly detection based on centroidal Voronoi diagram

WANG Lei, HOU Han-yu
(*School of Software, Hunan University, Changsha Hunan* 410082, *China*)

Abstract: Network anomaly detection has been an active research topic in the field of intrusion detection for many years. However, it has not been widely applied in practice due to high false alarm rate, etc. Based on the centroidal Voronoi diagram, a new algorithm of anomaly detection was proposed in this paper, in which the centroidal Voronoi diagram was used in the clustering of sample data first, and then the point density was computed out according to the results of clustering for each sample point, which was used to determine whether the sample data was abnormal or not. Finally, a series of experiments on well known KDD Cup 1999 dataset demonstrate that the new algorithm has low false positive rate while ensuring high detection rate.

Key words: clustering; intrusion detection; false detecting rate; detection rate; Receiver Operating Characteristic (ROC)

0　引言

随着互联网的普及和业务量的不断增长,网络的规模和传输速度也急剧增长,这使得如何在高速环境和海量数据中检测异常行为,成为入侵检测领域面临的又一大难题。网络异常会显著破坏和降低网络服务质量,而目前传统的误用检测和异常检测技术,无论在硬件设计还是在检测算法方面都难以满足对高速大规模网络进行有效检测的要求。因此,研究新的网络入侵检测算法对提高网络的可靠性和可用性具有重要意义[1-2]。

误用检测是建立在将某种模式或者特征描述方法对某种已知攻击进行表达的基础之上,然后将所监视的事件与知识库中的已知攻击模式进行匹配,当发现有匹配时则认为有入侵发生。误用检测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。而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,当存在入侵行为时,用户或系统的行为模式会发生一定程度的改变。因此,在检测入侵活动时,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。异常检测的优点是有可能检测出以前从未出现过的攻击方法,但缺点是误报率较高[3-4]。

为了进一步减低异常检测的误报率,近年来,国内外学者提出了一系列异常检测方法,大致可分为基于统计的方法[5]、基于数据挖掘的方法[6]和基于模式匹配的方法[7]等。其中,基于数据挖掘的方法由于它能从大量数据中提取人们感兴趣的、事先未知的知识和规律,而不依赖经验,且具有较高的检测性能和效率,正成为当前异常检测的主要方法之一。本文运用数据挖掘中的聚类分析方法[8],基于质心 Voronoi 图[9],提出了一种新的异常检测算法。在该算法中,首先利用质心 Voronoi 图来对样本数据进行聚类,然后基于聚类结果,计算出各个样本点的点密度,并以此来判断各个样本点是否异常。最后,通过基于 KDD Cup 1999 数据集的实验测试,验证了新算法的有效性。

1　质心 Voronoi 图

对给定的集合 $\Omega \subseteq \mathbf{R}^N$, $\{V_i\}_{i=1}^k$ 集合称为 Ω 的一个划分,当且仅当 $\forall i \neq j$, 有 $V_i \cap V_j = \varnothing$, 且 $\bigcup_{i=1}^{k} \overline{V}_i = \overline{\Omega}$, 其中 $\overline{V}_i$ 和 $\overline{\Omega}$ 分别表示集合 V_i 和 Ω 的元素个数。令 d 表示域 $\mathbf{R}^N$ 上的一个距离函数,给定点集 $\{z_i\}_{i=1}^k \subseteq \Omega$, 则点 Z_i 对应的 Voronoi 区域 $\hat{V}_i$ 定义为:

$$\hat{V}_i = \{x \in \Omega \mid d(x,z_i) < d(x,z_j);\ \forall j = 1,2,\cdots,k, j \neq i)\} \quad (1)$$

其中,点 z_i 称 Voronoi 区域 $\hat{V}_i$ 的生成元,而集合 $\{V_i\}_{i=1}^k$ 则称为 Ω 的一个 Voronoi 划分或 Voronoi 图。

给定点集 $\{v_i\}_{i=1}^k \subseteq \mathbf{R}^N$,以及域 $\mathbf{R}^N$ 上的一个距离函数 d,

收稿日期:2009 - 07 - 03 ;**修回日期**:2009 - 08 - 30。　**基金项目**:国家 863 计划项目(2006AA01Z2227)。

作者简介:王雷(1973 -),男,湖南长沙人,副教授,博士,主要研究方向:计算机网络、数据挖掘; 侯瀚雨(1982 -),男,湖南长沙人,硕士研究生, 主要研究方向:计算机网络。

则点集$\{v_i\}_{i=1}^k$中的点v^*称为点集$\{v_i\}_{i=1}^k$的质心,若v^*满足:

$$\min \sum_{i=1}^{k} d(v^*, v_i) \tag{2}$$

基于上述质心的定义,因此对给定的k个点$\{z_i\}_{i=1}^k \subseteq \Omega$,可以得到与这$k$个点对应的$k$个Voronoi区域合$\{\hat{V}_i\}_{i=1}^k$,以及与这$k$个Voronoi区域对应的$k$个质心$\{z_i^*\}_{i=1}^k$。

假定集合$\{\hat{V}_i\}_{i=1}^k$为Ω的一个Voronoi图,且每个Voronoi区域$\hat{V}_i$的生成元均为其质心,则称$\{\hat{V}_i\}_{i=1}^k$为Ω的一个质心Voronoi图。

2　基于质心 Voronoi 图的异常检测算法

在具体给出基于质心Voronoi图的异常检测算法(Anomaly Detection based on Centroidal Voronoi Diagram, ADCVD)之前,本文首先给出几个相关定义如下[8]。

定义1　点邻域。假定集合$\{\hat{V}_i\}_{i=1}^k$为Ω的一个质心Voronoi图,$\{z_i^*\}_{i=1}^k$为对应的生成元,对任意样本点q,称以q为中心,以$\bar{d}_c(q)$为半径的邻域内所包含的点集为样本点q的点邻域,记为$N(q)$。其中:

$$\bar{d}_o(q) = \frac{\sum_{i \in \hat{V}_c(q)} d(q,i)}{|\hat{V}_c(q)|} \tag{4}$$

其中,$\hat{V}_c(q)$表示点q所在的Voronoi区域。

定义2　点密度。假定集合$\{\hat{V}_i\}_{i=1}^k$为Ω的一个质心Voronoi图,$\{z_i^*\}_{i=1}^k$为对应的生成元,对任意样本点q,假定其对应的点邻域为$N(q)$,q所在的Voronoi区域为$\hat{V}_i$,$\hat{V}_i$对应的生成元为z_i^*,则定义其对应的点密度$\rho(q)$为:

$$\rho(q) = \alpha \times \frac{1}{d_1(q)} + \beta \times \frac{1}{d_2(q)} \tag{5}$$

其中,$\alpha + \beta = 1$, $\alpha \geqslant 0, \beta \geqslant 0; d_1(q) = d(z_i^*, q) + 1$, $d_2(q) = \frac{\sum_{i \in N(q)} d(q,i)}{|N(q)|} + 1$。

由以上定义1和定义2,结合前文中的质心Voronoi图的定义可知,对任意样本点p, q,若样本点q为异常点,则q距离q所在的Voronoi图的质心距离将比正常样本点p到p所在的Voronoi图的质心距离要远(即:$d_1(q)$较大),且q的点邻域中所包含的点集到q的平均距离也要比正常样本点p的点邻域中所包含的点集到p的平均距离远(即:$d_2(q)$较大),因此,由式(5)可知有$\rho(q) < \rho(p)$。所以可以用式(5)定义的点密度来作为识别异常样本点的依据。

基于以上分析,ADCVD算法可描述如下:

输入:样本点集Ω,点密度阈值ρ_0;

输出:正常样本点集Ω_1和可能的异常样本点集Ω_2。

步骤1　随机选取k个点构成点集C_1。

步骤2　以C_1作为生成元构造点集Ω的Voronoi图,利用式(2)求得对应的质心点集C_2。若$C_2 = C_1$,则转步骤3;否则,令$C_1 = C_2$,转步骤2。

步骤3　假定集合$\{\hat{V}_i\}_{i=1}^k$为Ω的一个质心Voronoi图,$\{z_i^*\}_{i=1}^k$为对应的生成元,对任意样本点q,利用式(5)求得其对应的点密度$\rho(q)$。若$\rho(q) \geqslant \rho_0$,则将点q划分到正常样本点集Ω_1;否则,将点q划分到异常样本点集Ω_2。

3　基于 ADCVD 算法的异常检测框架

本章基于第2章提出的ADCVD异常检测算法构建了一个网络异常检测框架,其结构如图1所示。

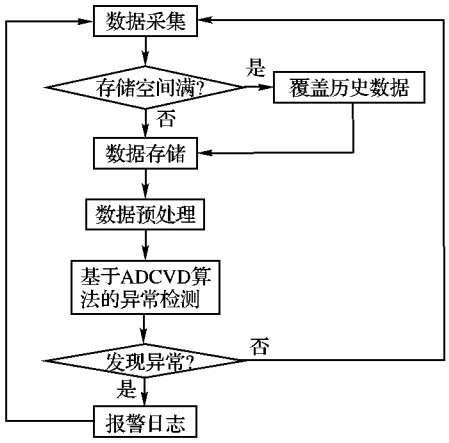

图1　基于ADCVD算法的异常检测框架

在图1中,各个主要环节的功能如下。

1)数据采集:每次从网络中采集定量数据,以供后续异常检测分析。

2)数据预处理:为了避免超高维空间的计算问题,需要对设定的特征进行特征选择来进行降维和数据格式化出来,然后,将所有数据映射到表征数据的统一的特征向量空间。

3)基于ADCVD算法的异常检测:利用第2章中的ADCVD算法对格式化数据进行异常检测,以发现异常样本点集。若存在异常样本点,则形成报警日志。

4　算法性能分析与仿真实验

KDD Cup 1999数据在每条记录的41个属性,对于t维数据空间上的任意两点x_i, x_j,本文采用Minkowski度量来定义x_i, x_j之间的距离:

$$d_p(x_i, x_j) = \left(\sum_{k=1}^{t} |x_{ik} - x_{jk}|^p \right)^{1/v} \tag{6}$$

4.1　实验数据集采集

为评价提出的网络异常检测算法,选用KDD Cup 1999网络数据集。该数据集由美国麻省理工学院Lincoln实验室仿真美国空军局域网环境而建立的测试数据集,训练数据集包含了7个星期网络流量,共有494 021个记录,其中正常记录数据97 278个,其余396 473个是异常数据。这些记录中包括了多种广泛的网络环境下的模拟入侵,其中每个实例包含42个属性而且均已标识为正常或特定的攻击类型。这些数据集共有22种攻击和正常类型.数据集中的入侵类型按攻击手段类型可大致划分为4类:拒绝服务攻击,如TCP同步报文洪泛攻击;远程权限获取(Remoting to Local, R2L),如猜测口令;各种权限提升(User to Root, U2R),如缓冲区溢出;各种端口扫描和漏洞扫描。

4.2　数据预处理

为了将ADCVD算法应用于真实网络数据中的入侵行为实际检测之中,首先需要对数据的属性值进行标准化,因为属性值之间由于采用不同的度量单位,其差别可能很大,造成对数据间距离的影响也不同,为了消除这种差别,我们采用了如下的标准化方法:对于给定的训练数据集,由式(1)、式(2)和式(3)分别求特征向量的均值m_f、计算绝对偏差的平均值S_f和标准化的特征属性值Z_{if}。

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf}) \tag{1}$$

$$S_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|) \tag{2}$$

$$Z_{if} = \frac{x_{if} - m_f}{S_f} \quad (3)$$

在对所有的数据都进行标准化理之后，即可利用统计特性将原始实例的特征属性映射到一个标准的属性空间，各维上的取值标准化到[0,1]。从而找出数据的特征表示，为聚类分析提供可靠的数据来源。

4.3 算法性能测试

实验在 Matlab6.5 环境下实现，使用从 KDD CUP 1999 网络连接数据集中选取的一个实验数据集，该数据集来源于 1998 DARPA 入侵检测评估程序，其中正常实例与异常实例之比为 9:1。对该数据集使用基于质心 Voronoi 图的异常检测算法（ADCVD）对数据集做 10 次检测，并记录每次的检测率和误检率。

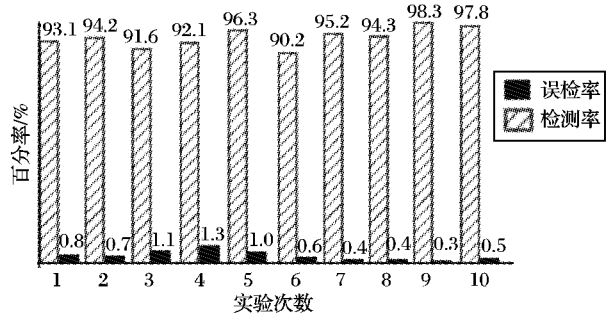

图 2 检测率和误检率检测实验结果

检测率和误检率是入侵检测系统中最重要的性能指标，在图 2 中通过检测率和误检率来表达入侵检测的实验情况。其中，检测率由算法检测出的入侵实例总数与数据集中的入侵实例总数之比得出，而误检率则是由数据集中正常实例个数和检测出的正常实例个数的差值与数据集中正常实例总数之比得出。检测率与误检率总是紧密相关，增加检测率常常要以误检率的增加为代价，而误检率偏高使系统对原本不是攻击的事件产生了错误的警报，将导致入侵检测系统的功效降低. 因此，既能增加检测率又能降低误检率是入侵检测系统最希望达到的目标。从图 2 可以看到 ADCVD 算法在具有较低的误报率同时，也具有良好的检测率。

4.4 与相关工作的对比实验

4.4.1 算法的入侵检测能力

为了验证本文的算法对入侵的检测能力，使用了随机提取出来的 200 000 条正常数据和 2 100 条攻击数据，把本文算法和文献[11]提出的算法进行比较，采用十折交叉验证的方法得到了如图 3 数据集的平均检测率（Detection Rate，DR）和误检率（False alarm Rate，FR）的 ROC（Receiver Operating Characteristic）曲线图。

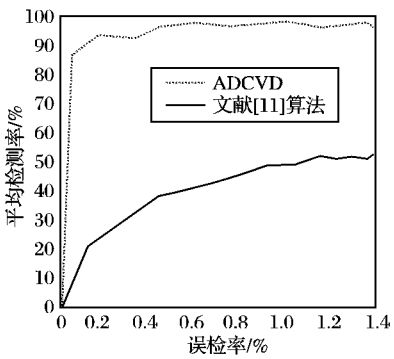

图 3 数据集的平均检测率

由图 3 中可知，在误检率较低的同时，检测率一直保持一个较高的效率；而最早将聚类分析用于入侵检测的检测系统[11]的误报率为 1% 时，检测率仅为 50%。这充分表明本文的算法对于未知入侵行为检测的可行性和有效性。

4.4.2 与 K-means 算法的比较

将 ADCVD 算法和 K-means 算法进行对比测试，测试共进行 3 次，每次测试各从实验数据中随机抽取 2600 条和2 100 条数据，分别形成训练集合 Train 和测试集合 Test。测试结果如表 1 所示。

表 1 两种算法在检测率和误检率方面的比较 %

算法	测试数据集 1		测试数据集 2		测试数据集 3	
	检测率	误检率	检测率	误检率	检测率	误检率
ADCVD	95.2	0.4	96.7	1.0	93.4	0.4
K-means	82.1	9.2	83.4	12.6	74.1	8.7

从测试结果可以看出，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。

4.4.3 与 FSVM 的比较

为了更有力地说明本文方法的有效性，本文采用之前的 2 100 条攻击数据作为独立测试集，针对 4 种类型的攻击，把本文的算法与模糊支持向量算法（Fuzzy Support Vector Machine，FSVM）[7]进行了详细对比，结果见表 2。

表 2 4 种攻击类型的检测率 %

攻击类型	ADVCD 算法	FSVM 算法	攻击类型	ADVCD 算法	FSVM 算法
Probing	96.1	92.7	U2R	83.0	72.6
Dos	89.7	83.4	R2L	51.2	54.2

从表 2 中可以看出，本文所述方法在几类具体攻击的检测效果上均优于 FSVM 算法，充分说明了本文所述方法的有效性。然而，在 R2L 攻击的检测效果上还有待提高。这主要是因为有很多 R2L 入侵是伪装合法用户身份进行攻击，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，造成了算法检测的困难，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。

5 结语

基于质心 Voronoi 图，提出了一种新的异常检测算法。在该算法中，首先利用质心 Voronoi 图来对样本数据进行聚类，然后基于聚类结果，计算出各个样本点的点密度，并以此来判断样本数据是否异常。最后，通过基于 KDD Cup 1999 数据集的实验测试，仿真结果表明，新算法在具有较低的误报率同时，也具有良好的检测率。为了进一步提高算法的性能，我们下一步拟结合一些已有的典型自动决定聚类数算法[10]，以让算法在进行质心 Voronoi 划分时更加快速收敛，从而提高算法执行的效率。

参考文献：

[1] 刘衍珩，田大新，余雪岗，等. 基于分布式学习的大规模网络入侵检测算法[J]. 软件学报，2008，19(4)：993－1003.

[2] 宿娇娜，李程，李巍，等. 基于改进 NB 分类方法的网络异常检测模型[J]. 计算机工程，2008，34(5)：148－149.

[3] 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.

[4] 王雷，林亚平，彭雅，等. 基于认知学习和最小风险的朴素贝叶斯邮件过滤算法[J]. 系统仿真学报，2004，16(3)：413－416.

（下转第 185 页）

实验数据中的 20% 作为测试数据，80% 作为训练数据。使用 LIBSVM[8] 实现支持向量机二类分类，利用训练数据分别训练 6 个分类器。经过实验测试后，确定了参数 $C = 50, 1/\sigma^2 = 10$。

本文实验主要关注正确判断某类程序类别的概率（Accuracy Rate，AR）。由前文所述根据不同 API 的 *IG* 值高低对 API 序列进行排序。从图 5 所示可以看出，选取 API 数量不同会对 4 类程序判定的准确率（Accurate Rate，AR）均造成影响。经过观察与分析，随着 API 数量的增加，AR 上升速度较快；当 API 数量接近 1 200 时，得到各类程序分类 AR 最高值；而当 API 数量继续增加时，AR 值则呈缓慢下降趋势，这是由于当所取 API 数量过大时，许多 *IG* 值较低的函数调用也被计算在内，造成区分度的降低。所以在病毒行为检测中，并非选取 API 越多越好。根据实验结果选取特征向量维数为 1 150 时四类程序检测 AR 取到近似最佳值，如表 1 所示为多类 SVM 分类方法与已有的病毒检测方法[2] 检测性能对比。

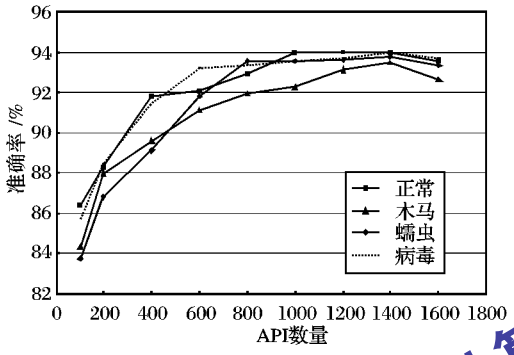

图 5　特征选择检测结果

表 1　多类 SVM 与已有的检测方法性能对比　%

检测方法		检测准确率
基于签名的特征码		49.28
RIPPER[2]		89.36
朴素贝叶斯		97.11
多重朴素贝叶斯		96.88
多类 SVM	木马	92.96
	蠕虫	93.59
	病毒	93.60

如表 1 可以发现，多类 SVM 检测方法的检测效率优于传统的基于签名的特征码检测以及 RIPPER 技术，但较贝叶斯方法准确率仍存在一定差距，这是在引入了多类分类之后，由于不同种类病毒程序确实存在一定相似之处，对其进行准确细分的难度大于仅仅区分病毒程序和正常程序的二类分类，因此这种差距应在可接受范围之内。

5　结语

由于现实环境中应用程序的 API 行为复杂多样，比之正常程序与病毒程序的 API 差异，正常程序间的 API 差异更大，这使得一般仅靠简单的 API 调用来识别病毒并不实际可行。

本文通过对大量病毒程序的功能、行为进行分析，从中寻找病毒程序的“特有”行为特征，并将这些特征放大，将不同程序的特征收集起来形成行为特征集。这些特征主要是由程序调用 API 函数所形成的。在获得了所需特征集之后，利用熵值的原理对各个特征进行处理，通过计算其信息增量来判断其对区分病毒的贡献程度，也就是“放大”的依据，从而缩小特征集维数，完成特征集的筛选。之后利用支持向量机的概念，构造超平面实现程序的二类划分的方法。针对这种方法存在的不足，进一步提出了由支持向量机推广的多类分类策略，设计了一种详细划分待测程序类别的改进方法。最后，选择了 1 000 多个程序其中包括 600 多病毒程序，共提取 2 000 多个 API，通过对这些 API 的统计、分析和计算，验证了思路的可性行，为病毒的检测提供了一种可行的方法。

参考文献：

[1] XU J Y, SUNG A H, CHAVEZ P, *et al*. Polymorphic malicious executable scanner by API sequence analysis[C]// Proceedings of the 4th International Conference on Hybrid Intelligent Systems. Washington, DC: IEEE Computer Society, 2004: 378 – 383.

[2] SCHULTZ M G, ESKIN E, ZADOK E, *et al*. Data mining methods for detection of new malicious executables[C]// Proceedings of the 2001 IEEE Symposium on Security and Privacy. Washington, DC: IEEE Computer Society, 2001: 38.

[3] KOLTER J Z, MALOOF M A. Learning to detect malicious executables in the wild[J]. Proceedings of the 10th ACM SIGKDD International Conference. New York: ACM, 2004: 470 – 478.

[4] WANG T Y, WU C H, HSIEH C C. A virus prevention model based on static analysis and data mining methods[C]// IEEE 8th International Conference on Computer and Information Technology Workshops. Washington, DC: IEEE Computer Society, 2008: 288 – 293.

[5] 王硕，周激流，彭博. 基于 API 序列分析和支持向量机的未知病毒检测[J]. 计算机应用，2007，27(8)：1942 – 1943.

[6] 余辉，赵晖. 支持向量机多类分类算法新研究[J]. 计算机工程与应用，2008，44(7)：185 – 189.

[7] ABE S, INOUE T. Fuzzy support vector machines for multiclass problems[EB/OL]. [2009 – 04 – 20]. http://www.dice.ucl.ac.be/Proceedings/esann/esannpdf/es2002-6.pdf

[8] CHANG C C, LIN C J. LIBSVM: A library for support vector machines[EB/OL]. [2009 – 06 – 20]. www.csie.ntu.edu.tw/ ~ cjlin/libsvm/.

（上接第 180 页）

[5] KRUGEL C, TOTH T, KIRDA E. Service specific anomaly detection for network intrusion detection[C]// Proceedings of the 2002 ACM Symposium on Applied Computing. New York: ACM Press, 2002: 201 – 208.

[6] LEE W, STOLFO S J. A framework for constructing features and models for intrusion detection systems[J]. ACM Transactions on Information and System Security, 2000, 3(4): 227 – 261.

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

[8] 陈治平，王雷. 基于密度梯度的聚类算法研究[J]. 计算机应用，2006，26(10)：2389 – 2392.

[9] DU QIANG, GUNZBERGER M, JU LILI, *et al*. Centroidal Voronoi tessellation algorithms for image compression, segmentation, and multichannel restoration[J]. Journal of Mathematical Imaging and Vision, 2006, 24(2): 177 – 194.

[10] 肖立中，邵志清，马汉华，等. 网络入侵检测中的自动决定聚类数算法[J]. 软件学报，2008，19(8)：2140 – 2148.

[11] PORTNOY L, ESKIN E, STOLFO S J. Intrusion detection with unlabeled data using clustering. [EB/OL]. [2009 – 04 – 15]. http://sneakers.cs.columbia.edu/ids/publications/cluster-ccsdmsa01.pdf.