基于 k－means 改进算法的入侵检测系统的研究

Research of IDS Based on Data Mining

关 心 王 新 （湛江师范学院 广东省湛江市 524048）

摘要：本文介绍了入侵检测系统的基本概念，分析了数据挖掘技术在入侵检测系统中的应用。本文主要研究了聚类分析中的 k－means 算法在入侵规则匹配中的应用，指出了该算法的不足，通过对传统 k－means 算法的改进解决了聚类算法固有的无法预知最佳聚类个数和分类过细的问题。提高了系统的规则匹配效率。

关键词：入侵检测 数据挖掘 聚类分析

1 引言

随着 Internet 的迅速发展，信息安全日益受到人们的关注，作为网络安全一个组成部分的入侵检测技术被重视起来，并逐渐成为保障网络信息安全不可缺少的部分。基于内容的网络安全解决方案成为人们研究的重点之一。

传统的信息安全方法采用严格的访问控制和数据加密策略来保护，在复杂系统中，这些策略是必要的，但不是充分的。入侵检测系统是检测企图破坏系统资源的可用性、真实性和完整性行为的软硬件。它能实时监测系统的活动，实时的发现攻击行为（甚至即将发生的攻击行为），并采取相应的措施来避免攻击的发生或尽量减少攻击产生的危害。

入侵检测系统作为一种积极主动的安全防护系统，它提供了对内部攻击、外部攻击和误操作的实时防范，在网络系统受到危害之前识别和拦截入侵行为。入侵检测原理如图 1 所示。

从图 1 可以看出：通过对历史行为、特定行为模式等的分析，可以制订出安全（入侵）规则库，进而指导当前系统的入侵检测，实现对入侵行为的识别和预防。

2 数据挖掘技术

为了应对入侵检测系统信息量大的问题，人们引入了数据挖掘技术。将数据挖掘应用于入侵检测能够广泛地审计数据来得到模型，从而精确地捕获实际的入侵和正常的行为模式。数据挖掘（DM，Data Mining），又称为数据库中的知识发现（KDD，Knowlage Discovery in Database），是指从大型数据库或数据仓库中提取隐含的、未知的、异常的及有潜在应用价值的信息或模式。它是数据库研究中的一个很有价值的新领域，融合了人工智能、机器学习、统计学等多个领域的理论和技术。

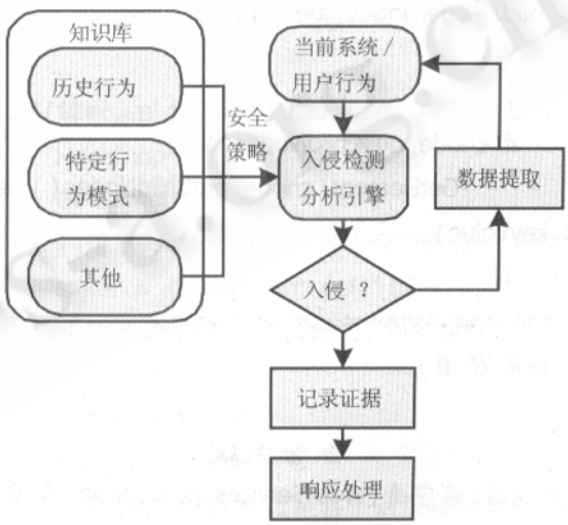

图 1 入侵检测原理

数据挖掘发现的知识形式通常有：概念（Concepts）、规则（Rules）、规律（Regularities）、模式（Patterns）、约束（Constraints）、可视化（Visualizations）等。这些知识可以直接提供给决策者，用以辅助决策过程；或者提供给领域专家，修正专家已有的知识体系；也可以作为新的知识转存到应用系统的知识存储机构中，

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

如专家系统(Expert System)、规则库(Rule Base)等。

数据挖掘必须具备以下几个方面的要素,才能获得满意的结果:

2.1 规模

要从数据中挖掘出规律,则数据源的规模必须大。如何从如此海量的数据中有效地提取出有用的信息,需要各方面技术的协调;

2.2 历史数据

数据挖掘必须对数据进行长期趋势的分析,但数据在时间轴上大的纵深性是数据挖掘的一个新难点;

2.3 数据集成和综合性

数据挖掘可能要面对的是关系非常复杂的全局模式的知识发现,但注意力可以更集中于数据采掘的核心处理阶段;

2.4 查询支持

一般由用户提出的即时随机查询,往往不能形成精确的查询要求,需要靠数据挖掘技术进行实时交互,使决策者的思维保持连续,才有可能挖掘出更深入、更有价值的知识;

2.5 模式的适用性

数据挖掘模式的发现主要基于大样本的统计规律,发现的模式不必适用于所有的数据,达到某阈值就可以为有效。

3 数据挖掘技术在入侵检测系统中的应用

典型的数据挖掘应用方案如下:

(1) 数据准备。在这个阶段,将从操作环境中提取并集成数据,解决语义二义的问题。

(2) 消除脏数据等等,然后对数据进行选择和预分析。在 IDS 中,将用户的历史行为数据和当前操作数据进行集成,并删除一些无用的数据和预处理,以备用于数据挖掘。

(3) 挖掘。在这个阶段里,综合利用关联分析、序列分析、分类分析、聚类分析等多种数据挖掘方法分析经过预处理的数据,从中提取有关特征和规则。

(4) 表达。数据挖掘将获取的特征和规则以便于理解和观察的方式反映给系统。在入侵检测系统中,通过数据挖掘发现有关的特征和规则后,再根据这些特征和规则将用户的异常模式和正常模式定义出来,然后存储在知识库中。另外系统还对当前的用户的行为数据进行挖掘后找出特征和规则,然后以一定的方式表达出来,系统将它与知识库中的模式进行匹配检测。

(5) 评价。可以对数据挖掘后所提取的网络安全异常模式或正常模式进行评价,如果能够有效地检测出入侵行为,就说明它是成功的,否则的话,就可以重复执行上述过程,直至得出满意的结果为止。

4 聚类分析与 k – means 算法

聚类就是将数据对象分组成为多个类或簇,划分的原则是在同一个簇中的对象之间具有较高的相似度,而不同簇中的对象差别较大。利用这个特性能够方便地把海量的网络信息分成若干的性质不同的簇,通过进一步的模式匹配来发现其中的攻击行为。本文使用了聚类规则中经典的 k – means 算法。

4.1 传统 k – means 算法的缺点

传统的 k – means 算法存在一些缺点:

(1) 必须预先确定最终聚类的个数 k,并指定同样个数的记录作为初始聚类中心,然后反复扫描整个记录集,不断地改变聚类中心和改变记录所归属的聚类,直到聚类中心不再改变为止。而聚类结果与聚类个数 k 的大小有直接关系,不同的聚类个数会产生不同的聚类结果,在实际应用中往往很难确定产生最佳聚类效果的聚类个数。

(2) 聚类结果中某些聚类可能为空。也就是没有对象和该聚类的聚类中心相似,即产生了空聚类。

(3) 在聚类初始时随机选择的初始聚类中心可能并不是最佳的,聚类结果受初始聚类中心的影响。我们知道传统的 k – means 算法在选取初始聚类中的时候采取任意选择的方式,而事实上,选择较好的初始聚类中心可以得到更好的聚类效果。

4.2 k – means 算法的改进方案

针对以上分析的 k – means 算法的不足,进行了一些改进,即在算法中增加两个聚类参数:聚类半径 r 和最邻近阈值 h。具体方法是:计算当前记录与所有聚类中心距离的最小值,若发现此最小值大于聚类半径 r,则将这个记录作为新的聚类中心。最终聚类个数 k 为数据集的最佳聚类个数,而不是预先确定的聚类个数。这种方法不需要预先确定聚类个数,能够把距离接近的记录归入到同一聚类中,而把大多数距离较

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

远的异常记录孤立到远离其它聚类的聚类中，能在一定范围内调节聚类的个数，并且能使异常记录归入单独的聚类，有利于对异类记录进行标记，减少了他们对计算聚类中心的影响。

当聚类结果中出现空聚类时，采用的方法是将离聚类中心最远的对象移出所在的聚类，来产生新的聚类中心，用新产生的聚类来替代这些空聚类。

在初始聚类中心的选择上采用 P. S. Bradley 和 Usama M. Fayyad 提出的一种初始聚类中心方法：首先从原始数据集中提取 t 个样本集，对每个样本集进行 k－means 聚类，生成 t 个初始中心集（每个集中包括 k 个元素）；然后分别以 C_i 为初始聚类中心，对所有的 t×k 个元素用 k－means 算法聚类，得到 t 个聚类中心集，选取效果最好的作为最终初始聚类中心。

这种方法的优点是提出了一种自动选择初始聚类中心的方法；通过选取样本而不是整个数据集上实现，降低了算法的时间复杂度；利用多个样本集，通过对预初始中心聚类，可以避免“孤立点”的影响，提高初始中心的代表性。

改进的 k－means 算法描述如下：

输入：包含 n 个数据的数据库

输入参数：初始聚类个数 M；聚类半径 r；最邻近阈值 h

输出：k 个聚类

（1）选取 M 个最佳初始聚类中心 $\{w_1, w_2, \cdots, w_m\}$，其中 $w_i = x_i, i \in \{1,2,\cdots,K\}, i \in \{1,2,\cdots,n\}$；

（2）使每个聚类 C_i 与 w_i 相对应；

（3）计算其它记录 $x_i (i \in \{1,2,\cdots,n\})$ 到聚类中心的距离的最小值 min；

（4）若 min＜r，将 x_i 分到最近的 $w_{i\cdot}$ 所在的聚类 $C_{i\cdot}$，即 $|x_i - w_{i\cdot}| \leq |x_i - w_i|, i \in \{1,2,\cdots,k\}$；否则，产生一个新聚类，将 x_i 作为新聚类的中心；

（5）返回 2，直到所有记录都完成；

（6）以每个聚类的平均值替代原来的聚类中心，即 $w_i = \sum_{x_i \in C_i} x_i / C_i$

（7）如果出现空聚类，则离聚类中心最远的点移出所在的聚类，来产生新的聚类中心，用新产生的聚类来替代这些空聚类；

（8）直到聚类中心值不变为止。

（9）最后计算所有聚类中心值中任意两个中心之间的距离，并与最邻近阈值 h 进行比较，如果这个距离值小于 h 则合并这两个聚类。

（10）重复 9，直到任意两个聚类中心的距离值都大于 h 为止。

4.3 实验结果分析

在本文的实验中采用了 KDD Cup 99 数据集，其中包括 4 大类型的攻击中 DOS 攻击（包括 smurf、neptune、back、teardrop、pod、land 等）和 PROBE 端口扫描攻击（包括 portsweep、ipsweep、satan 等）占了 80% 以上，分别对这两种攻击进行检测。选取数据中的 15 个关键数值属性进行聚类，在聚类过程中并不用到记录的类别标识，聚类集结果可以直接将数据聚成不同的类别，使得异常类和正常类数据分开。

通过实验可以发现，取不同的聚类数对结果的影响很大。而我们无法预知最佳的聚类个数。采用改进的 k－means 算法可以解决初始聚类个数选择难的问题。算法引入了聚类半径和最临近阈值，不需要预先输入聚类个数就能将数据聚成最佳的个数。

引入参数后的 DOS 攻击检测结果

M	η(%)	h	r	DOS 攻击		
				最终聚类个数	检测率(%)	误检率(%)
2	2	7	8	37	99.27	7.18
2	2	7	10	31	94.58	5.67
2	2	7	15	20	85.36	3.92
2	2	7	20	16	82.76	1.35

引入参数后的 PROBE 攻击检测结果

M	η(%)	h	r	PROBE 攻击		
				最终聚类个数	检测率(%)	误检率(%)
2	2	7	8	43	99.45	8.17
2	2	7	10	35	94.62	7.24
2	2	7	15	30	92.39	1.66
2	2	7	20	25	93.74	1.35

实验表明：改进的 k－means 算法对 DOS 攻击的检测率明显提高了，算法聚类效果更好，而且不需要提前指定最后的聚类个数，而且可以通过改变最临近阈值来控制聚类粒度。（下转第 109 页）

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

(上接第 84 页)

5 结束语

本文通过对入侵检测和数据挖掘技术的研究，针对入侵检测技术的领域特点，对 k－means 算法进行了改进和优化。并从理论和实验分别论证了算法改进的意义。

参考文献

1 毛国君、段立娟、王实，数据挖掘原理与算法 ，清华大学出版社.

2 朱树人、李伟琴，入侵检测技术研究[J]，计算机工程与设计，2001 年 04 期，13－18.

3 周皓峰、朱扬勇、施伯乐，一个基于兴趣度的关联规则算法，计算机研究与发展，2002.

4 Jiawei Han Micheline Kamber 著，数据挖掘概念与技术[M]，范明、孟小峰等译，北京：机械工业出版社，2001 年，35－210.

5 David Hand Heikki Mannila Padhraic Smyth 著. 数据挖掘原理[M]，张银奎、廖丽、宋骏等译，北京：机械工业出版社，中信出版社，2003 年，32 — 112.

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn