基于关联规则的网络入侵检测方法

陈洪泉，霍志凯

(海军大连舰艇学院信息与通信工程系 辽宁 大连 116018)

【摘要】介绍了基于关系代数理论的ORAR关联规则算法，分析了在KDD CUP 99中选择训练数据集和选择特征的基本方法，并在此基础上利用ORAR算法进行了频繁3、4、5、6项集入侵模式的挖掘，将挖掘结果应用于测试数据集的入侵检测，从检测的准确率和误检率两个方面较为系统地对不同的频繁项集检测的结果进行了比较，得到了检测效果最好的频繁项集，仿真结果对于入侵检测方法的进一步研究具有积极的借鉴意义。

关 键 词 频繁模式； 入侵检测； KDD99； ORAR

中图分类号 TP393.08 **文献标识码** A doi:10.3969/j.issn.1001-0548.2009.z1.018

Association Rules Based Network Intrusion Detection Method

CHEN Hong-quan and HUO Zhi-kai

(Departement of Information and Communication Engineering, Dalian Naval Academy Dalian Liaoning 116018)

Abstract ORAR association rules algorithm based on relation algebra theory is introduced. The basic method of selecting disciplined data set and features in KDD CUP 99 is analyzed. With ORAR algorithm, the mining aims at frequent three item sets, frequent four item sets, frequent five item sets, and frequent six item set. The mining patterns are used to test data collection, the results are compared according to the accuracy rate and true false rate, and the best frequent item set is achieved.

Key words frequent pattern; intrusion detection; KDD99; ORAR

1 网络入侵检测技术

网络入侵检测技术作为防火墙的有效补充，在网络安全防护中具有重要的作用，传统的基于入侵模式特征匹配的检测方法具有检测准确率高的优点，但其在检测未知入侵方式时则显得力不从心。

对入侵检测的研究可追溯到20世纪80年代。文献[1]第一次详细阐述了入侵检测的概念。文献[2]研究出了第一个入侵检测系统模型。自此之后，入侵检测技术的研究一直是学者研究的重点和热点。目前实际使用较多的入侵检测技术还是基于入侵模式匹配的传统的入侵检测方法[3]，该方法具有检测准确率高的优点，但在新的入侵方法不断涌现的网络空间，入侵特征的提取往往滞后于入侵方式的更新。为了适应入侵检测的需要，基于数据挖掘技术研究新的入侵检测方法成为研究的热点。本文基于数据挖掘中的关联规则方法，从大量的入侵数据中发现频繁入侵模式，并以此作为入侵检测标准对KDD99测试数据集进行了检测，得到了较为理想的效果，该方法对于入侵检测技术的进一步研究具有重要的参考价值。

2 关联规则算法

文献[4]最早提出关联规则挖掘；文献[5]提出经典的关联规则挖掘；文献[6]独立开发了使用剪枝方法的变形算法；文献[7]提出了统计分布CD和数据分布DD的并行Apriori算法。在关联规则挖掘的扩充中，文献[8]研究了关联规则的尺度化并行数据挖掘算法。作为数据挖掘中较为经典的关联规则算法，Apriori在数据挖掘中具有里程碑的作用，但其具有扫描次数多、候选项集大、运算时间长的缺点，不适合海量数据环境下的关联规则挖掘。针对Apriori算法的不足之处，文献[9]基于关系代数理论，利用关系矩阵及相关运算给出了搜索大项集的基于关系代数理论的优化的关联规则挖掘算法ORAR，该算法只需扫描数据库一次，克服了Apriori算法需要多次扫描数据库的缺点。该算法具有良好的并行性，本文将ORAR算法用于入侵频繁模式挖掘。

收稿日期： 2009 ↖09 ↖ 10

作者简介：陈洪泉(1977 ↖)，男，硕士，讲师，主要从事网络信息安全方面的研究.

3　基于ORAR的入侵检测仿真

3.1　数据集选取

为了降低仿真过程的复杂度，选用KDD CUP 99数据集[10]的“kddcup.data-10-percent”子集作为实验数据集。在“kddcup.data-10-percent”数据集中共有22种入侵类型，并且各种入侵方式的个数不同。

表1　仿真选取入侵类型及抽取的数目

所选入侵形式	数据集中的数目	抽样数目
back	2 203	2 000
ipsweep	1 247	1 000
neptune	107 201	2 000
portsweep	1 040	1 000
satan	1 589	1 000
smurf	280 790	2 000
warezclient	1 020	1 000

3.2　特征项的选取

特征选择对入侵检测的结果有很大的影响，不是所有的特征对数据处理都有用，有的还会对处理结果产生负面影响。因此，网络数据特征属性的正确选取对数据处理十分重要。

通过对数据集的分析可以发现，有些特征项与入侵类型的关联性较小。这些特征项的存在不仅会影响计算的速度，还会产生一些无用的规则影响检测效率。因此，在关联规则挖掘的过程中，要把这些项去掉。

仿真中选择了41个特征项中的18个特征进行频繁项集挖掘，这18个特征为：source bytes、destination bytes、hot、logged in、is guest login、Count、srv count、serror rate、srv serror rate、same srv rate、srv diff host rate、dst host count、 dst host srv count、dst host same srv rate、dst host diff srv rate、dst host same src port rate、dst host srv diff host rate、dst host serror rate。

3.3　仿真结果

在仿真过程中使用ORAR算法对数据集进行频繁项集挖掘。

为了使得到的频繁项集与入侵类型之间有较大的关联性，在计算的过程中将支持度设为t=0.95。t=0.95时所得到的各入侵类型相关的频繁3、4、5、6项集的数目如表2所示。

表2　不同入侵类型的频繁项集数目

入侵类型	频繁-3项集	频繁-4项集	频繁-5项集	频繁-6项集
back	35	35	21	7
ipsweep	20	15	6	1
neptune	120	210	252	210
portsweep	56	70	56	28
satan	56	69	52	22
smurf	286	715	1 287	1 716
warezclient	10	5	1	0

表2中，与各入侵类型相关的频繁项集可以作为一个规则库，根据这个规则库可以对数据的类型进行分类。在分类的过程中，让一条测试数据与规则库进行匹配。当匹配度大于一定的阈值时，就可以判断该数据的类型。

根据频繁项集所组成的规则库，可以判断一条测试数据的类型，从而实现入侵检测。在仿真的过程中，是通过从“kddcup.testdata.labeled”数据集中选取数据进行检测的。

通过入侵检测仿真实验，可以得到频繁项集构成的规则库对入侵类型判断的准确率及误检率。

准确率是指对一定数量的网络数据能够正确地判断其入侵类型的比例。

误检率是指将一种类型的数据判断为另一种类型的比例。

在计算准确率的过程中，将判断的阈值设为0.9，即当一个连接的数据与某种类型的规则库的匹配度达到0.9以上时，则将这条连接判定为该类型的入侵。然后根据不同类型的规则库分别对从测试集中提取的数据进行类型判断。

对于每种入侵类型，分别从测试集中抽取300条连接对规则库进行测试，根据测试的结果得出准确率。不同类型入侵的频繁3、4、5、6项集作为规则库时对连接类型判断的准确率如表3所示。

表3　不同频繁项集对入侵检测的准确率

入侵类型	频繁-3项集	频繁-4项集	频繁-5项集	频繁-6项集
back	0.44	0.44	0.44	0.45
ipsweep	0.98	0.98	0.95	0.95
neptune	0.96	0.96	0.96	0.95
portsweep	0.91	0.91	0.91	0.92
satan	0.99	0.99	0.99	0.98
smurf	0.99	0.99	0.99	0.98

准确率的表示形式为$r=L/M$，其中r为准确率；L为判断正确的连接数；M为总的连接数。从实验结果中可以发现，基于关联规则的入侵检测有较高的准确率。由于测试集中没有warezclient类型的连接，因此没有对warezclient类型连接判断的准确率。

对于一个性能较好的规则库来说，不仅要有较高的检测准确率，还要有较低的误检率。在计算误检率时，是用不同类型连接的规则库分别对normal类型的连接进行检测。

实验过程中，将报警的阈值设为0.2，即当一个normal类型的连接与某种类型规则库的匹配度大于0.2时，就认为是误检。

仿真时从测试集中抽取300条normal类型的连接进行测试。不同入侵类型的频繁3、4、5、6项集作为规则库时对连接类型判断的误检率如表4所示。

误检率的表示形式为$s=N/P$，其中s为误检率；N为发生误检的连接数目；P为抽取的连接的数目。

表4 不同频繁项集对入侵检测的误检率

入侵类型	频繁-3项集	频繁-4项集	频繁-5项集	频繁-6项集
back	0.40	0.08	0.07	0.05
ipsweep	0.28	0.28	0.02	0.01
neptune	0.17	0.02	0.02	0.01
portsweep	0.22	0.22	0.02	0.02
satan	0.03	0.03	0.01	0.01
smurf	0.18	0.06	0.03	0.03
	0.44	0.16	0.16	0.16

通过对表3数据的分析可以发现，随着频繁项集的项数的增加，检测准确率有下降的趋势，但是不明显。因此就准确率来说，用频繁3、4、5、6项集作为规则库对入侵进行检测时，效果相差不大，都有很高的准确率。

通过对表4数据的分析发现，随着频繁项集项数的增加，误检率相应减小，其中3、4项集的误检率较高，5、6项集的误检率较低。因此，从误检率方面考虑用频繁5、6项集作为规则库的效果较好。

综合以上两方面的的分析结果可以得出，用频繁5、6项集作为规则库时有较好的性能。但考虑到频繁-6项集是在频繁-5的基础上计算产生的，而对于现实网络中的大量数据来说，进行频繁项集挖掘的计算量是巨大的，用频繁-5项集作为规则库较为合适。

4 结 论

本文将ORAR算法用于网络入侵检测，在对训练数据集进行频繁3、4、5、6项集挖掘的基础上，将挖掘得到的规则应用于测试数据集进行入侵检测测试，针对各项集的检测准确率及误检率综合评价项集检测的性能，得知性能最好的检测项集为频繁5-项集。仿真结果也表明，基于关联规则进行网络入侵检测具有较高的检测效果。

参 考 文 献

[1] ANDERSON J P. Computer security threat monitoring and surveillance[R]. Fort Washington, P A: Jamep Anderson Co., 1980.

[2] DENNING D. An intrusion detection model[J]. IEEE Transactions on Software Engineering, 1987, 2: 222.

[3] ROESCH M. Snort: lightweight intrusion detection for networks[C]//USENIX Lisa ’99. [S. l.]: [s.n.], 1999: 229-238.

[4] AGRAWAL R, IMICLINSKI T, SWAM I A. Database mining: a performance perspective[J]. IEEE Trans Knowledge and Data Enginnering, 1993, 5: 9142925.

[5] AGRAWAL R, SRIKANT R. Fast algorithm for mining association rules[C]//Proceeding 1994 International conference Very Large Data Bases. Santiago, Chile: The ACM SIGMOD Anthology, 1994.

[6] MANNILA H, TOVIVONEN H, VERKAMO A I. Efficient algorithm for discovering association rules[C]//Proceedings AAAI ’94 Workshop Knowledge Discovery in Databases. Seattle WA: The ACM SIGMOD Anthology, 1994.

[7] AGRAWAL R, SHAFER J C. Parallel mining of association rules: design, implementation, and experience[J]. IEEE Trans Knowledge and Data Engineering, 1996, 8: 9622969.

[8] HAN Eui-hong, GEORGE K, KUMAR V. Scalable parallel data mining for association rules[C]//Proceeding of the ACM SIGMOD ’97. AZ USA: The ACM SIGMOD Anthology, 1997.

[9] 陈 莉, 焦李成. 基于关系代数的关联规则挖掘算法[J]. 西北大学学报(自然科学版), 2005, 35(6): 692-694.
CHEN Li, JIAO Li-cheng. Association rule mining algorithm based on relation algebra theory[J]. Journal of Northwest University (Natrural Science Edition), 2005, 35(6): 692-694.

[10] KDD CUP 99. KDD Cup 99 dataset[EB/OL]. [2009-08-20]. http://kdd.ics. uci. edu/databases/kddcup99/kddcup99. html.

编 辑 蒋 晓