

入侵检测中的混合特征选择算法研究

黎银环, 钟艳花

(广东省江门职业技术学院, 广东 江门 529000)

摘 要: 在开放式网络中, 高维混合特征的冗余或不相容属性会降低网络入侵检测的效率. 为提高入侵检测系统的响应性能, 提出一种混合特征选择方法, 利用粗糙集形式化描述入侵检测的特征选择, 采用信息熵和平均权重分别定义数值型和字符型特征的重要度. 算法产生降序特征序列, 采用 K-means 聚类算法评估出优化特征子集. 在 KDD CUP99 数据集上的仿真实验表明, 算法有效选择特征子集并缩短了检测时间.

关键词: 粗糙集; 特征选择; 信息熵; 平均权重; 重要度

中图分类号: TP393.08

文献标识码: A

文章编号: 1000-7180(2012)09-0051-04

Research on Hybrid Feature Selection Method in Intrusion Detection

LI Yin-huan, ZHONG Yan-hua

(Jiangmen Vocational & Technical Polytechnic, Jiangmen 529000, China)

Abstract: In an open network, redundant or incompatible attributes of high-dimensional mixed features reduce the efficiency of network intrusion detection. In order to improve the response performance of intrusion detection system, this paper proposes a hybrid feature selection method. Rough set theory is used to description for intrusion detection feature selection. Information entropy and average weight are used to define importance of numeric features and character features. After generating a descending feature sequence, K-means clustering algorithm is used to evaluate the optimal feature subset. Simulation experiment is done in KDDCUP99. It shows that the method is effective to select feature subset and shorten the detection time.

Key words: rough set; feature selection; information entropy; average weight; importance

1 引言

入侵检测系统属于积极主动的安全防护技术, 在网络安全中至关重要, 在不影响网络性能的情况下对网络进行实时监测. 网络中的原始数据属于多特征、高噪声的混合型数据集, 海量的审计数据和冗余特征会增加入侵检测的系统开销和干扰分类结果, 甚至影响网络性能. 为保证开放式网络入侵检测的时效性和提高系统的检测性能, 要对数据集的特征加以选择. 主要根据信息熵的重要度、条件特征和类别特征的互信息等在输入特征集中选择一个对目标概念最相关的特征子集, 并对特征选择算法进行性能评估. 特征选择已成为网络入侵检测的研究热

点并取得许多成果^[1-7]. 文献[2]提出采用互信息的方法消除冗余特征和用归纳值约简算法消除冗余属性值. 文献[3]利用粗糙集中的属性约简算法剔除冗余属性. 文献[4]建立一种基于条件信息熵的网络攻击特征选择算法. 但上述算法并未考虑到网络数据为混合型数据集的特点, 导致检测效率不高. 本文尝试综合粗糙集和信息熵在特征选择中速度较快的优点, 采用粗糙集理论形式化描述了网络入侵检测的特征选择, 利用信息熵定义数值型特征的重要度和平均权重定义字符型特征的重要度, 形成基于重要度排序的特征序列, 采用 K-means 聚类算法对所选择特征子集的测试数据集进行聚类, 根据聚类效果的准确性选择出优化的特征子集.

收稿日期: 2012-02-06; **修回日期:** 2012-03-26

基金项目: 国家自然科学基金项目(60903025)

2 基于重要度的混合特征特征选择

2.1 用粗糙集理论描述特征选择

粗糙集理论的主要思想是:在保持知识库分类能力不变的前提下,通过知识约简,剔除其中不相关或冗余的特征,并且不丢失基本信息,从而导出概念的分类规则,提高决策的精度^[8].在处理模糊和不确定性问题中有着广泛的应用.入侵检测的过程是根据所获取的大量数据推测系统行为的过程,特征选择的目的是设计分类方法来降低描述数据集的特征维数,可以用粗糙集的相关理论来描述入侵检测特征选择.参考文献[4,6,8],本文的相关定义和定理修正如下.

定义 1 入侵检测特征选择决策表是一个五元组 $S=\langle U, C, D, V, f \rangle$, 其中, U 为网络中告警记录集合, $C=\{c\}$ 为告警记录的条件特征集, $D=\{d\}$ 为类别决策特征性集, 表示系统当前处于被攻击或正常状态. $V=\{V_x\}$ 是特征值的集合, V_x 为特征 x 的值域, $f:U \times (C \cup D) \rightarrow V$ 是一个信息函数, 指定告警记录集合 U 中的每个记录的特征值.

定理 1 特征选择决策表 $S=\langle U, C, D, V, f \rangle$ 存在约简关系 H , 对于 $\exists c \in C$, 使得 $H(\langle d \rangle | C) = H(\langle d \rangle | C - \{c\})$, 即决策特征 d 的取值不会受到 C 中的某个特征 c 的影响, 则认为特征 c 是 C 中的冗余特征, 可以被约简; 否则 c 是 C 中必要的特征; $B = \{C - \{c\} \mid (B \text{ 为非空集})\}$ 由所有必要的特征组成, B 是 C 相对于 $\{d\}$ 的优化特征子集.

粗糙集理论的数据约简算法要求特征值是逻辑型或者离散型的, 而网络数据通常包括连续型特征和离散型特征, 如协议和服务名称等为字符型离散特征. 因此, 特征选择前要先对连续型数据特征预处理.

2.2 特征选择算法中的相关定义

定义 2 信息熵: 假设告警数据集 U 根据条件特征集 C 被分为 n 个部分, $U = \{U_j\} (j=1, 2, \dots, n)$; 分类值为 $D = \{D_i\} (i=1, 2, \dots, k)$, 则 U 被划分后的分类信息熵 $E(U)$ 定义为:

$$E(U) = - \sum_{j=1}^n p(U_j, U) \left(\sum_{i=1}^k p(D_i, U) \log p(D_i, U) \right) \quad (1)$$

其中, $p(U_j, U) = \frac{|U_j|}{|U|}$ 是告警数据集 U_j 占总数据集 U 的比例; $p(D_i, U) = \frac{|D_i|}{|U|}$ 是告警数据集 U

中类别特征值为 D_i 的告警数据记录所占的比例, $|D_i|$ 是指类别特征值为 D_i 的记录数.

因不同数值特征值的取值范围和精度不同, 划分计算区间目的是对数值型特征值进行离散化处理, 本文参考文献[6,9]对计算区间的划分修正定义如下.

定义 3 根据数值特征集 $X = \{x_j\} (j=1, 2, \dots, n) (X \in C)$ 中 n 个数值型特征值升序排序, 将整个数值域划分为 n 个区间 $R = \{r_j\} (j=1, 2, \dots, n)$, 每个区间 r_j 包括两个观察值. 权值表示在该区间的特征值数 $|x_j|$; 类值表示某个特征在权类区间的记录数所属的类型 D_i .

第 j 个区间的权值和类值关系表示为:

$$R_j(X | D) = (|x_j|, D_i) \quad (2)$$

定义 4 结合信息熵和计算区间的划分, 数值型特征的重要度修正计算如下:

$$E = \sum_{j=1}^n p(r_j, U) \left(\sum_{i=1}^k p(D_i, r_j) \left| \log \left(\frac{p(D_i, r_j)}{p(D_i, U)} \right) \right| \right) \quad (3)$$

其中, $D_i (i=1, 2, \dots, k)$ 表示分类; $p(r_j, U)$ 计算区间 r_j 在数据集 U 中所占的比例; $p(D_i, r_j)$ 计算某一分类 D_i 在区间 r_j 中所占的比例; $p(D_i, U)$ 计算某一分类 D_i 在数据集 U 中所占的比例. E 表示特征对分类 D_i 影响的重要程度. 特征与分类越相关, 则其重要度越大, 当重要度趋于零时表明特征对分类 D_i 没有影响, 此特征可以约简.

文献[10]在对文本信息的集中度和分散度分析时发现某本文特征项与其在类别中出现的频率间存在一定关系: 特征项在越少的类别中出现, 特征项的权重越大; 出现频度小的特征具有较大的互信息. 网络数据包中的协议和服务名称等属性为文本, 参照此特点, 本文修正定义字符特征值的平均权重计算公式如下:

定义 5 设字符型特征数据集 $X = \{x_i\}$ (特征值个数 $i=1, \dots, n$) ($X \in C$), 则字符型特征在各分类中的平均权重为:

$$W = \frac{k \sum_{i=1}^n h_i MI(x_i, D_i)}{\sum_{i=1}^n k_i} \quad (4)$$

其中 h_i 为特征 x_i 在分类 D_i 中出现的次数, K 为数据集中类别总数, K_i 为字符型特征出现的类别个数.

$MI(x_i, D_i)$ 为特征 x_i 与分类 D_i 的互信息. 字

符型特征值在某类别中出现的频率越少,其互信息越大,出现的类别个数越小,此特征项的平均权重越大,对分类的影响就越大。

2.3 混合特征选择算法

特征选择算法通常包括选择特征子集与特征子集评估两个阶段。针对网络数据包的特点,混合特征选择算法先对数值型和字符型特征预处理,分别计算数值型和字符型特征的重要度,特征按其重要度形成降序特征序列。在对所选特征子集进行评估时要依据一定评估标准,文献[11]利用 K-means 算法聚类,采用类间平均离散度与类内离散度之和的平均值之比作为评估值筛选特征子集,达到无监督学习的效果。此算法的计算量较大,为提高系统的响应性能,评估阶段的算法需简洁有效。本文利用 K-means 算法对所选特征子集的数据集进行聚类,将聚类结果与实际类别属性比较统计分类精度。根据分类精度 V_i 进行评估,分类精度越大,表明分类越准确,选出优化特征子集。

算法描述如下:

第一阶段:混合特征预处理形成特征序列

输入:训练记录集,特征集

输出:特征序列链表

① 初始化各链表

CreateList(Dlist); // 创建数值特征链表

CreateList(Clist); // 创建字符特征链表

CreateList(rlist); // 创建区间链表

CreateList(list); // 创建特征链表

② 处理数值型特征,建立计算区间

For($i=2$; $i \leq p$; $i++$) // p 为特征维数

如果是数值型特征

For($j=2$; $j \leq m$; $j++$) // m 为记录数

{按定义 3 将数值特征值划分计算区间存入区间链表 rlist,统计权值并标记类别}

③ 计算特征重要度并排序

For($i=2$; $i \leq p$; $i++$)

For($j=2$; $j \leq m$; $j++$)

如果是数值型特征

{根据公式 3 计算数值特征的重要度 E ;

按 E 将数值特征降序插入序列 Dlist}

否则是字符型特征

{根据公式 4 计算字符特征的平均权重 W ;

按 W 将字符特征降序插入序列 Clist}

④ 从字符和数值特征链表中取出特征依次加入特征序列 list;

⑤ 输出基于重要度降序排列的特征序列 list。

第二阶段:子集评估产生优化特征子集

输入:特征序列,测试记录集

输出:优化特征子集

For($i=2$; $i \leq p$; $i++$)

① 从测试数据集中选择特征序列中前 i 个特征值组成新测试数据子集 U_i ;

② 利用 K-mean 算法对新测试记录集 U_i 进行聚类,得到聚类结果集;

③ 根据聚类结果集统计各分类的分类精度作为特征子集的评估值;

For($q=1$; $q \leq k'$; $q++$)

// k' 为本次聚类数

{ $V_1[q]=0$;

$V_2[q]=0$;

For($j=1$; $j \leq m$; $j++$)

根据聚类结果判别测试记录所属类别 d_1 , 记数 $V_1[q]++$;

For($j=1$; $j \leq m$; $j++$)

{获取测试数据集 U_i 中记录的实际类别属性 d_2 ; 如果 $d_1=d_2$, 则记数 $V_2[q]++$ }

$$V_i = \sum_{q=1}^{k'} V_2[q] / \sum_{q=1}^{k'} V_1[q]$$

④ 筛选出各分类中分类精度 V_i 最大的子集作为优化特征子集。

算法复杂性分析:在本算法中,记录数为 m 条,特征维数为 p ;数据预处理形成特征序列的最大时间复杂度约为: $O(2mp^2)$;在评估特征子集算法中使用简洁的 K-means 算法,计算复杂度约为 $O(mKt)$,其中 K 为聚类数, t 为迭代次数,评估特征子集阶段的计算复杂度约为 $O(2pmKt)$ 。算法总的计算复杂度较低。

3 实验结果与分析

仿真实验硬件平台采用 CPU2.0GHz,内存为 2GB 的个人计算机,软件平台为 Windows XP,数据库为 MS SQL Server 2005,在 VC++6.0 编程环境下实现程序设计。以 10% KDDCUP99 数据集作为验证数据源进行仿真实验,数据集总有 494 015 条记录,每一条记录由 41 个条件属性和 1 个类别属性构成,其中条件特征包含 34 个数值特征和 7 个字符特征。入侵类型总 22 种攻击,其中主要包括 4 大类网络攻击类型:Denial-Of-Service(DoS), Surveillance or probe(Probe), User to Root(U2R), Re-

mote to Local(R2L). 正常记录约占 19.68%, 异常数据记录占约 80.32%. 在实际网络环境中大量记录是正常的行为记录, 入侵行为记录只占大约 1.5%~2%. 因此, 仿真实验参照实际网络数据的比例抽取 2 组数据集, 每组 50000 条记录分别作训练数据和测试数据, 如表 1, 测试数据需要包含类别属性. 针对 4 大攻击类型, 本文算法产生的优化特征子集序号如表 2 所示.

表 1 数据集

样本集	正常记录数	异常记录数	异常比例(%)
训练集	48945	1055	2.11
测试集	48960	1040	2.08

表 2 各分类的优化特征子集

攻击类型	最优特征子集序号	特征数
R2L	3,22,35,5,23,37,24,33,40	9
U2R	3,13,16,32,17,29,15	7
Probe	3,40,27,4,5,26,28,34	8
DoS	2,3,24,36,38,37,35,26,5	9

实验中分别选择 26 个特征属性和表 2 的优化特征子集各组成新数据集进行效率测试, 得出各类网络攻击在不同特征数下的检测效率, 如表 3 所示.

表 3 各类攻击在不同特征数下的检测效率

攻击类型	特征数	检测率 (%)	检测率对比 (%)	测试时间 (ms)	时间对比下降 (%)
R2L	26	77.21	0.01	3318	15.34%
	9	77.23		2803	
U2R	26	85.71	0.51	3204	18.42%
	7	86.24		2612	
Probe	26	97.27	0.04	4511	17.73%
	8	97.30		3710	
DoS	26	99.15	0.13	14807	32.28%
	9	99.24		10023	

由表 3 中看出利用优化特征集与使用 26 个特征进行检测评估相比, 各类攻击的检测率没有下降, U2R 的检测率相对提高了 0.51%, 而各类攻击的检测时间明显降低, Dos 攻击的检测时间下降了 32.28%, 表明本文算法能有效约简特征并缩短了系统的检测时间. 另外, 通过本实验表明在判断入侵行为时, 特征维数并不是越多检测结果越准确. 因为冗余和不相关特征不仅会加重入侵检测系统的存储负

担和耗费系统的检测时间, 也可能会对聚类判决带来不必要的干扰, 影响检测性能.

4 结束语

在开放式的复杂网络环境中, 入侵检测是保障系统安全的关键技术之一. 特征选择是入侵检测系统的重要环节, 特征选择算法的性能优劣直接影响入侵检测系统的检测性能. 本文引入粗糙集描述入侵检测特征选择, 采用信息熵和平均权重分别计算数值型和字符型特征的重要度形成特征序列, 基于 K-means 算法聚类效果评估特征子集. 仿真实验验证了本算法是快速有效的. 今后的研究将进一步优化算法, 提高系统的检测率和降低误检率.

参考文献:

- [1] 陈友, 程学旗, 李洋, 等. 基于特征选择的轻量级入侵检测系统[J]. 软件学报, 2007, 18(7): 1639-1651.
- [2] 许晓东, 古一, 朱士瑞. 入侵检测中的数据约简研究[J]. 计算机工程, 2011, 37(11): 170-172.
- [3] 叶清, 确陈亚莎, 黄高峰. 基于粗糙集和证据推理的网络入侵检测模型[J]. 计算机工程, 2011, 37(5): 164-166.
- [4] 陈波, 于冷, 吉根林. 基于条件信息熵的网络攻击特征选择技术[J]. 小型微型计算机系统, 2008, 29(3): 428-432.
- [5] 余生晨, 王树, 高晓燕, 等. 网络入侵检测系统中的最佳特征组合选择方法[J]. 计算机工程, 2008, 34(1): 150-152.
- [6] 史志才, 夏永祥. 基于知识约简的网络入侵特征提取[J]. 计算机工程, 2011, 37(5): 134-136.
- [7] 张雪芹, 顾春华. 一种网络入侵检测特征提取方法[J]. 华南理工大学学报, 2010, 38(1): 81-86.
- [8] Pawlak Z. Rough set theory and its application to data analysis[J]. Cybernetics and Systems, 1998, 29(7): 661-688.
- [9] 李杨寰, 高峰, 李腾, 等. 特征选择中信息熵的应用[J]. 计算机工程与应用, 2009, 45(15): 54-57.
- [10] 陈艳秋, 熊耀华. 新型快速中文文本分类器的设计与实现[J]. 计算机工程与应用, 2009, 45(22): 53-55.
- [11] 吴剑. 基于特征选择的无监督入侵检测方法[J]. 计算机工程与应用, 2011, 47(26): 79-82.

作者简介:

黎银环 女, (1975—), 硕士, 讲师. 研究方向为网络与信息安全.

钟艳花 女, (1974—), 硕士, 副教授. 研究方向为智能算法、嵌入式系统.