文章编号：2095-2163(2019)04-0290-03　　中图分类号：TP311　　文献标志码：A

基于 Weka 平台的 R2L 攻击关联分析

郑继刚[1]，张静梅[2]

（1 保山学院 信息学院，云南 保山 678000；2 保山学院 图书馆，云南 保山 678000）

摘　要：远程主机的用户未授权访问攻击是网络攻击类型之一，该攻击类型出现在 KDDCup 数据集中，运用 Weka 对特征属性进行分析，挖掘结果显示不同属性特征间的联系。

关键词：Weka；数据挖掘；特征属性；关联分析

Analysis of R2L attack association based on Weka platform

ZHENG Jigang[1], ZHANG Jingmei[2]

（1 School of Information, Baoshan University, Baoshan Yunnan 678000, China;
2 Library of Baoshan University, Baoshan Yunnan 678000, China）

【Abstract】 Unauthorized access attacks of users on remote hosts are one of the types of network attacks, which occur in KDDCup datasets. Weka is used to analyze the characteristic attributes, and the mining results show the relationship between different attribute features.

【Key words】 Weka; data mining; characteristic attribute; association analysis

0　引　言

数据挖掘也称数据库中知识发现(knowledge discovery in database, KDD)[1]，从提出到现在一直得到了研究和应用领域的广泛关注。是目前重要研究课题之一。其从大量原始数据中挖掘出隐含的、有用的、尚未发现的信息和知识，帮助决策者寻找数据间潜在的有用知识。

远程主机用户未授权访问攻击(Remote to Local, R2L)，攻击是基于数据包负载的，数据包头部没有明显的频繁模式，单个数据包和正常连接区别不大，若采用传统检测方法很难提高检测率[2]。

本文采用数据挖掘 Weka 平台的关联规则算法[3]，依据 KDDCUP99 数据集的“KDDCUP. data_10_percent”子集[4]，挖掘出数据集中 R2L 攻击隐含的用户行为特征或规律，以指导入侵检测系统依据规则库对用户行为进行检测，根据检测结果采取不同的应对措施。

1　R2L 数据预处理

“KDDCUP.data_10_percent”子集中共有 1 126 条 R2L 攻击类型记录，每条记录有 41 个固定的特征属性和最后一个攻击类型标识。下载的数据集是 xls 格式的 excel 工作表，另存为 CSV 文件类型，在 Weka“Exploer”模块中，打开该 CSV 文件另存为 ARFF 文件类型，可视化结果如图 1 所示。

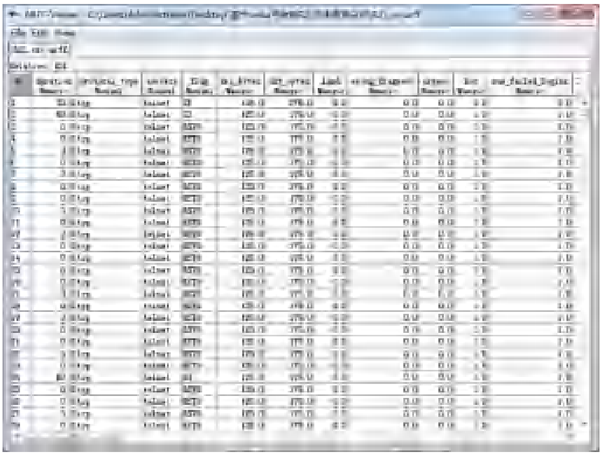

图 1　R2L 攻击类型 ARFF 格式示意图
Fig. 1　ARFF format diagram of R2L attack type

2　关联分析

在 Explorer 模块的关联规则(Associate)标签下，可以实现对数据集的关联分析操作，这里提供了 Apriori、FilteredAssociator、GeneralizedSequentialPatterns、HotSpot、PredictiveApriori、Tertius 等多种关联分析算法，本文选择 Apriori 算法对实验数据集进行了关联分析[5]。

每条攻击记录共有 42 个特征属性。除第 2、3、4、42 个属性是离散型外，其余 38 个属性均为数值型，如图 2 所示。

作者简介：郑继刚(1983-)，男，硕士，教授，主要研究方向：数据挖掘。
收稿日期：2019-04-22

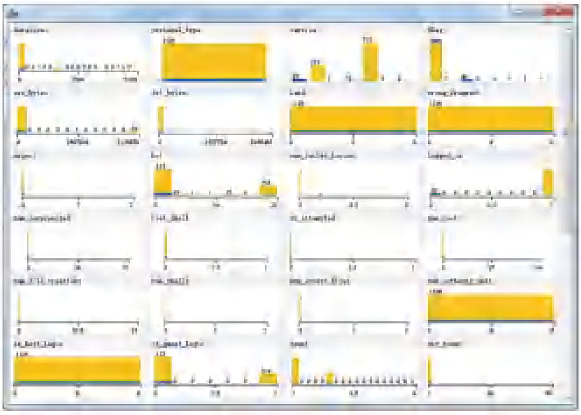

图2　R2L攻击类型可视化示意图

Fig. 2　Visualization of R2L attack types

借助Weka的“Filter树”，在“weka.filters.unsupervised.attribute.Discretize”中，设置attributeIndices属性为“1,5-41”，“bins”改成“3”，即设置为3段离散化值。勾选记录值完全相同的protocol_type、land、wrong_fragment、num_outbound_cmds、is_host_login等5个属性，2 s时间内与当前连接的流量特征、具有相同目标主机前100个连接。即第23~41个属性，并“Remove”以删除剩余18个属性。设置参数为“Apriori –N 20 –T 0 –C 0.9 –D 0.1 –U 1.0 –M 0.5 –S -1.0 –C -1”，前20条挖掘结果见表1。

表1　关联分析挖掘结果

Tab. 1　Result of association analysis mining

序号	挖掘结果
1	num_compromised='(-inf-12.666667]' 1123 ==> num_root='(-inf-18]' 1123　conf:(1)
2	num_shells='(-inf-0.666667]' 1123 ==> su_attempted='(-inf-0.333333]' 1123　conf:(1)
3	num_shells='(-inf-0.666667]' 1123 ==> num_root='(-inf-18]' 1123　conf:(1)
4	num_root='(-inf-18]' num_shells='(-inf-0.666667]' 1123 ==> su_attempted='(-inf-0.333333]' 1123　conf:(1)
5	su_attempted='(-inf-0.333333]' num_shells='(-inf-0.666667]' 1123 ==> num_root='(-inf-18]' 1123　conf:(1)
6	su_attempted='(-inf-0.333333]' num_root='(-inf-18]' 1123 ==> num_shells='(-inf-0.666667]' 1123　conf:(1)
7	num_shells='(-inf-0.666667]' 1123 ==> su_attempted='(-inf-0.333333]' num_root='(-inf-18]' 1123　conf:(1)
8	num_failed_logins='(-inf-1.666667]' num_compromised='(-inf-12.666667]' 1122 ==> num_root='(-inf-18]' 1122　conf:(1)
9	num_failed_logins='(-inf-1.666667]' num_shells='(-inf-0.666667]' 1122 ==> su_attempted='(-inf-0.333333]' 1122　conf:(1)
10	num_failed_logins='(-inf-1.666667]' num_shells='(-inf-0.666667]' 1122 ==> num_root='(-inf-18]' 1122　conf:(1)
11	num_compromised='(-inf-12.666667]' su_attempted='(-inf-0.333333]' 1122 ==> num_root='(-inf-18]' 1122　conf:(1)
12	num_compromised='(-inf-12.666667]' num_shells='(-inf-0.666667]' 1122 ==> su_attempted='(-inf-0.333333]' 1122　conf:(1)
13	num_compromised='(-inf-12.666667]' su_attempted='(-inf-0.333333]' 1122 ==> num_shells='(-inf-0.666667]' 1122　conf:(1)
14	num_compromised='(-inf-12.666667]' num_file_creations='(-inf-7]' 1122 ==> num_root='(-inf-18]' 1122　conf:(1)
15	num_compromised='(-inf-12.666667]' num_shells='(-inf-0.666667]' 1122 ==> num_root='(-inf-18]' 1122　conf:(1)
16	num_file_creations='(-inf-7]' num_shells='(-inf-0.666667]' 1122 ==> su_attempted='(-inf-0.333333]' 1122　conf:(1)
17	num_file_creations='(-inf-7]' num_shells='(-inf-0.666667]' 1122 ==> num_root='(-inf-18]' 1122　conf:(1)
18	num_failed_logins='(-inf-1.666667]' num_root='(-inf-18]' num_shells='(-inf-0.666667]' 1122 ==> su_attempted='(-inf-0.333333]' 1122　conf:(1)
19	num_failed_logins='(-inf-1.666667]' su_attempted='(-inf-0.333333]' num_shells='(-inf-0.666667]' 1122 ==> num_root='(-inf-18]' 1122　conf:(1)
20	num_failed_logins='(-inf-1.666667]' su_attempted='(-inf-0.333333]' num_root='(-inf-18]' 1122 ==> num_shells='(-inf-0.666667]' 1122　conf:(1)

根据挖掘结果，从中可以获取隐含在R2L攻击类型中不同属性特征间的联系：被迫妥协出现的次数num_compromised<13，超级用户root访问的数量num_root<18，shell提示符的数量num_shells<1，不执行“su”命令su_attempted，登录失败的次数，num_failed_logins<2，执行文件创建的数量num_file_creations<7。这些挖掘规则的置信度均为100%，如果降低置信度进行挖掘，会有更多的联系出现。

（下转第294页）