联合挖掘发现网络安全事件

贺　蓉　赵振西　周学海　（中国科学技术大学软件学院　安徽合肥）
陈尚义　赵　巍　（中国软件通用产品研发中心信息安全实验室　北京）

摘要：针对关联规则 Apriori 算法在实际应用中存在一些缺陷，本文在分析聚类分析和关联规则这两种挖掘算法的基础上，结合网络安全事件的特点，讨论了将这两种独立的挖掘方法集成起来的联合挖掘应用于安全事件发现中。

关键词：数据挖掘　安全事件　关联规则　概念聚类　联合挖掘

1　引言

随着网络技术的突飞猛进，电脑和互联网的广泛应用，造成了各种网络数据数量巨大，其中包含了许多噪音数据，因此，如何从海量的数据中发现事件的相关性，查找事件的根源，成为安全管理的一个瓶颈。

数据挖掘是从大量的数据中抽取出潜在的、有价值的知识（模型或规则）的过程，是一门新兴的交叉性学科，其应用领域十分广泛多样。在网络安全管理方面，目前国内外已开始研究数据挖掘技术在安全管理系统中的应用，例如哥伦比亚大学的 Wenke Lee[1] 等人从 1995 年开始首先将数据挖掘技术应用于入侵检测，提出了各种安全事件检测的方法。近年的研究表明，数据挖掘技术在安全管理中具有广阔的应用前景，一些高校和网络安全公司都开始了这方面的研究工作。但由于很多都是基于单一的数据挖掘算法，在实际应用中存在许多问题。

2　关联规则和聚类算法

2.1　关联规则

关联规则比较经典的算法有 Apriori 算法、AprioriTid 算法。

Apriori[2] 算法利用了一个层次顺序搜索的循环方法来完成频繁项集的挖据工作，使用 Apriori 算法进行关联规则挖掘，可以比较有效地产生关联规则，但存在着以下两种缺陷：首先，算法产生太多的冗余规则；其次，算法在效率上存在着问题。因此当数据库或 K 太大时，算法的时耗太大或无法完成，故算法的可扩展性也不强，难以推广。

2.2　聚类算法

聚类算法[3] 是一个将数据集划分成若干个聚类的过程，使得同一聚类内的数据具有较高的相似性，而不同聚类中的数据不具有相似性。本文采用的聚类算法是概念聚类算法，它是一种机器学习方法，不仅能产生基于某种度量的分类，而且能为每种类别找出有意义的描述。概念聚类有 2 个重要优点：(1) 聚类的分层结构由领域知识得到。(2) 概念聚类特别擅长于处理像 IP 地址、端口地址这样的分类属性。

结合关联规则挖掘算法的缺陷和聚类分析挖掘算法的特点，本文提出了聚类分析和关联规则两种算法相结合的联合挖掘。

3　联合挖掘在安全事件管理中的应用

联合挖掘的基本思想是将聚类分析作为关联规则算法的一个预处理步骤，即先对数据库中的数据按照一定的方法进行聚类，将数据按照用户感兴趣的方向进行数据区域细化，将数据集放在相应的类型中，用户根据其关心区域的选定数据类进行关联分析，使得在关联规则分析的过程中数据集的范围大大缩小，从而提高挖掘的效率。

已知安全事件数据库 T（$\{A_1, A_2, \cdots\cdots A_m\}$）

(1) 首先，由领域知识得到地址的概化分层图。如图 1。

(2) 采用概念聚类算法进行数据预处理。

其算法具体如下：

For(T 中每个事件 e)
e. Count = 1; /* 初始化,设每个警报的计数值 Count 为 1 */

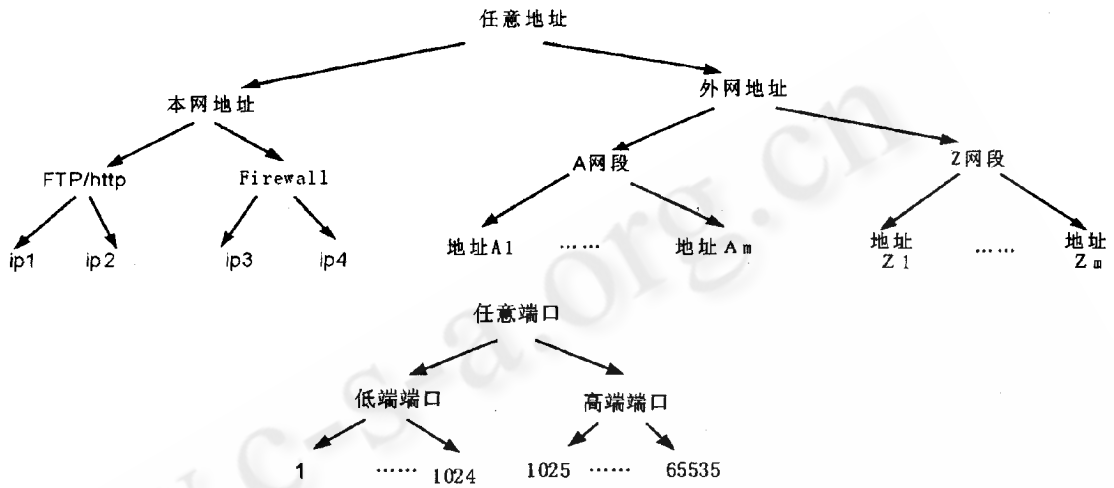

图 1

For(每个事件属性 Ai)
删除不能概化的属性;
While(T 中事件适合概化)Do
使用启发式算法选择合适的属性 Ai;
For(T 中的每个事件 e)
e. Ai = the father(e. Ai); /* 对相应的属性执行概化操作 */
For (T 中任意事件 ei,ej)
If (ei = ej)
ei. C = ei. C + ej. C
delete ej; /* 合并相同属性 */
If (存在合适的概化事件)
提交合适的概化事件给用户;
从 T 中删除提交的概化事件;
for (T 中每个事件 e)
e = T. e ; /* 将剩余警报恢复为未概化状态 */

(3) 用户根据自己的需求,对特定类采用 Apriori 算法进行关联规则分析。

具体算法如下:

扫描事件库 T 寻找含有用户指定属性的纪录来生成 1 - 项目集 L1 = {frequent 1 - itemsets};
k = 2; /* k 为项目集的长度 */
while LK - 10 do begin /* 进行关联 */
/* 取 LK - 1 的每对项目集进行关联,形成候选 k - 项目集 */
for ∀ L1K - 1 = {i1, i2, ⋯, ik - 1}, L2K - 1 = {j1, j2, ⋯, jk - 1} ∈ LK - 1 and L1K - 1 ≠ L2K - 1
where
their first k - 2 items are the same do
begin
ck = { i1, i2, ⋯, ik - 1, jk - 2 }; /* 生成候选 K - 项目集 */
if ((∃s, |s| = k - 1 and s ⊂ ck and s ∉ LK - 1)或 ck 中不包含轴属性) then
delete ck ; /* 删除不符合条件的候选 k - 项目集 */
else
Ck = Ck ∪ { ck }; /* 符合条件的并入候选 k - 项目集的集合中 */
end
for each ck ∈ Ck count the support (ck)
LK = {ck | support (ck) ≥ minimum_support};
k = k + 1;
end

```
for all lk , k >2 do / * 以下为规则生成过程 * /
 begin
  for all subset am ∈ lk do
   begin
    conf = support( lk) / support( am) ; / * 计算规则的可信度 * /
    if conf ≥ minimum_confidence then / * 规则的可信度大于用户给定的阈值 * /
    R = R∪ { am →( lk - am) , [ confidence = conf, support = support ( lk) ] } ;
   end
 end
```

4 实验

本文中实验数据来自实际的基于商业用途的 IDS 的历史警报的日志。扫描之间历史一周。包含了 156，380 个日志数据。以 SYN flood 对 windows NT 系统攻击的纪录为例：

表 1　示例数据

Event - ID	Source - IP	Source - Port	Dest - IP	Dest - Port	Event - Type	计数值 Count
899	ip1	21	ip3	3500	TCP SYN host sweep	300
1199	ip1	21	地址 A1	8001	http	1
1200	ip1	21	地址 B1	8002	http	1
1201	ip1	21	地址 B2	8002	http	1
1202	ip1	21	地址 C1	8003	http	1
1203	ip1	21	地址 D1	8004	http	1
.....	.....	.....	.....	.....	.....	.....
3128	ip2	21	ip3	21	FTP fragment attempt	30
3100	地址 A1	8001	ip3	21	FTP fragment attempt	11
3112	地址 B1	8002	ip3	21	FTP fragment attempt	12
3124	地址 C1	8002	ip3	21	Unknown protocol field in IP packet	10

设定各项阈值为 20，经过概念聚类，可以得到：(ip1,21,ip3,3500)，(ip1,21,外网地址，高端端口)，(ip2,21,ip3,21)，(外网地址，高端端口，ip3,21)，由用户选择在(ip2,21,ip3,21)，(外网地址，高端端口，ip3,21)类中进行关联；得到关联规则：

最小支持度	关联规则数	关联规则
84%	3	Dest - IP = ip3 ∧ Dest - Port = 21⇒ Event - Type = FTP fragment attempt

5 结束语

笔者在研究关联规则与聚类分析两种挖掘算法的基础上，针对单一的数据挖掘算法在安全管理中存在的缺陷，提出结合两种挖掘算法联合发现安全事件的方法，实验结果表明该算法的有效性。由于采用将聚类分析作为关联规则预处理的方法，准确率较高，但要比单独使用关联规则或聚类花费较多的时间，因此如何提高其高效性，将是笔者今后研究的一个重点。

参考文献

1 Wenke lee; Stolfo. S. J; K. W. Mok. A Data Mining Framework for Building Intrusion Deteciton Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium

2 R. Agrawal, T. Imielinaki, A. Swami. Mining association rules between sets of items in large database In proc of the ACM SIGMOD Conference on Management of Data, 1993:207 -216

3 David Hand, Heikki Mannila, Padhraic Smyth. Principles of Data Mining