改进数据挖掘算法在入侵检测系统中的应用

赵艳君[1],魏明军[2]
ZHAO Yanjun[1], WEI Mingjun[2]

1.河北联合大学 理学院,河北 唐山 063009
2.河北联合大学 信息学院,河北 唐山 063009
1.College of Science, Hebei United University, Tangshan, Hebei 063009, China
2.College of Information Engineering, Hebei United University, Tangshan, Hebei 063009, China

ZHAO Yanjun, WEI Mingjun. Application and realization of improved data mining algorithm in intrusion detection system. Computer Engineering and Applications, 2013, 49(18):69-72.

Abstract: Aiming to the existing problem of the powerless, high false negative rate, low detection efficiency and the lack of the rule base automatic extension mechanism to unknown aggressive behavior for existing detection mechanisms, combining the relevant knowledge of data mining technology, this paper designs one improved network intrusion detection system model based on data mining, combining misuse detection and anomaly detection. The model selects the *K*-means algorithm in clustering analysis and the Apriori algorithm in association rule mining and improves it. It applies the improved *K*-means algorithm to achieving normal behavior classes and data separation module, then utilizes the improved Apriori algorithm to achieve automatic extension of the rule base. By the experiment it verifies the function of the two algorithms.
Key words: data mining; intrusion detection; improved; *K*-means algorithm; Apriori algorithm

摘 要: 针对已有检测机制存在的对于未知攻击行为无能为力、漏报率较高、检测效率低以及缺少规则库自动扩充机制等问题,结合数据挖掘技术的相关知识,设计了基于数据挖掘的改进网络入侵检测系统模型。模型中选取聚类分析*K*-means算法和关联规则挖掘Apriori算法,并对其进行改进。用改进的*K*-means算法实现正常行为类及数据分离模块,用改进Apriori算法实现规则库的自动扩充功能,并通过实验验证了两个算法的功能。
关键词: 数据挖掘;入侵检测;改进;*K*-means算法;Apriori算法
文献标志码: A **中图分类号:** TP301 doi:10.3778/j.issn.1002-8331.1304-0309

1 引言

随着网络安全问题在人们生活中的重要性不断增强,作为新一代网络安全技术的入侵检测技术在网络安全中也发挥着越来越重要的角色。与此同时,现有入侵检测系统存在问题日益突出。现有入侵检测大多采用模式匹配的方式检测攻击,需要将待检测的数据包与规则库中的数据一一匹配,对已知攻击行为检测率较高,误报率较低,但对于已知攻击的变种或未知攻击却不能检测出来。而且,模式匹配需要事先建立一个已知攻击的检测规则和模式库,并需要安全领域专家不断进行规则库的更新和维护,否则就会造成系统的漏报率升高。因此,提高现有入侵检测系统的检测率、降低漏报率具有非常重要的现实意义。

依据检测所使用方法的不同,可以将传统入侵检测模型分为基于误用的入侵检测模型和基于异常的入侵检测模型两种[1-2]。基于误用的入侵检测模型需要建立一个已知攻击的规则库,并需要不断对知识库进行更新,才能跟踪攻击技术的发展,及时将新的攻击检测出来。因此,误用检测模型检测效果的好坏很大程度上依赖于模式库的及时更新。由此,可以看出,误用检测只能对已经发现的攻击进行防护,它不具备感知未知攻击的能力。而基于异常的入侵检测模型是根据统计数据,给正常行为建立一个模式库,一旦发现某种行为超出了正常行为的范围,就会将其当做入侵,做出相应反应[3-5]。此种检测模型的好处是对于未知攻击有着天生的良好感知能力。但是,由于系统的活动行为是在不断变化的,因此,需要对于正常行为模式库也需要不断调整,难于计算。对比这两种检测模型可发现,异常模型难于进行定量分析,不易实现;而误用模型会按照事先定义好的规则,将待检测数据与规则库中的数

基金项目: 河北省自然科学基金(No.F2012209019)。
作者简介: 赵艳君(1977—),女,讲师,主要研究领域为数据挖掘,网络技术应用;魏明军(1969—),男,副教授,主要研究领域为计算机网络安全。E-mail:zhaoyanjun@heuu.edu.cn
收稿日期: 2013-04-22 **修回日期:** 2013-06-14 **文章编号:** 1002-8331(2013)18-0069-04

据做模式匹配,实现起来相对简单。因此,目前大多数的入侵检测系统采用的是基于误用的入侵检测模型,而基于异常检测的入侵检测系统和二者结合的还比较少,多处于研究阶段。

为了改善现有入侵检测系统的性能,本文将误用检测与异常检测结合起来,构建一个基于误用检测和异常检测相结合的混合入侵检测模型。数据挖掘的最大特点在于它能够从繁杂的数据中发现人们未知的知识和规律,并且具有分析过程自动化、快速等优点。本模型中采用数据挖掘技术[6]实现以上提出的对入侵检测系统的改进。利用数据挖掘中的聚类分析算法对网络数据进行处理,建立正常行为类,以排除大部分正常数据。利用关联规则算法实现规则集的自动扩充机制。最后使用实验数据对两个改进算法进行了功能验证。

2 构建基于改进数据挖掘算法的网络入侵检测系统模型

2.1 系统组成

基于以上设计思路,本文构建了一个基于改进数据挖掘算法的网络入侵检测系统模型,如图1所示。

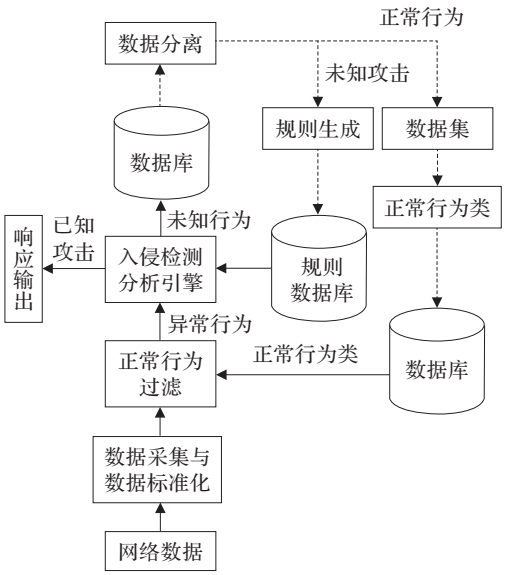

图1 基于数据挖掘的入侵检测系统模型

图1中,实线表示实时部分,虚线表示非实时的部分。

模块的基本功能与设计说明如下:

(1)数据采集与数据标准化模块:该模块主要负责采集网络数据并将数据进行标准化处理,为攻击检测做好准备。

(2)正常行为类模块:该模块的作用是将收集的数据作为训练数据,通过采用改进的聚类分析算法*K*-means,把训练数据归为几类,提取出形成的正常行为类的特征作为类的标志,形成正常行为类模式,并将其做为正常行为过滤模块的依据。

(3)正常行为过滤模块:该模块作为网络数据在进入检测分析引擎前的预处理程序。它利用正常行为类模块生成的正常行为类模式,对即将要进入检测分析引擎的数据包进行过滤,将符合正常行为类的数据过滤出来,这样可以大大降低分析引擎的工作量。

(4)入侵检测分析引擎模块:该模块将来自正常行为过滤模块的异常行为进行进一步分析,主要采取模式匹配的方法与规则数据库中的已知攻击规则进行匹配,若为攻击则响应输出;若不是,则是未知行为,存入数据库待下一步处理。

(5)数据分离模块:该模块处理的数据是经过分析引擎处理过的数据,这部分数据中会包含未知攻击数据和正常数据,需要将两部分数据进行分离,将未知攻击数据提供给规则生成模块,正常数据保存起来,用以更新正常行为类。采用的算法仍然是改进的*K*-means算法。

由于该模块的功能与正常行为类模块的实质相同,只是所处理的数据源不同,流程及过程不再重复,只需将数据换为日志记录即可,将正常数据保存到数据集中,攻击数据传送给规则生成模块。

(6)规则生成模块:该模块采用改进的Apriori算法,对来自数据分离模块的未知攻击数据进行关联规则挖掘,将发现的未知入侵行为模式表示成规则,并将其保存到规则库。

(7)规则数据库:规则数据库用于存放入侵检测分析引擎进行模式匹配所需要的规则。

(8)数据集:开始时使用的是初始数据集KDD CUP1999。

2.2 系统工作流程

以上提出的基于数据挖掘的入侵检测系统模型的运行过程可以分为以下几步:

(1)建立正常行为类:①利用基于误用的入侵检测系统,如snort来收集网络正常行为数据作为前期的训练数据。②利用数据挖掘中聚类分析算法,对收集的数据进行处理,将其聚类成几类,形成正常行为类,将其存储到数据库中。

(2)入侵检测:①利用网络嗅探器收集网络数据包。②将数据包解码,并将数据字段存入相应的数据结构当中。③数据进行标准化处理,为正常行为过滤做准备。④将数据与正常行为类进行比较,如果属于其中的某类,则表明是正常数据,将其丢掉;如果不是,则将其转交给检测引擎进行模式匹配,作进一步分析。⑤模式匹配,如果成功,则为攻击,那么相应模块会做出设定措施;若不是,则该数据中可能包含新攻击,将数据存储起来作为产生新规则的数据集。

(3)添加新规则:利用聚类分析算法对存储的数据进行聚类,排除小部分正常数据。然后,利用数据挖掘中的关联规则算法作关联分析,发现新规则并将其添加到规则库中。

3 入侵检测系统中算法的实现

基于改进数据挖掘算法的网络入侵检测系统中采用的算法是聚类分析算法(*K*-means)和关联规则算法(Apriori),

并对两种算法分别进行了改进。

3.1　改进 K-means 算法

K-means 算法[7-11]是硬聚类算法，是典型的基于聚类准则函数的聚类方法的代表，它把每个数据对象到各个类中心的某种距离之和作为进行优化的目标函数，同时采取函数求极值的方法获得进行迭代运算的调整规则。该算法的最终目的是根据输入的聚类个数 k，将数据对象划分为 k 个类。

K-means 算法思想简单、计算复杂度小，能够满足入侵检测对于实时性的要求，实现起来比较简单。但该算法本身也存在着一些急需解决的问题：

（1）无法自主确定聚类个数，需要事先输入确定的 k 值。在 K-means 算法开始聚类前，它必须要预先确定聚类的个数 k，同时随机选择相同个数的数据对象作为初始聚类中心。这样的话就会造成划分的类不是很准确，而且自主确定聚类个数也很困难，有时还需要结合相关领域的先验知识作为参考。同时，网络入侵检测的过程是实时的，所以可能没有办法事先知道聚类的个数 k，这样也就无法选择用做初始聚类中心的 k 个数据对象。

（2）通过 K-means 算法聚类出来的类不能确定哪个类是正常的，哪个类是异常的。但是在进行检测的过程中却需要通过正常行为类来排除正常数据包，减轻检测引擎的工作量。

针对 K-means 算法存在的缺点，作出一些改进。

在改进的 K-means 算法中，引入了聚类引导函数 $f(x_i)$，该函数可以帮助确定聚类向着点密度高的方向进行聚类。

定义 1（曼哈顿距离）

$$d(x_i,x_j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\cdots+|x_{is}-x_{js}| \tag{1}$$

式中，$x_i=(x_{i1},x_{i2},\cdots,x_{is})$，$y_j=(y_{j1},y_{j2},\cdots,y_{js})$ 都是 s 维的数据对象。

聚类引导函数中 r 的计算：

$$r=\frac{1}{m}\times\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}d(x_i,y_j) \tag{2}$$

其中，m 为正数，可调节，通常情况下为 2。

定义 2（聚类引导函数）

$$f(x_i)=\{p|Dist(p,x_i)\prec r,p\in X,x_i\in X\} \tag{3}$$

式中，X 是数据对象集合，$Dist(p,x_i)$ 为数据对象 p 到数据对象 x_i 的距离，r 代表距离半径，采用曼哈顿距离计算。

数据集中的每个对象可以根据以上定义的计算方法计算出对应的聚类引导函数值，该值将会被用做聚类中选择聚类对象的依据。

改进的 K-means 算法基于找到一个小范围内具有最大聚类引导函数值 $f(x_i)$ 的对象，即点密度最高的对象作为所在类的代表，每个类将会是由代表点所代表的对象和属于该代表点的对象组成。

改进 K-means 算法的整个算法描述如下：

输入：$X=\{x_1,x_2,\cdots,x_n\}$ 是包含 n 个对象的数据集合；

输出：k 个类。

过程：

根据输入的数据对象，计算出 r 及每个对象的聚类引导函数；

将每个对象看做一类，这样就形成了 n 个类，每个类中只有一个代表点；

Do

对于每个类代表点 x_i，寻找对象 x_j，$i\neq j$，x_j 到 x_i 的距离小于 r，且 $f(x_j)$ 是所有小于 r 对象中 $f(x_j)$ 最大的；

比较 $f(x_i)$ 和 $f(x_j)$，若 $f(x_i)\prec f(x_j)$，则类 x_j 归入类 x_i 中，否则，类 x_i 不变；

Until 对于每个 $i=1,2,\cdots,n$，类 x_i 不再发生变化。

3.2　改进 Apriori 算法

Apriori 是由 R.Agrawal 等人提出的数据挖掘中经典的关联规则算法，Apriori 采取多次扫描数据库的方法来产生频繁项目集[12]。具体做法是：第一次扫描后只产生宽度为 1 的候选项目集，然后通过比较每个项目的支持度与最小支持度，将不低于最小支持度的 1-阶候选项目集添加到频繁项目集中，此时形成了只包含 1-阶项目的最初的频繁项目集。此后，每一次扫描，都要根据前一次产生的频繁项目集，先构造出本次的候选项目集，然后再扫描数据库，从候选项目集中确定出本次的频繁项目集。重复以上过程，直到不再产生新的频繁项目集为止。

Apriori 算法是一个通用的数据挖掘算法，并不支持入侵检测的多属性问题。Apriori 算法中的各个事务中的项目之间不存在相互关系，一个项目的存在不会影响另一个项目，任何项目组合都可以。但是，对于入侵检测数据库中的数据并不是如此。如果直接使用该数据库的数据进行关联规则，由于中间过程中会产生无效候选项，可又需要扫描数据库计算其支持度，会大大降低该算法的效率。

根据入侵检测领域知识，可知同一特征下的不同属性支持度为 0 这一性质，可对 Apriori 算法作如下改进：

将 Apriori 算法 cand-gen（）子程序改为：

```
输入：频繁项目集 Lk-1
输出：初始候选项目集 Ck
Begin
for i=1 to Lk-1 项目集数
  for j=1 to Lk-1 项目集数
{
   if li 和 lj 只有一个项目不同 且不同的这两个项目不属于同一特征
      li ∪ lj ∈ Ck；
}
End
```

4　验证算法功能

4.1　K-means 算法性能验证实验

实验数据集：采用入侵检测领域比较权威的测试数据 KDD cup99[13] 中的“kddcup.data_10.percent”10% 数据集。

该实验数据集中主要包括 DoS、U2R、R2L 和 Probe 四大类攻击数据。以下分别针对这四类攻击数据对算法性能进行检测。

为了满足检测算法中两个假设的需要,每次实验从相应数据集中选择2 000条记录用于实验,其中正常数据1 966条,入侵数据34条,正常数据在所有数据中所占比例为98.30%,满足了检测算法对于正常数据的数量要远大于入侵数据数量的假设的需要。将各数据集分别在*K*-means和改进*K*-means上进行实验,比较算法性能。检测率越高,误检率越低,说明算法的检测能力越好。

以下将采用两种不同的角度对算法的性能进行实验,检测算法的性能。一种是采用单一攻击数据集,即数据集中的攻击数据均属于单一类别。另一种就是混合攻击检测,即数据集中的攻击数据是各种攻击数据的混合,类别更丰富。

(1)算法对单一攻击的检测性能:*K*-means算法需要事先指定聚类个数,且不同的聚类数对算法的聚类结果有很大的影响,所以需要进行反复实验。通过调整聚类个数得出*K*-means算法在各种攻击数据集上的聚类性能如表1所示。

表1 *K*-means算法单一攻击性能

数据集	聚类个数	误报率/(%)	检测率/(%)
DOS	15	0.864	100.000
R2L	11	2.339	91.176
Probe	20	6.256	100.000
U2R	15	4.629	8.823

改进*K*-means算法不需要事先指定聚类个数,可以根据数据集特定自主确定聚类个数。最终得出改进*K*-means算法对各种攻击的聚类性能如表2。

表2 改进*K*-means算法单一攻击性能

数据集	聚类个数	误报率/(%)	检测率/(%)
DOS	35	0.712	100.000
R2L	35	4.425	100.000
Probe	50	0.966	94.118
U2R	15	0.560	5.882

实验结果分析:从表1和表2中可以看出,和*K*-means算法相比,改进*K*-means算法的单一攻击性能中误报率更低、检测率更高。

(2)算法对综合攻击的检测性能:*K*-means算法聚类性能如表3所示。

表3 *K*-means算法的综合攻击性能

实验序号	聚类个数	误报率/(%)	检测率/(%)
1	36	1.424	20.588
2	40	0.102	67.647
3	44	2.747	88.235
4	50	5.239	91.177
5	55	6.205	91.177

改进*K*-means算法性能如表4。

表4 改进*K*-means算法的综合攻击性能

实验序号	聚类个数	误报率/(%)	检测率/(%)
1	44	1.984	91.177

实验结果分析:从表3和表4中可以看出,和*K*-means算法相比,改进*K*-means算法的综合攻击性能中误报率更低、检测率更高。

综上所述,改进*K*-means算法在不需要输入聚类个数的前提下,能够根据数据特点找出比较适当的聚类数量。它对单种入侵和混合入侵均表现出较低的误报率和较高的检测率,因此,改进*K*-means算法是可行的。

4.2 Apriori算法性能验证实验

利用改进的Apriori算法实现规则自动扩充能力。首先从数据集中挖掘出频繁项目集,进而形成关联规则并将其存入规则库中。

本次实验输出规则采用的Snort规则的格式,因此,选择duration,protocol_type,service,flag,src_bytes和dst_bytes这六个和Snort规则相关的字段。

实现规则自动扩充功能的程序界面如图2所示。

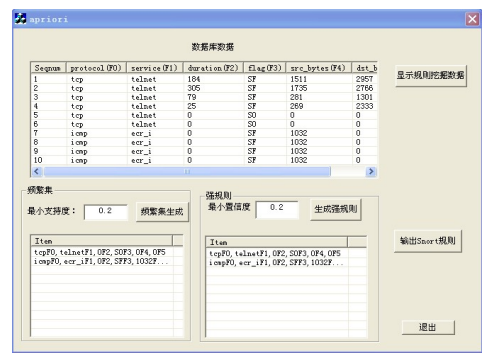

图2 规则生成

根据图2所示,生成规则的过程为:当设定最小支持度为0.2时,接着点击“频繁集生成”按钮,在下面的列表框里会显示出生成的满足最小支持度阶数最高的频繁集,如图2,生成两个支持度大于0.2的频繁集。然后,设置最小置信度为0.2时,点击“生成强规则”按钮,在下面列表框里将会显示生成的满足最小置信度的规则。如图2,刚刚产生的频繁集都是强规则。最后,点击“输出Snort规则”,程序就会将刚刚生成的强规则按Snort规则的格式写到文本文件rules.txt中。

结论,通过规则生成程序对未知攻击数据进行关联规则挖掘,可以实现入侵检测系统规则库的扩充。

5 结束语

本文设计了一个基于数据挖掘的、将误用检测和异常检测相结合的入侵检测系统模型,并对相关模块的工作流程及工作步骤进行了详细的介绍。针对模型中重点模块要实现的功能,在数据挖掘算法中选择了合适的算法。用改进的*K*-means算法实现正常行为类及数据分离模块,用改进Apriori算法实现规则库的自动扩充功能,并通过实验验证了两个算法的功能。

基于改进数据挖掘算法的入侵检测系统模型的研究实现了入侵检测系统的检测率的提高和漏报率的降低,改善了整个检测系统的检测性能。

(下转115页)

6 属性排序与关联规则提取

定义13 给定关联规则 $r:P \to Q$,记 $supp(r)=|g(P \cup Q)|/|G|$ 为该规则的支持度,$conf(r)=|g(P \cup Q)|/|g(P)|$ 为该规则的置信度。

定理5 若属性 b_i 和 b_j 是可比的,并且 $b_i < b_j$,那么一定存在置信度为1的关联规则 $b_i \to b_j$。

证明 因为 $b_i < b_j$,所以有 $g(b_i \cup b_j)=g(b_i)$,$conf(b_i \to b_j)=|g(b_i \cup b_j)|/|g(b_i)|=1$。

定理6 若属性 b_i 和 b_j 是不可比的,即 $b_i \| b_j$,那么关联规则 $b_i \to b_j$ 或者 $b_j \to b_i$ 的置信度一定存在小于1。

证明 因为 $b_i \| b_j$,所以有 $|g(b_i \cup b_j)| < |g(b_i)|$ 和 $|g(b_i \cup b_j)| < |g(b_j)|$,根据置信度定义,易知定理成立。

根据定理5,可以直接根据属性的排序序列直接得到置信度为1的关联规则。例如,在概念格 $L(K)$ 中,根据属性排序序列5<(((4,6)<2)||(9<(8<7||3))<1,可以得到 $9 \to 3$,$2 \to 1$,$8 \to 1$,$8 \to 7$ 等关联规则。

用定理5获得的关联规则一定是最简的。根据内涵缩减计算得到的关联规则实际上蕴涵了这些最简规则。例如,在概念格 $L(K)$ 中,(23,1278)的内涵缩减为{2,8},因此可以得到关联规则 $28 \to 17$,$28 \to 17$ 蕴涵了 $2 \to 1$,$8 \to 1$,$8 \to 7$。但是,根据序列5<(((4,6)<2)||(9<(8<7||3))<1 却无法得到 $28 \to 17$,而这个序列却不包含这样的信息,这样的信息只能在概念格中找到,也就是若要得到所有任意属性组合之间的关系,则需要建立概念格,而后计算内涵缩减得到关联规则。

7 结束语

本文深入研究了对象概念和属性概念,分析了对象概念和属性概念与不可约元的关系,提出了对象概念和属性概念的识别算法,进而得到了以属性概念为递归终止条件的内涵缩减计算方法,在最后研究了对象和属性的比较及其在规则提取中的应用。进一步需要研究的是对象概念和属性概念与奇异点的关系,以及对象概念和属性概念与概念稳定性的关系等内容。

参考文献:

[1] Ganter B,Wille R.Formal concept analysis:mathematical foundation[M].New York:Springer-Verlag,l999.

[2] Scaife M,Rogers Y.External cognition:how do graphical representations work[J].International Journal of Human Computer Studies,1996,45:185-213.

[3] 姜峰,范玉顺.基于扩展概念格的Web关系挖掘[J].软件学报,2010,21(10):2432-2444.

[4] 丁卫平,顾春华.基于形式概念分析的不完备电子病历系统粗糙挖掘研究[J].计算机科学,2009,36(10):230-233.

[5] Passquier N,Taouil R,Bastide Y,et al.Generating a condensed representation for association rules[J].Journal of Intelligent Information Systems,2005,24:29-60.

[6] Roth C,Obiedkov S,Kourie D G.Towards concise representation for taxonomies of epistemic communities[C]//Yahia S B,Nguifo E M.Proc CLA 4th International Conference on Concept Lattices and their Applications,2006,4923:240-255.

[7] 智东杰,智慧来,刘宗田.概念格的内涵缩减研究[J].计算机工程与应用,2009,45(1):42-44.

[8] 谢志鹏,刘宗田.概念格节点的内涵缩减及其计算[J].计算机工程,2001,27(3):9-11.

[9] 智慧来,智东杰,刘宗田.从合取范式到析取范式的转换研究[J].计算机工程与应用,2012,48(2):15-17.

[10] Skowron A,Rauszer C.The discernbility matrices and functions in information systems[M]//Intelligent Decision Support,Handbook of Applications and Advances of the Rough Sets Theory.The Netherlands:Kluwer,1992:331-362.

(上接72页)

参考文献:

[1] 王艳,肖维民.数据挖掘技术在入侵检测系统中的应用研究[D].马鞍山:安徽工业大学,2010.

[2] Verwoerd T,Hunt R.Intrusion detection techniques and approaches[J].Computer Communications,2002,25(15):1356-1365.

[3] 于琨.基于高频统计的异常检测方法的设计与实现[D].北京:北京邮电大学,2006.

[4] 蔡坚.基于人工神经网络的入侵检测系统的研究与实现[D].贵阳:贵州大学,2005.

[5] 武涛,王新房.基于数据挖掘的入侵检测研究[D].西安:西安理工大学,2010.

[6] 陈宇晖,傅明.基于数据挖掘的入侵检测方法研究[D].长沙:长沙理工大学,2010.

[7] 李洋.*K*-means聚类算法在入侵检测中的应用[J].计算机工程,2007,33(14):154-156.

[8] 张建萍.基于聚类分析的*K*-means算法研究及应用[J].计算机应用研究,2007,24(5):166-168.

[9] Chinrungrueng C,Sequin C H.Optimal adaptive *k*-means algorithm with dynamic adjustment of learning rate[J].IEEE Trans on Neural Networks,1995,6.

[10] 严晓光.聚类在网络入侵的异常检测中的应用[J].计算机系统应用,2005(10):34-37.

[11] 马晓春,高翔,高德远.聚类分析在入侵检测系统中应用研究[J].微电子学与计算机,2005,22(4):134-136.

[12] Sun Ying.Using data mining technology solve intrusion detection of network[J].Computer Knowledge and Technology,2010,6(23):6463-6464.

[13] 张新有,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.