基于粗糙集关联规则挖掘的入侵检测研究

马洪江

(阿坝师范高等专科学校计算机科学系　四川 623000)

摘　要　本文用作为数据挖掘的工具，对入侵检测中采集的网络数据集进行分析，知识约简，从中挖掘出一些对用户有用的潜在规则，不再依赖专家的经验，使入侵检测具有更好的灵活性，智能性。

关键词　粗糙集，数据挖掘，关联规则，入侵检测

The Research of Association Rules in Intrusion Detection Based on RST

MA Hong-Jiang

(Computer Science Department of Aba Teachers College, Sichuan 623000)

Abstract　This paper uses RST as tool of data mining to analyze network data set of intrusion detection so that mine some potential rules useful to user from these data, the proposed method no longer rely on experience of expert and lead to the intrusion detection have better flexibility, intelligence.

Keywords　Rough set, Data mining, Association rules, Intrusion detection

1　引言

计算机网络给人们带来了巨大的帮助，随着网络的普及和新技术的不断出现，网络安全问题成为了人们不得不去面对的问题。目前常用的防火墙技术并不能完全解决这个问题，因为防火墙可以限制外部网络到内部网络的访问，但对于内部网络的入侵，就无能为力了。于是人们提出了入侵检测技术作为其它经典手段的加强和补充，是任何一个安全系统中的最后一道防线。

入侵检测是指对潜在的有预谋的未经授权的访问信息，操作信息以及致使系统不可靠，不稳定或无法使用的企图的检测和监视。根据检测方法，入侵检测可分为异常检测和滥用检测，异常检测是根据非正常行为(系统或用户)和使用计算机资源非正常情况检测出入侵行为，该方法能检测未知的攻击类型，但误检率比较高。滥用检测根据已经知道的入侵方式来检测入侵，该方法的检测准确率高，但只能检测到已知的攻击类型。

由于上述入侵检测方法在误检率，实时性，智能性等方面存在一系列问题，数据挖掘技术在入侵检测中得到广泛应用，在检测中的数据属于海量数据，要解决的问题是如何从这些海量数据中挖掘出对用户有用的信息，使用户及时知道网络数据中出现的异常情况，而数据挖掘技术就能解决这些问题。数掘挖掘的方法很多，比如关联规则，分类，聚类等等，本文采用关联规则的方法，用粗糙集理论进行知识约简，从知识表中挖掘出决策规则。使其检测性能得到进一步的提高。

2　粗糙集的基本理论

粗糙集理论是一种研究不精确，不确定性知识的工具。由波兰科学家 Z. Pawlak 在 1982 年首先提出。粗糙集理论是离散数据推理的一种新方法。粗糙集理论作为一种处理不完备信息的有力工具，可以不需要任何辅助信息(如统计学中的概率分布，模糊集理论中的隶属度等)，仅依据数据本身提供的信息就能够在保留关键信息的前提下，对数据进行化简并求得知识的最小表达，从而建立决策规则，发现给定数据集中的隐含知识。

2.1　不可区分关系

设 R 是 U 上的一个等价关系，U/R 表示 R 的所有等价类(或者是 U 上的分类)构成的集合，$[x]_R$ 表示包含元素 $x\in U$ 的 R 等价类。一个知识库就是一个关系系统 $K=(U,R)$，其中 U 为非空有限集，称为论域，R 是 U 上的一簇等价关系。

若 $P\subseteq R$，且 $P\neq\Phi$，则 $\cap P$(P 中所有等价类的关系)也是一个等价关系，称为 P 上的不可区分关系，记为 ind(P)，且有

$$[x]_{\mathrm{ind}(P)}=\bigcap_{R\in P}[x]_R$$

这样，$U/\mathrm{ind}(P)$(即等价关系 ind(P)的所有等价类)表示与等价关系族 p 相关的知识，成为 K 中关于 U 的 P 基本知识。

2.2　知识表达系统与知识库

在粗糙集中，一个知识表达系统 S 被定义为：

$S(U,A,V,f)$

其中，U：论域；A：属性的全体；V：$V=\bigcup_{a\in A}V_a$，V_a 是属性的值域；f：$U\times A\to V$ 是一个信息函数，它为每个对象的每个属性赋予一个信息值，即 $\forall a\in A, x\in U, f(x,a)\in V$。

知识表达系统也称为信息系统，通常也用 $S=(U,A)$ 来代替 $S=(U,A,V,f)$。知识表达系统的数据以关系表的形式表示，关系表的行对应要研究的对象，列对应对象的属性，对象的信息通过指定对象的各属性值来表达。

在知识表达系统 K 中，$A=C\cup D$，$C\cap D=\phi$，C 称为条件属性集，D 称为决策属性集。具有条件属性集和决策属性集的知识表达系统称为决策表。

3　基于粗糙集的数据挖掘

若两个或多个变量的取值之间存在某种规律性，就成为

马洪江　副教授，硕士，主要从事计算机网络及计算机应用研究。

关联。关联规则是寻找在同一个条件中的不同项的相关性，比如在一次购买活动中所购买不同商品的相关性。

关联规则挖掘的方法很多，其中最著名的是 Apriori 算法，其基本思想可分为两步：第一步是找出所有的频繁项目集：要求所有频繁项目集的支持度不低于设定的最小支持度；第二步是从所有的频繁项目集中挖掘出强关联规则：要求这些规则必须同时满足不低于最小支持度和最小置信度。其中，求所有频繁项目集的计算量最大。

由于入侵检测系统中的数据属于海量数据，用传统的求频繁项目集的方法效率很低，致使实时性不强，因此，本文用粗糙集理论对数据中的属性进行知识约简，删去冗余的属性，从中找出决策规则，再找出强关联规则。代替了传统的 Apriori 方法，使其效率、准确性都能得到提高。下面将会比较详细地介绍粗糙集的关联规则挖掘的思路以及其中所用到的一些粗糙集的方法。

3.1 构造决策表

把收集到的网络数据作为信息系统，比如一个连接记录包含以下属性：

(time, duration, service, src_host, dest_host, scr_bytes, dst_bytes, flag)

其中，time：表示连接开始的时间；duration：表示连接从开始到结束所经历的时间；service：即连接的应用协议如 WWW, FTP, DNS, Telnet 等；src-host：源主机；dst-host：目的主机；flag：连接状态标记，包括正常结束状态和连接请求被拒绝的状态。

本文把源主机端口，连接协议，连接开始的时间，经历的时间，目的端口，连接状态标记作为条件属性，连接的应用协议作为决策属性。每行表示一个连接记录，每列表示一个属性，这样就构成了一个关于网络数据流的决策表。

3.2 确定各属性的重要性

决策表中并不是每个属性都是同等重要的，有些对决策结果起主要作用，而有些是不重要的，判断属性重要性的基本思想是根据该属性对分类结果的影响，若去掉该属性对分类影响大，说明该属性是重要的，反之，是不重要的。设决策表中的条件属性和决策属性集分别为 C，D，对于任意属性的重要性[3]定义为：$sig^{Y}_{X-\{x\}}(x)=(|S_X(Y)|-|S_{X-\{x\}}(Y)|)/|U|$，其中 $S_X(Y)=Y^{(U/X)^-}=Y_{V\in U/X, V\subseteq Y}V$。由此，可以计算网络数据中的每一个属性的重要性，以便进行知识约简。

3.3 知识约简[3]找出决策规则

知识约简是粗糙集理论的核心内容之一。网络数据流决策表中并不是每个属性都是同等重要的，甚至其中有些是冗余的，这些属性会产生一些无用的规则。因此必须在保持知识库的分类能力不变的条件下，删除其中不相关或不重要的知识。在网络数据决策表中，如果重要性小于给定的最小重要性域值，说明该属性为不重要属性，从决策中删除。并且属性值完全相同的行也应合并为一行。经过约简，便得到约简后的网络数据流决策表，然后根据简化后的决策表，得出一系列的决策规则。

3.4 关联规则的产生

对得到的每个决策规则再计算其支持度和置信度，如果该规则不小于指定的最小支持度阈值和最小置信度阈值，就是强关联规则，保留下来，否则就丢弃。假设通过求属性重要性，删去不重要的属性 time, duration, flag 后，得到如下一条关联规则：src_host=202.96.7.5,dst_host=202.108.35.210→service=WWW,10,80，这条规则表示在所有的网络连接中有 10%符合源主机端口的 IP 地址为 202.96.7.5，目的主机端口 IP 地址为 202.108.35.210 且所提供的服务是 WWW。在源主机端口的 IP 地址为 202.96.7.5，目的主机端口 IP 地址为 202.108.35.210 的网络连接中，有 80%提供的是 WWW 服务。

结束语 粗糙集理论作为一种新型的数据挖掘工具，已经很好的体现出了它的优势，本文提出了基于粗糙集理论的关联规则挖掘算法，能比较迅速地挖掘出潜在规律，提高了入侵检测的性能。

参 考 文 献

1 周明全，吕林涛，李军怀. 网络信息安全技术. 西安电子科技大学出版社，2003

2 张云涛，龚玲. 数据挖掘原理与技术. 电子工业出版社，2004

3 张文修，吴伟志. 粗糙集理论与方法. 科学出版社，2003

4 周庆敏，李永生，等. 基于粗集理论的数据挖掘应用. 南京工业大学学报，2003(2)：44～47

5 宁玉杰，郭晓淳. 基于数据挖掘技术的网络入侵检测系统. 计算机测量与控制，2002，10(3)：189～191

(上接第 11 页)

附录 1 通过 Email 附件传播的 VBS 脚本病毒主要代码(仅用于实验没有危害)

```
Function mail_virus_test()
  On error resume next
  wscript. echo
  Set outlookApp = CreateObject("Outlook. Application") //创建
      //一个 outlook 应用对象
  If outlookApp = "Outlook"Then
     Set mapiObj = outlookApp. GetnameSpace("MAPI") //获取
        //MAPI 的名字空间
     Set addiList = mapiObj. AddressLists //获取地址表的个数
     For Each addr In addrList
        If addr. AddressEntries. Coun()0 Then
          AddrEntCount = addr. AddressEntries. Count //获取每
             //个地址表的
                                  // Email 记录
        For addrEntIndex = 1 To addrEntCount
        Set item = outlookApp. CreatItem(0) //遍历地址表的
                //Email 地址
        Set addrEnt = addr. AddressEnties ( addrEntIndex ) //获取
                //具体 Email 地址
        item. To = addr. Address //填入收件人地址
        item. Subject = "邮件病毒传播实验" //写入邮件标题
        item. Body = "收到此信不要担心，这仅是实验，对电脑没有
                                  危害"
                                  //真正病毒的破坏部分，
                                  这里仅是感染标记
        Set attachments = item. Attachment //定义附件
        Attachments. Add fileSysObj. GetSpecialFolder ( 0 ) &.
        "test. jpg. vbs"
        Item. DeleteAfterSubmit = Ture //信件提交后自动删除
        If item. To ( )"" Then
           Item. Send //发送邮件
           ShellObj. regwrite "HKCU\softerware\Mailtest\mailed"
           , " 1 "
                //病毒感染标记，以免重复感染
        Endif
        Next
        End if
        Next
        End if
End Function
```