

基于快速属性约简的网络入侵特征选择

牟琦, 龚尚福, 毕孝儒, 匡向阳

(西安科技大学计算机学院, 西安 710054)

摘要: 高维网络数据中的无关属性和冗余属性会导致入侵检测速度慢及效率低下。为解决该问题, 提出一种基于快速属性约简的网络入侵特征选择方法。以网络数据的条件属性与类别属性之间的互信息为度量去除无关属性, 采用基于粗糙集正区域的属性重要性计算公式作为启发信息, 设计一种快速属性约简算法去除网络数据的冗余属性, 实现网络入侵特征子集的优化选择。在 KDD CUP1999 数据集上的仿真实验结果表明, 该方法能有效去除网络数据中的无关属性和冗余属性, 具有较高的入侵检测率和较低的误报率。

关键词: 互信息; 粗糙集; 属性约简; 特征选择; 网络入侵检测

Network Intrusion Feature Selection Based on Fast Attribute Reduction

MU Qi, GONG Shang-fu, BI Xiao-ru, SHE Xiang-yang

(School of Computer, Xi'an University of Science and Technology, Xi'an 710054, China)

【Abstract】 Aiming to problem that independent and redundant attributes of high dimensional network data cause classification algorithms' slow detection speed and low detection rate in network intrusion detection, this paper presents a feature selection method for network intrusion based on fast attribute reduction. It adopts Mutual Information(MI) between condition and label attributes of network data as measure to discard independent attributes, then a formula for measuring attribute importance based on positive region of rough set is applied as heuristic information to design a fast attribute reduction algorithm, which removes redundant attributes of network data to realize optimal selection of feature subset of network intrusion. Simulation experiment is done in KDDCUP1999. Result shows that the method is more effective in discarding independent and redundancy attributes and it has higher intrusion detection rate and lower false positive rate.

【Key words】 Mutual Information(MI); rough set; attribute reduction; feature selection; network intrusion detection

DOI: 10.3969/j.issn.1000-3428.2011.17.037

1 概述

在网络入侵检测中, 高维数据中无关属性和冗余属性的存在使分类检测算法的检测速度慢、检测率低。粗糙集理论能够在保证原始数据集分辨能力不变的前提下, 去除其中的无关属性和冗余属性。因此, 一些学者将其应用到网络入侵特征选择中。文献[1]采用基于条件信息熵的属性约简算法(Entropy Attribute Reduction, EAR)实现网络入侵特征选择。文献[2]采用基于改进条件信息熵的属性约简算法(Improved Entropy Attribute Reduction, IEAR), 完成网络入侵特征选择。文献[3]提出基于量子粒子群优化(Quantum Particle Swarm Optimization, QPSO)的属性约简算法, 实现网络入侵特征子集的优化选择。但以上算法都存在时间复杂度高的缺陷, 其中 EAR 算法时间复杂度为 $O(|C|^2|U|^2)$, IEAR 算法时间复杂度为 $O(|C|^2|U|^2)$, QPSO 算法引入随机搜索策略, 其时间复杂度较前 2 个算法更高。

本文提出一种基于快速属性约简(Fast Attribute Reduction, FAR)的网络入侵特征选择方法, 该方法在计算属性的互信息量^[4]的基础上, 设计了一个时间复杂度为 $\max(O(|C^-||U|), O(|C^-|^2|U/C^-|))$ 的快速属性约简算法, 实现网络入侵特征的优化选择。

2 基于快速属性约简的网络入侵特征选择

2.1 粗糙集基本概念

定义 1 定义五元组 $S = (U, C, D, V, F)$ 是一个网络连接决

策表, 其中, $U = \{u_1, u_2, \dots, u_n\}$ 为网络连接的非空样本集合; C 为条件属性, 即网络连接样本的属性集合; D 为类别属性(决策属性), 即网络连接的攻击类型集 $a \in C \cup D$; V_a 是属性 a 的值域; $f: U \times (C \cap D) \rightarrow V$ 是一个信息函数, 它指定 U 中每一样本 u 的属性值。

定义 2 在网络连接决策表 $S = (U, C, D, V, F)$ 中, 设 $U/D = \{D_1, D_2, \dots, D_k\}$ 为决策属性 D 对样本集 U 的划分, $U/P = \{P_1, P_2, \dots, P_m\}$ 为条件属性子集 $P(P \subseteq C)$ 对样本集 U 的划分, 称 $POS_P(D) = \cup P(D_i), D_i \in U/D$ 为 P 关于 D 的正区域。

定义 3 在网络连接决策表 $S = (U, C, D, V, F)$ 中, 若 $\forall B \subseteq C, POS_B(D) = POS_C(D)$ 且 B 相对于 D 是独立的, 则称 B 是 C 相对于 D 的属性约简。

定理 1 在网络连接决策表 $S = (U, C, D, V, F)$ 中:

$$POS_C(D) = \bigcup_{x \in U / C \wedge \forall x, y \in X \Rightarrow f(x, D) = f(y, D)} X$$

定义 4 在网络连接决策表 $S = (U, C, D, V, F)$ 中, 设:

$$U/C = \{[u_1]_C, [u_2]_C, \dots, [u_m]_C\}$$

$$U' = \{u_1', u_2', \dots, u_m'\}$$

由定理 1 可设:

基金项目: 陕西省自然科学基金资助项目(2009JM7007)

作者简介: 牟琦(1974—), 女, 副教授, 主研方向: 网络安全, 数据库技术; 龚尚福, 教授; 毕孝儒, 硕士研究生; 匡向阳, 副教授

收稿日期: 2011-03-04 **E-mail:** bi_xiao_ru@sina.com

$$POS_C(D) = [u_{i1}]_C \cup [u_{i2}]_C \cup \dots \cup [u_{it}]_C$$

其中, $\forall u'_{is} \in U' \wedge |[u'_{is}]_C / D| = 1 (s = 1, 2, \dots, t)$, 记为:

$$U'_{POS} = \{u'_1, u'_2, \dots, u'_m\}$$

$$U'_{NEG} = U' - U'_{POS}$$

称 $S' = (U', C, D, V, F)$ 为简化的网络连接决策表。

定义 5 在网络连接决策表 $S = (U, C, D, V, F)$ 中, $S' = (U', C, D, V, F)$ 为其简化决策表, 对于 $\forall B \subseteq C$, 定义:

$$POS_B'(D) = \bigcup_{X \in U'/B \wedge X \subseteq U'_{POS} \wedge |X/D|=1} X$$

定理 2 在网络连接决策表 $S = (U, C, D, V, F)$ 中, $S' = (U', C, D, V, F)$ 为其简化决策表, 对于 $\forall B \subseteq C$, 如果 $POS_B'(D) = U'_{POS}$, 则 $POS_B(D) = POS_C(D)$ 。

定理 3 在网络连接决策表 $S = (U, C, D, V, F)$ 中, $P \subseteq C$, $\forall a \in (C - P)$, 则 $U/(P \cup \{a\}) = \bigcup_{X \in U'/P} (X/\{a\})$ 。

定理 4 在网络连接决策表 $S = (U, C, D, V, F)$ 中, $S' = (U', C, D, V, F)$ 为其简化决策表, 则 $\forall a \in Core(C)$ 为核属性的充分必要条件为 $POS'_{C-\{a\}}(D) \neq U'_{POS}$ 。

2.2 以互信息量为尺度的无关属性去除

设 $C = \{c_1, c_2, \dots, c_m\}$ 为网络连接决策表的属性集, 则条件属性 c_i 和类别属性 D 的互信息(Mutual Information, MI) $I(c_i; D)$ 反映了两属性之间的共有信息量, 值越大, 则两者之间的相关性越强, 对网络入侵的正确检测的贡献越大。因此, 可设互信息量阈值 θ , 定义 $I(c_i; D) < \theta$ 的条件属性为无关属性, 并予以舍弃。

2.3 属性重要性的定义与推导

在网络连接决策表 $S = (U, C, D, V, F)$ 中, $S' = (U', C, D, V, F)$ 为其简化决策表, 则属性 $\forall a \in (C - P)$ 重要性定义为:

$$sig_P(a) = |U'_{P \cup \{a\}} - U'_P|, P \subseteq C$$

由定理 3 可推出:

$$sig_P(a) = \left| \bigcup_{\substack{X \in U'/P \\ ((X \subseteq U'_{POS} \wedge |X/D| \neq 1) \vee X \subseteq U'_{POS} \wedge \\ y \in X/\{a\} \wedge y \subseteq U'_{POS} \wedge |y/D|=1)}} X \cup \right. \\ \left. \bigcup_{\substack{X \subseteq U'_{NEG} \wedge y \in X/\{a\} \wedge y \subseteq U'_{NEG}}} X \right|$$

2.4 属性约简算法

属性约简算法具体如下:

输入 网络连接决策表 $S = (U, C, D, V, F)$, MI 阈值 θ

输出 网络入侵最优属性子集 R

- (1) $S' = \text{ReduceIndependentAttribute}(S, \theta)$
- (2) $U/C' = \text{CalculateInd}(C')$
- (3) $U' = \text{GetReducedObjectSet}(U/C')$
- (4) $Core(C') = \text{CalculateCore}(S')$
- (5) $R = Core(C')$
- (6) for each $\alpha \in C' - R$
- (7) $sig_R(\alpha) = \text{CalculateAttributeSignificance}(R, \alpha)$;
- (8) end;
- (9) $sig_R(\alpha') = \max\{sig_R(\alpha)\}$
- (10) $U' = U' - B_R(\alpha') - NB_R(\alpha')$
- (11) $C' = C' - R$
- (12) if $U' \neq \emptyset$
- (13) $U'_{POS} = U'_{POS} - B_R(\alpha')$;
- (14) $U'_{NEG} = U'_{NEG} - NB_R(\alpha')$;
- (15) go to (5)
- (16) else
- (17) output R
- (18) end

算法中主要函数功能如下:

ReduceIndependentAttribute (S, θ): 计算条件属性与决策属性之间的互信息量, 将 MI 小于阈值的条件属性作为无关属性去除, 并将新的网络连接决策表 $S' = (U, C', D, V, f)$ 返回; C' 为舍弃无关属性后的网络数据特征子集。

CalculateInd (C'): 根据基数排序思想计算条件属性集对样本集 U 的划分 U/C' , 并将其返回。

GetReducedObjectSet (U/C'): 根据定理 4 计算简化决策表 $S' = (U', C', D, V, f)$ 的核属性, 并将其返回。

CalculateAttributeSignificance (R, α): 根据式(1)计算属性 α 重要性, 并将其作为返回值。

2.5 算法时间复杂度

该算法的时间复杂度分析如下:

(1) 计算条件属性与决策属性之间的互信息时间复杂度为 $O(|C|)$ 。

(2) 计算划分 (U/C') 时间复杂度为 $O(|C' - 1| \|U\|)$ 。

(3) 计算核属性时间复杂度为 $O(|C' - 1|^2 |U/C'|)$ 。

(4) 计算 $sig_R(\alpha)$ 时间复杂度为 $O(|U' - U'_R|)$, 算法第(6)步~第(8)步时间复杂度为 $O(|C' - R| \|U' - U'_R\|)$, 第(6)步~第(15)步时间复杂度为:

$$O(|C' - 1| \|U'\|) + O(|C' - 1| \|U' - U'_R\|) + \dots + \\ O(|C' - R_k| \|U' - U'_{R_k}\|)$$

其中, R_k 为约简属性集, 因此其最坏时间复杂度为:

$$O(|C' - 1| \|U'\|) + O(|C' - 1| \|U'\|) + \dots + O(|C' - 1| \|U'\|) = \\ O(|C' - 1|^2 \|U'\|) = O(|C' - 1|^2 |U'/C'|)$$

因此, 本文算法的最差时间复杂度为:

$$\max(O(|C' - 1| \|U'\|), O(|C' - 1|^2 |U'/C'|))$$

3 实验与结果分析

3.1 实验数据集与参数设置

实验采用 KDDCUP1999^[5]数据集, 在对选取训练和测试样本离散化的基础上, 形成实验数据集, 如表 1、表 2 所示。

表 1 实验数据集 1

攻击类型	训练样本集			测试样本集		
	样本数	正常样本数	攻击样本数	样本数	正常样本数	攻击样本数
DoS	12 967	9 764	3 203	2 596	1 953	643
Probe	7 551	6 631	920	1 512	1 327	185
R2L	5 840	5 157	683	1 171	1 034	137
U2R	1 378	1 358	20	692	679	13

表 2 实验数据集 2

攻击类型	训练样本集			测试样本集		
	样本数	正常样本数	攻击样本数	样本数	正常样本数	攻击样本数
DoS	3 749	2 449	1 300	751	490	261
Probe	2 582	2 289	293	517	458	59
R2L	2 498	2 257	241	501	452	49
U2R	1 250	1 225	25	262	246	16

实验采用支持向量机(Support Vector Machine, SVM)作为分类检测算法, 在此基础上采用 Matlab 7.0 实现了本文算法。其中, SVM 的核函数采用 RBF 函数, 核参数 g 和惩罚系数 C 采用交叉验证参数寻优方法获取。

3.2 FAR 算法有效性实验与分析

在实验数据集 1 上, 应用本文算法实验结果见表 3、表 4。

表 3 去除无关属性后的特征子集序号

攻击类型	MI 阈值	特征子集序号
DoS	0.12	{2, 3, 4, 5, 6, 12, 13, 25, 26, 29, 37, 38, 39, 40}
Probe	0.25	{4, 5, 6, 27, 28, 34, 40, 41}
R2L	0.10	{3, 5, 6, 10, 22, 23, 24, 33, 35, 37, 40}
U2R	0.05	{1, 3, 10, 13, 14, 16, 17, 29, 33}

表4 采用 FAR 算法属性约简后的最优特征子集

攻击类型	网络数据最优特征子集
DoS	protocol-type, service, dst-bytes, flag
Probe	src-bytes, dst-bytes, flag, srv-count, dst-host-same-srv-count
R2L	service, dst-bytes, dst-host-srv-count
U2R	duration, service, dst-bytes, dst-host-count

由表3可以看出,在4种攻击类型中,去除无关属性后的特征集数目明显减少,尤其是在Probe攻击中,其特征数目由41个减少到8个,减少了80.49%。由表4可知,经FAR算法属性约简后,在4种攻击类型中,去除无关和冗余属性后,网络数据特征集数目由41个减少为3个~5个。

为测试FAR算法的有效性,实验分别采用未进行属性约简的表1样本集和经FAR属性约简后的样本集对SVM训练,获取检测结果如表5、表6所示。

表5 未进行属性约简的SVM检测结果

攻击类型	属性个数	训练时间/s	检测时间/s	检测率/(%)	误报率/(%)
DoS	41	4.546	1.594	99.89	0.01
Probe	41	1.735	0.625	96.24	0.35
R2L	41	1.687	0.568	52.17	0.00
U2R	41	0.313	0.281	50.00	0.00

表6 采用 FAR 算法属性约简后的SVM检测结果

攻击类型	属性个数	训练时间/s	检测时间/s	检测率/(%)	误报率/(%)
DoS	4	1.281	0.250	100.00	0.00
Probe	5	0.786	0.175	100.00	0.03
R2L	3	0.652	0.172	91.96	3.29
U2R	4	0.031	0.016	91.67	0.00

通过表5、表6可以看出,经FAR属性约简后,SVM分类算法在4类攻击检测中的训练和检测时间大幅减少,特别是在DoS攻击检测中,SVM训练和检测时间分别减少了71.82%和84.32%。同时,SVM算法在4类攻击检测上的检测率有明显提高,并保持了很低的误报率,尤其是在R2L和U2R攻击检测中,SVM检测率分别提高了78.19%和83.34%,且误报率保持在0~3.29%的低水平。

3.3 3种算法在网络入侵特征选择中的对比实验

实验分别采用EAR算法、IHAR^[6]算法和FAR算法对数据集2进行属性约简,实现网络入侵特征选择,结果如表7所示。在此基础上,测试3种算法对SVM检测性能的影响,结果如表8所示。

表7 3种算法的属性约简结果

攻击类型	属性约简时间/s			约简后的特征子集数		
	EAR	IHAR	FAR	EAR	IHAR	FAR
DoS	801.10	12.55	0.071	8	6	4
Probe	441.36	8.45	0.063	13	6	2
R2L	204.28	6.34	0.204	14	10	4
U2R	1.90	1.71	0.047	10	8	3

表8 3种算法对SVM的网络入侵检测结果

攻击类型	属性约简算法	训练时间/s	检测时间/s	检测率/(%)	误报率/(%)
DoS	EAR-SVM	0.656	0.203	99.12	0.01
	IHAR-SVM	0.437	0.135	99.11	0.00
	FAR-SVM	0.435	0.013	100.00	0.14
Probe	EAR-SVM	0.235	0.078	99.10	0.01
	IHAR-SVM	0.218	0.063	100.00	0.00
	FAR-SVM	0.087	0.070	100.00	0.00
R2L	EAR-SVM	0.282	0.095	70.00	0.88
	IHAR-SVM	0.156	0.084	96.00	7.28
	FAR-SVM	0.108	0.009	96.00	0.77
U2R	EAR-SVM	0.094	0.016	85.71	4.86
	IHAR-SVM	0.024	0.015	85.44	4.98
	FAR-SVM	0.007	0.010	90.88	4.90

由表7可知,与EAR、IHAR算法对比,在4种攻击类型中,FAR算法的约简时间最少,特别是在DoS攻击检测中,FAR算法的约简时间是0.071s,较EAR算法和IHAR算法下降801.03s和12.48s。由表8可以看出,经本文FAR算法对网络入侵特征选择后,SVM的检测率、误报率均优于采用EAR算法和IHAR算法,特别是U2R攻击检测中,在本文算法下,SVM检测率分别提高了6.03%和6.37%,而误报率未明显上升。同时,经本文FAR算法进行网络入侵特征选择后,SVM训练时间和检测时间均少于采用EAR算法和IHAR算法,尤其是在Probe攻击检测中,SVM训练时间分别减少62.98%和60.09%;检测时间分别减少91.03%和88.89%。

4 结束语

针对高维网络数据中的无关属性和冗余属性致使分类算法入侵检测性能不高的问题,本文提出一种基于快速属性约简的网络入侵特征选择方法,实验结果显示该方法能有效地约简网络数据特征,改善分类算法的入侵检测性能。今后将确定无关属性互信息量阈值,并对分类算法的可扩展性进行实验和分析。

参考文献

- [1] 段丹青,陈松乔,杨卫平,等.使用粗糙集和支持向量机检测入侵[J].小型微型计算机系统,2008,29(4):627-630.
- [2] 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.
- [3] 汪世义,陶亮,王华彬.基于QPSO属性约简在NIDS中的应用研究[J].微电子学与计算机,2010,27(1):120-122.
- [4] 何绍荣,梁金明,何志勇.基于互信息和关系积理论的特征选择方法[J].计算机工程,2010,36(13):257-259.
- [5] Maxion R A. KDD99Cupdataset[EB/OL]. [2010-07-07]. <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>.
- [6] 徐章艳,杨炳儒,宋威.一种快速计算HU差别矩阵属性约简算法[J].小型微型计算机系统,2008,29(10):1820-1827.

编辑 陆燕菲

(上接第112页)

参考文献

- [1] Akyildiz L F, Wang Xudong, Wang Weilin. Wireless Mesh Networks: A Survey[J]. Computer Networks, 2005, 47(4): 445-487.
- [2] 黄东平,王华勇.动态门限秘密共享方案[J].清华大学学报:自然科学版,2006,46(1):102-105.
- [3] 徐颖.无线Mesh网的接入控制和密钥管理[D].西安:西安电子科技大学,2007.
- [4] Pedersen T. A Threshold Cryptosystem Without a Trusted Party[C]//Proc. of the 10th Annual International Conference on Theory and Application of Cryptographic Techniques. Berlin, Germany: Springer-Verlag, 1991: 522-526.

- [5] Shamir A. How to Share a Secret[J]. Communications of the ACM, 1979, 22(11): 612-613.
- [6] Okamoto T, Pointcheval D. The Gap-problems: A New Class of Problems for the Security of Cryptographic Schemes[C]//Proc. of Public Key Cryptography Conference. Berlin, Germany: Springer-Verlag, 2001.
- [7] Shamir A. Identity-based Cryptosystems and Signature Schemes[C]//Proc. of International Cryptology Conference. Berlin, Germany: Springer-Verlag, 1985: 47-53.
- [8] 李光松,韩文报.基于签密的Ad Hoc网络密钥管理[J].计算机工程与应用,2005,41(12):160-164,167.

编辑 陆燕菲

