



基于模糊理论与关联规则的入侵检测模型

章坚武¹, 黄佳森¹, 周迪²

(1. 杭州电子科技大学, 浙江 杭州 310018; 2. 浙江宇视科技有限公司, 浙江 杭州 310018)

摘要: 利用 BV-Apriori 算法生成匹配规则库, 引入模糊集合技术解决连续型数据划分过程中边界过硬的问题, 完成特征之间关系的实时分析与规则库的更新, 搭建入侵检测 BVA-IDS (Boolean vector Apriori-intrusion detection system) 模型。研究表明, 相比顺序生成频繁项集的 Apriori 算法与已有文献的 Apriori-BR 算法, 本文的 BV-Apriori 算法挖掘效率显著地提高; 相比已有文献的检测模型, 本文的 BVA-IDS 模型在入侵检测指标上表现较好, 有较高的检测准确性与较低的误检率和漏检率。

关键词: 入侵检测; Apriori 算法; 布尔向量; 模糊理论

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019077

Intrusion detection model based on fuzzy theory and association rules

ZHANG Jianwu¹, HUANG Jiasen¹, ZHOU Di²

1. Hangzhou Dianzi University, Hangzhou 310018, China

2. Zhejiang Uniview Technologies Co., Ltd., Hangzhou 310018, China

Abstract: An intrusion detection model based on fuzzy theory and improved Apriori algorithm was proposed. The BV-Apriori algorithm was used to generate the matching rule base, and the problem of excessive boundary in the continuous data partitioning process was solved by fuzzy set technology. The real-time analysis of the relationship between features and the update of the rule base were completed, and the intrusion detection model BVA-IDS (Boolean vector Apriori-intrusion detection system) was built. The results show that the mining efficiency of the BV-Apriori algorithm is significantly improved when compared with the existing Apriori-BR algorithm, in addition, the BVA-IDS model also performs well on intrusion detection indicators with high detection accuracy, and low false positive rate and false negative rate.

Key words: intrusion detection, Apriori algorithm, Boolean vector, fuzzy theory

收稿日期: 2018-10-08; 修回日期: 2019-05-01

基金项目: 国家自然科学基金资助项目 (No.61772162, No.U1866209); 国家重点研发计划经费资助项目 (No.2016YFB0800201); 浙江省自然科学基金资助项目 (No.LY16F020016)

Foundation Items: The National Natural Science Foundation of China (No.61772162, No.U1866209), The National Key Research Development Program of China (No.2016YFB0800201), The Natural Science Foundation of Zhejiang Province of China (No.LY16F020016)



1 引言

随着网络技术的不断发展以及应用的深入,大量的多类型安全威胁正在增加,因此,有效的入侵检测系统^[1]已引起人们的关注。关联规则挖掘最早是 Agrawal 等人^[2]在 1993 年 SIGMOD 大会上提出的一个数据挖掘的重要研究课题,用于发现隐藏在大型数据集中令人感兴趣的联系。随后在 1994 年,由 Agrawal 和 Srikant^[3]提出第一个也是最经典的关联规则算法——Apriori 算法。

针对经典 Apriori 算法在生成频繁项集过程中,存在大量候选项集和多次遍历扫描数据库的缺陷: Aditya 等^[4]为了改进 Apriori 算法,提出了一种新方法来生成频繁模式。通过查找模式代替返回数据库,使得总组合的数量从 2^N 降低到 2^{N-2} , 以此减少内存使用量和执行时间。Li 等^[5]提出了一种称为 MAR-DPS 的算法,该算法具有一些深度修剪策略,包含 3 种方法来压缩频繁项集的大小,并减少生成新频繁项集的连接数,在面对不同的数据集时会选择适当的方法来生成 2-频繁项集,以提高算法效率。Xiao 等^[6]提出了一种基于标记者事务压缩的改进 Apriori 算法,该算法引入了新的限制参数 (support>1/2), 优化了关联规则的参数,减少了候选项集的数量和比较次数,使得计算时间和成本降低。黄玉蕾等^[7]提出一种基于位运算和倒排索引的 Apriori-BR 算法,该算法对频繁项集的倒排索引按照事务长度进行分组,利用位运算加快子集的检测,并在必要时动态删除无效的低维事务,以此提高挖掘效率。

关联规则挖掘在很多入侵检测系统中得到了应用。Chana 等^[8]提出了基于模糊关联规则的入侵检测和防御系统,该系统利用规则基和模式基进行检测,预防和预测 Web 服务攻击,以确定是否明确授予对正常事务的访问权限。张春琴等^[9]针对云环境提出一种基于模糊推理的网络入侵检测方法,该方法利用互信息特征选择对样本特征降

维,再利用改进的模糊 C 均值聚类和规则参数优化方法进行规则库的建立,最后基于规则库对输入连接数据进行模糊推理,具有较好的可行性和有效性。Jiao 等^[10]根据模糊逻辑理论,将模糊频繁项集转换为模糊关联规则,并生成模糊关联规则库来搭建异常检测模型,该模型有效地提高了检测准确率。Mabu 等^[11]描述了一种基于 GNP (genetic network programming) 的新型模糊类关联规则挖掘方法用于网络入侵检测。通过模糊集理论与 GNP 相结合,该方法可以处理既包含离散属性又包含连续属性的混合数据集,并提取有助于提高检测能力的重要类关联规则,以提高检测准确率。高妮等^[12]提出一种基于自编码网络的支持向量机入侵检测模型 AN-SVM。该模型采用多层无监督的限制玻尔兹曼机 (RBM) 将原始数据映射至低维空间,建立双向映射自编码网络结构,再采用 SVM 分类算法对所学习到的最优低维表示进行入侵识别。

在现有的网络入侵检测中,少有考虑流量自身特征之间相互关联的系统,而基于规则匹配的系统需要提取网络流量的特征和建立网络数据规则库,由于网络数据的增速十分快,因此需要对特征之间的关联关系进行高效挖掘。所以,为了建立流量自身特征之间关联关系的入侵检测系统,本文改变了 Apriori 算法由 1-频繁项集低到高顺序生成频繁项集,直到生成 k -频繁项集的方式,提出一种基于布尔向量插叙生成频繁项集的方式,再利用模糊集合技术解决连续性变量集合划分过程中边界过硬的问题,高效地生成关联规则库,完成特征之间关系的实时分析,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 BVA-IDS 模型。

2 Apriori 算法及其改进

2.1 Apriori 算法

定义 1 假设 $I=\{I_1, I_2, I_3, \dots, I_n\}$ 是所有项目的

集合, 若存在 $X \subset I$, 则称 X 为项目集 (itemset, 简称项集), 若是其中项目的计数为 k , 则称之为 k -项集。若其支持度不小于最小支持度, 则称之为 k -频繁项集。

定义 2 假设 $I = \{I_1, I_2, I_3, \dots, I_n\}$ 是目标事务库中所有项目构成的集合, $D = \{T_1, T_2, T_3, \dots, T_m\}$ 是事务库, 其中有 $T_i = \{I_{i1}, I_{i2}, I_{i3}, \dots, I_{ik}\}$, 并且 $I_{ij} \in I$, 称 T_i 为数据库中的事务。

定义 3 设 A, B 是不同的项集, 表达式 $A \Rightarrow B$ 称为规则, 其中 A, B 必须同时满足 $\{A, B | A \subset I, B \subset I, A \cap B = \emptyset\}$, A 为该规则的前提, B 为结果。

定义 4 如果目标事务库中有 x 条事务同时包含项目集 A 和 B , 那么称 x 为该项目集的支持度计数 (support_count, 简称支持数)。若 x 占总事务数 N 为 $S\%$, 那么称 $S\%$ 为 $A \Rightarrow B$ 的支持度 (support), 它反映该规则的有用性, 如式 (1) 所示:

$$\text{Support}(A \Rightarrow B) = \frac{\text{Support_count}(A \cup B)}{N} = S\% \quad (1)$$

定义 5 如果在目标事务库中包含项目集 A 的事务中有 $C\%$ 的事务同时也包含项目集 B , 那么称 $C\%$ 为 $A \Rightarrow B$ 的置信度 (confidence), 它反映该规则的确定性, 如式 (2) 所示:

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Support_count}(A \cup B)}{\text{Support_count}(A)} = C\% \quad (2)$$

定义 6 如果在目标事务库 D 中, A, B 为项集, 若存在 $A \Rightarrow B$ 的支持度 $S\%$ 和置信度 $C\%$ 分别不小于预先设定的最小支持度 minSup 和最小置信度 minCon, 则称 $A \Rightarrow B$ 为强关联规则。

性质 1 若 k -项集是频繁项集, 则该项集所有非空子集都是频繁项集。

性质 2 若 k -项集是非频繁项集, 则该项集所有超集都是非频繁项集。

Apriori 算法是一种自底向上的规则挖掘方

法, 算法实现的两个过程如下。

步骤 1 寻找支持度不小于最小支持度 minSup 的所有频繁项集, 该步骤中又包含相互融合的连接步和剪枝步, 由 $(k-1)$ -频繁项集生成 k -频繁项集, 最后生成所有的频繁项集。

步骤 2 由频繁项集产生强关联规则: 根据式 (2) 计算步骤 1 中所有频繁项集的置信度, 如果频繁项集的规则置信度不小于预先设定的最小置信度 minCon, 即强关联规则。

Apriori 算法的两个步骤中, 其计算量主要花费在第一步的挖掘频繁项集中, 因此, 本文仅针对 Apriori 算法第一步挖掘频繁项集的过程, 提出改进的 BV-Apriori (Boolean vector Apriori) 算法。

2.2 BV-Apriori 算法相关定义

定义 7 每个事务 T_x 的向量定义为:

$$T_x = (d_{x1}, d_{x2}, \dots, d_{xy}) \quad (3)$$

每个项 I_y 的向量定义为:

$$I_y = (d_{1y}, d_{2y}, \dots, d_{xy})^T \quad (4)$$

其中, d_{xy} 为第 x 个事务的第 y 项, $d_{xy} = \begin{cases} 0, & I_y \notin T_x \\ 1, & I_y \in T_x \end{cases}$,

I_y 的支持度计数为:

$$\text{Support_count}(I_y) = \sum_{k=1}^x d_{ky} \quad (5)$$

定义 8 将事务库 D 记为:

$$D = (I_1 \ I_2 \ \dots \ I_n) = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ d_{m1} & d_{m2} & \dots & d_{mn} \end{bmatrix} \quad (6)$$

其中, 依据定义 7 可以得到 d_{ij} 的值 (0 或 1)。

定义 9 项集 $\{I_i, I_j\}$ 的向量定义为 D_{ij} :

$$D_{ij} = D_i \wedge D_j = \begin{pmatrix} d_{1i} \wedge d_{1j} \\ d_{2i} \wedge d_{2j} \\ \vdots \\ d_{mi} \wedge d_{mj} \end{pmatrix} \quad (7)$$



其中,“ $\wedge$ ”是“逻辑与”运算符。这样,2-项集 $\{I_i, I_j\}$ 的支持度计数为:

$$\text{Support_count}\{I_i, I_j\} = \sum_{k=1}^m (d_{ki} \wedge d_{kj}) \quad (8)$$

定义 10 k -项集 $\{I_1, I_2, \dots, I_k\}$ 的向量定义为:

$$\begin{aligned} D_{12\dots k} &= D_1 \wedge D_2 \wedge \dots \wedge D_k \\ &= \begin{pmatrix} d_{11} \wedge d_{12} \wedge \dots \wedge d_{1k} \\ d_{21} \wedge d_{22} \wedge \dots \wedge d_{2k} \\ \vdots \\ d_{m1} \wedge d_{m2} \wedge \dots \wedge d_{mk} \end{pmatrix} \end{aligned} \quad (9)$$

这样, k -项集 $\{I_1, I_2, \dots, I_k\}$ 的支持度计数为:

$$\begin{aligned} \text{Support_count}\{I_1, I_2, \dots, I_k\} \\ = \sum_{p=1}^m (d_{p1} \wedge d_{p2} \wedge \dots \wedge d_{pk}) \end{aligned} \quad (10)$$

定义 11 将项 I_y 的向量中值为“1”所对应的事务 T_x 的向量做“逻辑与”运算:

$$\begin{aligned} Z'_y &= [(d_{k1}, d_{k2}, \dots, d_{ky}) \wedge \\ &(d_{k1}, d_{k2}, \dots, d_{ky}) \wedge \dots \wedge (d_{k1}, d_{k2}, \dots, d_{ky})] \end{aligned} \quad (11)$$

其中, k 的值是项 I_y 在数据库 D 中向量值为“1”对应行的数值。再将 Z'_y 向量中, I_y 前所对应的向量元素置0, 得到 Z_y , 定义 Z_y 是项 I_y 的布尔向量。

定义 12 在项 I_y 的布尔向量 Z_y 中, I_y 位后所有的0值对应的项组成一个集合, 该集合的所有非空子集中, 若为频繁项集, 则定义这些频繁项集为 I_y 的0-项集。

2.3 BV-Apriori 算法

BV-Apriori 算法思想: 扫描一次数据库将其转换成布尔矩阵(0或1), 对其列中“1”的个数进行累加, 计算其支持数, 删除小于最小支持数的无用列(项)并按支持度计数从小到大向右排序, 构造一个1-频繁项矩阵。利用定义11求出项 I_y 对应的布尔向量 Z_y , 其“1”值对应的项组成的集合是频繁项集, 根据性质1频繁项集的非空子集都是频繁项集, 那么该集合中含项 I_y 的非空子集均为频繁项

集(为避免重复生成频繁项集, 所以只取含项 I_y 的子集), 且它们的支持数均等于 I_y 的支持数。每次利用布尔向量 Z_y 生成频繁项集后, 将在其之前生成的符合条件的0-项集与之组合, 生成新的项集并计算支持数, 与最小支持数比较后得到频繁项集。依次遍历所有布尔向量, 至此生成所有的频繁项集。具体算法流程如下所示。

算法 1 BV-Apriori 算法

输入: 事务数据库 D 、最小支持度 minSup

输出: 满足最小支持度的所有频繁项集

$D = \text{trans2matrix}(\text{事务库 } D)$ //把事务库转换为0、1矩阵 D

$M = \text{ones}(1, \text{size}(D, 1)) \times D$ //计算各事物项的支持数

$ND = \text{getnewmatrix}(M, D)$ //删除、排序后生成 ND

$ND = \text{getnewmatrix}(M, D)$ //删除、排序后生成 ND

For $i = 1:y$ Do //循环生成布尔向量

$Z_i = \text{getbooleanvector}(ND, i)$

Return Z_i

For $j = 1:i$ Do //由布尔向量直接生成频繁项集

If $d_{xy} == 1$

Add I_y To $\{\}$

Frequent0 = combination($\{\}$)

Frequent1 = combination(Frequent0, 0-项集)

End If

If $\text{minSup of Frequent0, Frequent1} \geq \text{minSup}$

Return Frequent0, Frequent1

End If

End For

End For

根据Apriori算法的性质, 给出如下优化策略。

策略 1 由布尔向量 Z_y 直接生成的频繁项集Frequent0, 若其支持度等于最小支持数, 则不参与生成Frequent1的运算。

策略 2 在生成Frequent1的运算中, 若某0-

项集与某频繁项集组合生成非频繁项集,那么该0-项集及其超集不再与该频繁项集及其超集进行组合。

2.4 BV-Apriori 算法实例

输入 事务数据库 D (见表1), minSup=0.3

输出 频繁项集 (包含频繁项集及其支持数)

表1 事务数据库 D

TID	事务项集
001	$I_1, I_2, I_3, I_5, I_6, I_{11}$
002	I_1, I_3, I_7
003	I_5, I_9
004	I_1, I_3, I_4, I_5, I_7
005	$I_1, I_3, I_5, I_7, I_{11}$
006	I_5, I_{10}
007	I_1, I_2, I_3, I_5, I_6
008	I_1, I_4
009	I_1, I_3, I_5, I_7
010	I_1, I_5, I_8, I_9

步骤1 扫描数据库得到矩阵 D, 并计算得到每项的支持度计数:

$$D = \begin{matrix} & \begin{matrix} I_1 & I_2 & I_3 & I_4 & I_5 & I_6 & I_7 & I_8 & I_9 & I_{10} & I_{11} \end{matrix} \\ \begin{matrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \\ I_9 \\ I_{10} \\ I_{11} \end{matrix} & \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix} \\ & \begin{matrix} 8 & 2 & 6 & 2 & 8 & 2 & 4 & 1 & 2 & 1 & 2 \end{matrix} \end{matrix} \quad (12)$$

步骤2 因为最小支持度为 0.3, 事务数为 10, 所以最小支持度计数为 3。删除小于最小支持数的项 $I_2, I_4, I_6, I_8, I_9, I_{10}, I_{11}$, 并对生成的矩阵按支持度计数从小到大向右排序得到 1-频繁项矩阵 ND。

$$ND = \begin{matrix} & \begin{matrix} I_7 & I_3 & I_1 & I_5 \end{matrix} \\ \begin{matrix} I_7 \\ I_3 \\ I_1 \\ I_5 \end{matrix} & \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \\ & \begin{matrix} 4 & 6 & 8 & 8 \end{matrix} \end{matrix} \quad (13)$$

步骤3 将项 I_7 向量中值为 1 对应的事务 T_2, T_4, T_5, T_9 的向量做“逻辑与”运算, 得到:

$$Z_1 = [(1 \ 1 \ 1 \ 0) \wedge (1 \ 1 \ 1 \ 1) \wedge (1 \ 1 \ 1 \ 1) \wedge (1 \ 1 \ 1 \ 1)] = [1 \ 1 \ 1 \ 0] \quad (14)$$

同理可得:

$$Z_2 = [0 \ 1 \ 1 \ 0], Z_3 = [0 \ 0 \ 1 \ 0], Z_4 = [0 \ 0 \ 0 \ 1] \quad (15)$$

步骤4 为了更直观地得到算法结果, 将下述生成的频繁项集依次加入表 2 中。

(1) 布尔向量 Z_4 中, 数值为“1”的位所对应的项组成的集合为 $\{I_5\}$, 含项 I_5 的非空子集是 $\{I_5\}$, 支持度计数为 8;

(2) 布尔向量 Z_3 中, 数值为“1”的位所对应的项组成的集合为 $\{I_1\}$, 含项 I_1 的非空子集是 $\{I_1\}$, 支持度计数为 8;

(3) 布尔向量 Z_2 中, 数值为“1”的位所对应的项组成的集合为 $\{I_3, I_1\}$, 含项 I_3 的非空子集是 $\{I_3\}, \{I_3, I_1\}$, 支持度计数均为 6;

(4) 布尔向量 Z_1 中, 数值为“1”的位所对应的项组成的集合为 $\{I_7, I_3, I_1\}$, 含项 I_7 的非空子集是 $\{I_7\}, \{I_7, I_3\}, \{I_7, I_1\}, \{I_7, I_3, I_1\}$, 支持度计数均为 4。

步骤5

(1) 在 Z_3 之前生成的频繁项集为 $\{I_5\}$, 是 0-



项集, 将其加入 $\{I_1\}$, 得到新的项集 $\{I_1, I_5\}$, $\text{Support_count}\{I_1, I_5\}=1 \wedge 1+1 \wedge 0+0 \wedge 1+1 \wedge 1+1 \wedge 1+0 \wedge 1+1 \wedge 1+1 \wedge 0+1 \wedge 1+1 \wedge 1=6$, 符合支持度要求, 是频繁项集。

(2) 在 Z_2 之前生成的频繁项集中, $\{I_5\}$ 是 0-项集, 将其加入 $\{I_3\}$, $\{I_3, I_1\}$, 得到新的项集 $\{I_3, I_5\}$ 、 $\{I_3, I_1, I_5\}$, 同理得 $\text{Support_count}\{I_3, I_5\}=5$, $\text{Support_count}\{I_3, I_1, I_5\}=5$, 均是频繁项集。

(3) 在 Z_1 之前生成的频繁项集中, $\{I_5\}$ 是 0-项集, 将其加入 $\{I_7\}$ 、 $\{I_7, I_3\}$ 、 $\{I_7, I_1\}$ 、 $\{I_7, I_3, I_1\}$, 得到新的项集 $\{I_7, I_5\}$ 、 $\{I_7, I_3, I_5\}$ 、 $\{I_7, I_1, I_5\}$ 、 $\{I_7, I_3, I_1, I_5\}$, 同理得 $\text{Support_count}\{I_7, I_5\}=3$, $\text{Support_count}\{I_7, I_3, I_5\}=3$, $\text{Support_count}\{I_7, I_1, I_5\}=3$, $\text{Support_count}\{I_7, I_3, I_1, I_5\}=3$, 均是频繁项集。

至此, 生成所有频繁项集。

3 基于 BV-Apriori 算法的入侵检测模型

3.1 入侵检测模型总体架构

入侵检测技术传统上分为两大类, 分别为异常入侵检测 (anomaly detection) 和误用入侵检测^[13] (misuse detection)。其中, 异常入侵检测是建立一个用户行为的正常模式轮廓, 如果当前系统中的网络数据轮廓值与正常值的差异超出指定的阈值, 则存在入侵或者潜在的攻击行为, 该方法不依赖于攻击特征, 只需依据受检测的目标即可发现入侵行为。误用入侵检测是建立一个攻击模式

库或特征库, 该库中包含了所有已知的各种入侵方式、系统的漏洞和违规操作等信息, 将实时网络数据与模式库进行特征匹配, 如果匹配成功, 表明该数据是攻击数据, 系统中有非法行为存在。

本文将异常入侵检测和误用入侵检测两种方法相结合, 并引入模糊集合相关理论知识, 提出了一种新的入侵检测模型 BVA-IDS (Boolean vector Apriori-intrusion detection system), 其结构如图 1 所示。

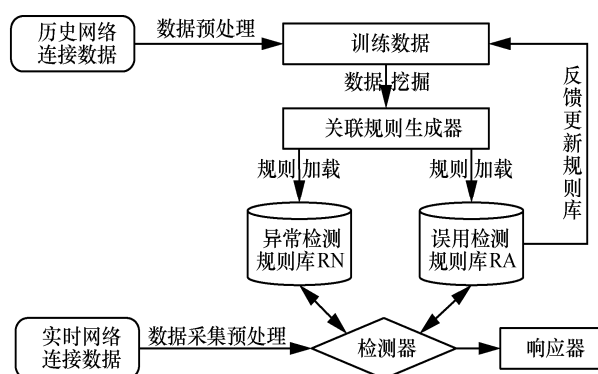


图1 入侵检测 BVA-IDS 模型

3.2 模糊离散化

为了解决连续性变量集合划分过程中边界过硬的问题, BVA-IDS 模型在数据离散化阶段引入模糊集合技术。根据定义的模糊隶属函数, 将连续型变量划分到特定的线性区间, 并根据隶属函数计算出相应的隶属度, 根据其隶属度按特定区间划分变量属性。

在复杂的网络连接数据中, 连接特征包含了

表2 频繁项集及支持数

项 (布尔向量)	$I_5 (Z_4)$	$I_3 (Z_3)$	$I_1 (Z_2)$	$I_7 (Z_1)$
对应频繁项集: 支持数	$\{I_5\}: 8$	$\{I_1\}: 8$ $\{I_1, I_5\}: 6$	$\{I_3\}: 6$ $\{I_3, I_1\}: 6$ $\{I_3, I_5\}: 5$ $\{I_3, I_1, I_5\}: 5$	$\{I_7\}: 4$ $\{I_7, I_3\}: 4$ $\{I_7, I_1\}: 4$ $\{I_7, I_3, I_1\}: 4$ $\{I_7, I_5\}: 3$ $\{I_7, I_3, I_5\}: 3$ $\{I_7, I_1, I_5\}: 3$ $\{I_7, I_3, I_1, I_5\}: 3$

连续时间、协议类型、传送的字节数等基本属性, 它们有离散型也有连续型。在已有的研究中, 大多数对连续型属性的处理是设定不同区间, 并将属性划分到特定的区间里。本文则将模糊集合概念引入离散化处理中, 由参考文献[11]可知, 模糊分类函数相关参数定义如下:

- β , 属性 I 在数据集中的平均值;
- γ , 属性 I 在数据集中的最大值;
- α , $\alpha = 2\beta - \gamma$ 。

3 个线性区间 (low, mid, high) 的隶属函数如下:

$$f_{\text{low}}(x) = \begin{cases} 1.0 & , x \leq \alpha \\ \frac{x - \beta}{\alpha - \beta} & , \alpha < x < \beta \\ 0 & , x \geq \beta \end{cases} \quad (16)$$

$$f_{\text{mid}}(x) = \begin{cases} 0 & , x \leq \alpha \\ \frac{x - \alpha}{\beta - \alpha} & , \alpha < x < \beta \\ 1.0 & , x = \beta \\ \frac{x - \gamma}{\beta - \gamma} & , \beta < x < \gamma \\ 0 & , x \geq \gamma \end{cases} \quad (17)$$

$$f_{\text{high}}(x) = \begin{cases} 0 & , x \leq \beta \\ \frac{x - \beta}{\gamma - \beta} & , \beta < x < \gamma \\ 1.0 & , x \geq \gamma \end{cases} \quad (18)$$

由上述隶属函数得到的模糊隶属函数如图 2 所示。

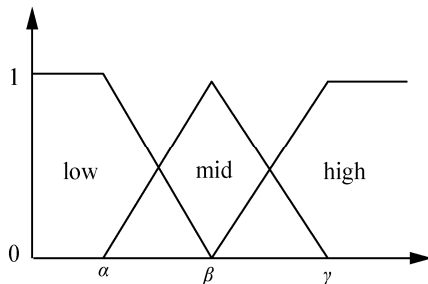


图2 模糊隶属函数

将数据预处理阶段得到的正常连接数据集 D_n

和攻击连接数据集 D_a 按上述隶属函数离散化处理后, 得到网络正常连接数据集 D_{nf} 和网络攻击连接数据集 D_{af} 作为规则库挖掘阶段的数据输入, 并通过 BV-Apriori 算法挖掘出关联规则库。

3.3 检测分类

本文的 BVA-IDS 系统包含使用规则库 RN 的异常检测和使用规则库 RA 的误用检测两部分。它们有相同的检测过程: 将测试数据集 D_t (由正常和攻击连接数据组成) 中的每一条网络连接数据先后与异常检测规则库 RN 与误用检测规则库 RA 对比, 若满足规则库 RN 则为正常连接数据, 若满足规则库 RA 则为攻击连接数据, 若两者均不满足则标记为未知连接, 进一步分析。

以网络连接数据与规则库 RN 比对为例, 具体的检测流程如图 3 所示。

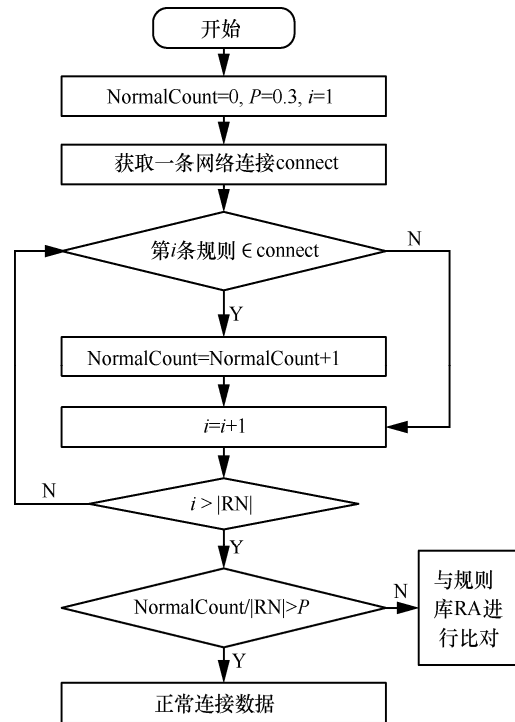


图3 检测流程

(1) 初始值: 满足规则数 $\text{NormalCount} = 0$, 规则数占比 $P = 0.3$, 规则编号 $i = 1$ 。

(2) 获取一条网络连接数据 connect。



(3) 判断第 i 条规则是否为 connect 的子集，是则跳至步骤 (4)，否则跳至步骤 (5)。

(4) NormalCount = NormalCount+ 1。

(5) $i = i + 1$ 。

(6) 判断 $i > |RN|$ (规则库中规则总数)，是则跳至步骤 (7)，否则跳至步骤 (6)。

(7) 判断 NormalCount/ $|RN| \geq P$ ，是则为正常连接数据，否则与规则库 RA 再进行比对。

3.4 系统评价标准

衡量入侵检测系统性能有检测率 (detection rate, DR)、误检率 (false positive rate, FPR) 和漏检率 (false negative rate, FNR) 这 3 个重要指标。对这 3 个指标的计算，需要用到下列变量。

- TN (true negative): 检测判断为正常网络连接行为，实际上为正常网络连接行为的数目。
- FP (false positive): 检测判断为攻击网络连接行为，实际上为正常网络连接行为的数目。
- TP (true positive): 检测判断为攻击网络连接行为，实际上为攻击网络连接行为的数目。
- FN (false negative): 检测判断为正常网络连接行为，实际上为攻击网络连接行为的数目。

(1) 检测率

检测率是入侵检测系统对网络连接行为判断正确的数目占总测试集数目的比例，一定程度上反映了入侵检测系统的准确性。计算式为：

$$DR = \frac{TN + TP}{TP + TN + FP + FN} \times 100\% \quad (19)$$

(2) 误检率

误检率是入侵检测系统错误地将网络连接正常行为判断成攻击连接行为的数目占总测试集中正常连接行为数目的比例。计算式如下：

$$FPR = \frac{FP}{TN + FP} \times 100\% \quad (20)$$

(3) 漏检率

漏检率是入侵检测系统错误地将网络连接攻击行为判断成正常连接行为的数目占总测试集中

攻击连接行为数目的比例。计算式如下：

$$FNR = \frac{FN}{FN + TP} \times 100\% \quad (21)$$

4 实验结果与分析

4.1 BV-Apriori 算法验证

为了验证 BV-Apriori 算法的性能，在 Intel Core i7-6 700K，CPU 主频 4.00 GHz，8 GB 内存，Ubuntu 16.04 LTS 操作系统环境下，采用 MATLAB 实现经典 Apriori 算法、参考文献[7]中的 Apriori-BR 算法与本文所提的 BV-Apriori 算法。实验数据为 accidents 数据集，该数据取自 FIMI 的数据库，包含 1991—2000 年 Flanders 地区的交通事故记录，大小为 33.8 MB，共 340 183 条事务数，有 572 个项目，平均事务长度为 45。

实验一：在相同的支持度 minSup=0.45 条件下，使用 accidents 数据集的 5 万、10 万、15 万、20 万、25 万、30 万条事务数进行测试，实验结果如图 4 所示。

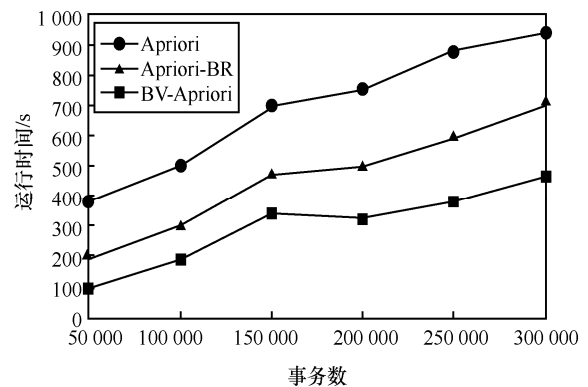


图 4 accidents 数据集 0.45 支持度测试结果

由图 4 可知，实验一中的 3 种算法的运行时间均随事务数的增加而增长，但当事务数从 15 万增加到 20 万时，3 种算法时间增加得并不明显，是因为两个数据集产生的频繁项集相差不多，因此连接产生候选项集与计算支持度耗时也相差不多。在不同的事务数时，BV-Apriori 算法的运行时间均少于经典 Apriori 算法和 Apriori-BR 算法，

其挖掘效率显著地提高。

4.2 BVA-IDS 模型验证

为了验证 BVA-IDS 模型的性能,使用数据集 KDD CUP 99 中 10% 的训练数据子集和测试数据子集。数据集中每一条网络连接用 41 个特征来描述。但在 41 个特征中,其大多数特征属性对于入侵检测是不适用的,多余的属性不但没有意义而且会消耗更多的计算空间与时间资源。因此,选择了能体现原始数据集的最小子集属性,进行关联规则挖掘和入侵检测。根据已有研究^[14]的研究成果得知,对于 KDD CUP 99 中适用于入侵检测的普遍最优属性集索引为{1, 3, 4, 5, 6, 10, 14, 23, 25, 30, 35}的特征属性。

本节设计 3 个实验,在相同的参数条件下与已有文献中的 Crisp Data Mining (Crisp DM) 模型、Fuzzy Data Mining (Fuzzy DM) 模型^[11]和 AN⁵-SVM 模型^[12]的入侵检测指标进行对比分析。

实验二:从 10% 的训练子集中随机筛选 9 137 条正常连接,设置 minCon=1, minSup=0.4, $P=0.3$,经过训练建立异常检测规则库,从 10% 的测试子集中随机筛选 194 条正常连接和 579 条攻击连接,与规则库进行匹配,并计算相应的检测指标。重复操作 10 次,最后实验结果由平均值求得。

实验三:从 10% 的训练子集中随机筛选 1 637 条攻击连接,设置 minCon=0.7, minSup=0.4, $P=0.6$,经过训练建立误用检测规则库,从 10% 的测试子集中随机筛选 750 条正常连接和 240 条攻击连接,与规则库进行匹配,并计算相应的检测指标。重复操作 10 次,最后实验结果由平均值求得。

实验四:参考文献[12]的表 1,随机抽样生成数据量相同的 5 个样本集,设置 minCon、minSup、 P 3 个参数分别与实验二、三相同,进行误用和异常检测实验。

3 个实验的结果分别见表 3~表 5。

表 3 异常检测系统实验结果

检测指标	模型类别		
	BVA-IDS	Crisp DM	Fuzzy DM
DR	96.38%	90.30%	94.40%
FPR	5.15%	10.30%	7.20%
FNR	3.11%	9.50%	5.00%

表 4 误用检测系统实验结果

检测指标	模型类别		
	BVA-IDS	Crisp DM	Fuzzy DM
DR	98.99%	98.30%	98.70%
FPR	0.40%	0.67%	0.53%
FNR	3.01%	5.00%	3.75%

表 5 异常和误用检测系统实验结果

模型/指标		数据集				
		DS1	DS2	DS3	DS4	DS5
AN ⁵ -SVM	DR	98.23%	94.73%	97.48%	98.31%	96.82%
	FPR	1.225%	4.48%	3.67%	1.71%	3.11%
BVA-IDS	DR	98.39%	95.51%	97.60%	98.57%	97.51%
	FPR	1.09%	3.71%	3.57%	1.46%	2.93%

由表 3 和表 4 可知,在异常检测系统中,BVA-IDS 入侵检测模型的检测率高于 Crisp Data Mining 模型和 Fuzzy Data Mining 模型;在误用检测系统中,也略微高于其他两个模型。并且 BVA-IDS 入侵检测模型的误检率和漏检率均低于两者。由此可看出本文提出的 BVA-IDS 入侵检测模型具有较高的检测率与较低的误检率和漏检率。

由表 5 可知,在整体的检测系统中,BVA-IDS 入侵检测模型的检测率在不同数据集规模均一定程度上优于 AN⁵-SVM 模型,且有较低的误检率。

在不同的阈值条件下,得到 BVA-IDS 模型、Crisp Data Mining 模型和 Fuzzy Data Mining 模型的接收器特性曲线(receiver operating characteristic, ROC),如图 5 所示。

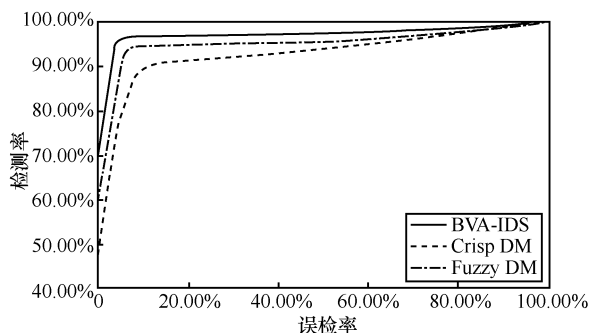


图5 ROC曲线

ROC曲线是误检率随检测率的变化曲线，能评估入侵检测的性能，其曲线下面积越大，系统的性能越好。由图5可知，BVA-IDS系统的曲线下面积最大，而且曲线始终在另两个系统曲线的上面，因此其性能始终为最优。

综上所述，本文提出的BVA-IDS入侵检测模型相比已有文献的检测模型，提升了入侵检测的整体准确性。

5 结束语

本文利用矩阵和位运算，提出了改进的BV-Apriori算法，该算法扫描一次数据库将其转换成一个布尔矩阵，对其列中“1”的个数进行累加计算支持度，删除小于最小支持度的列构造一个1-频繁项矩阵，利用位运算求出1-频繁项对应的布尔向量，借助布尔向量生成频繁项集，依次与0-项集组合，按插叙的方式生成所有的频繁项集。再利用改进的BV-Apriori算法和模糊集合技术生成高效的关联规则库，搭建入侵检测BVA-IDS模型。最后，设计对比实验验证了BV-Apriori算法效率，并计算BVA-IDS模型的入侵检测性能指标，与已有文献的相应指标对比分析，在异常检测和误用检测中，BVA-IDS模型均表现出较好的检测准确性，尤其是降低了误检率和漏检率。

参考文献:

[1] DENNING D E. An intrusion-detection model[J]. IEEE Trans-

actions on Software Engineering, 1987, SE-13(2): 222-232.

- [2] AGRAWAL R, IMIELINSKI T, SWAMI A N. Mining association rules between sets of items in large databases[C]//The 1993 ACM SIGMOD International Conference on Management of Data, May 26-28, 1993, New York, USA. New York: ACM Press, 1993: 207-216.
- [3] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[M]. New York: ACM Press, 1994: 487-499.
- [4] ADITYA S P, HEMANTH M, LAKSHMIKANTH C K, et al. Effective algorithm for frequent pattern mining[C]//International Conference on IoT and Application, May 19-20, 2017, Nagapattinam, India. Piscataway: IEEE Press, 2017.
- [5] LI L, LI Q, WU Y, et al. Mining association rules based on deep pruning strategies[J]. Wireless Personal Communications, 2018(2): 1-25.
- [6] XIAO M, YIN Y, ZHOU Y, et al. Research on improvement of apriori algorithm based on marked transaction compression[C]//Advanced Information Technology, Electronic and Automation Control Conference, March 25-26, 2017, Chongqing, China. Piscataway: IEEE Press, 2017: 1067-1071.
- [7] 黄玉蕾, 罗晓霞, 林青. 基于位运算和倒排索引的关联规则挖掘算法[J]. 电信科学, 2015, 31(11): 81-86.
- [8] HUANG Y L, LUO X X, LIN Q. An association rule mining scheme based on bit operation and reverse index[J]. Telecommunications Science, 2015, 31(11): 81-86.
- [9] CHANA G Y, CHUAA F F, LEEB C S. Fuzzy association rules vs fuzzy associative patterns in defending against Web service attacks[C]//International Conference on Fuzzy Systems and Knowledge Discovery, August 13-15, 2016, Changsha, China. Piscataway: IEEE Press, 2016: 524-529.
- [10] 张春琴, 谢立春. 云环境中改进 FCM 和规则参数优化的网络入侵检测方法[J]. 电信科学, 2018, 34(1): 72-79.
- [11] ZHANG C Q, XIE L C. Network intrusion detection method based on improved FCM and rule parameter optimization in cloud environment[J]. Telecommunications Science, 2018, 34(1): 72-79.
- [12] JIAO W, LI Q. Anomaly detection based on fuzzy rules[J]. International Journal of Performability Engineering, 2018, 14(2): 376-385.
- [13] MABU S, CHEN C, LU N, et al. An intrusion-detection model based on fuzzy class-association-rule mining using genetic network programming[J]. IEEE Transactions on Systems Man & Cybernetics Part C, 2010, 41(1): 130-139.
- [14] 高妮, 高岭, 贺毅岳, 等. 基于自编码网络特征降维的轻量级入侵检测模型[J]. 电子学报, 2017(3): 730-739.
- [15] GAO N, GAO L, HE Y Y, et al. A lightweight intrusion detection model based on auto encoder network with feature reduction[J]. Acta Electronica Sinica, 2017(3): 730-739.
- [16] KUMAR S. Classification and detection of computer intru-

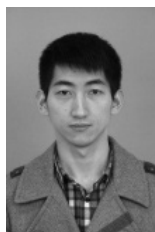
sions[D]. West Lafayette: Purdue University, 1995.

- [14] MADBOULY A I, GODY A M, BARAKAT T M. Relevant feature selection model using data mining for intrusion detection system[J]. International Journal of Engineering Trends & Technology, 2014, 9(10): 804-811.

[作者简介]



章坚武（1961—），男，杭州电子科技大学通信工程学院教授、博士生导师，主要研究方向为移动互联网、多媒体通信技术、网络安全等。



黄佳森（1992—），男，杭州电子科技大学通信工程学院硕士生，主要研究方向为数据挖掘、网络安全等。



周迪（1975—），男，浙江宇视科技有限公司高级工程师、宇视研究院院长，主要研究方向为视频安全、人工智能等。