第31卷 第20期 计 算 机 工 程 2005年10月
Vol.31 № 20 Computer Engineering October 2005
·软件技术与数据库· 文章编号：1000—3428(2005)20—0090—03 文献标识码：A 中图分类号：TP311

基于粗糙集理论的关联规则挖掘研究及应用

王旭仁[1,2]，许榕生[1]
(1. 中科院高能物理所计算中心，北京 100039；2. 首都师范大学信息工程学院，北京 100037)

摘 要：提出了一种基于粗糙集理论的关联规则算法，使用粗糙集理论对数据进行预处理，同时使用属性限制避免挖掘无用的关联规则，挖掘出来的关联规则是分类规则，可以对未知数据进行分类；使用规则过滤去除冗余规则，只保留本质的、一般的规则。通过对网络安全审计数据的分析的试验表明，该方法是行之有效的。

关键词：数据挖掘；粗糙集理论；关联规则；入侵检测

Research and Application of Association Rule Mining Based on Rough Set Theory

WANG Xuren[1,2], XU Rongsheng[1]
(1.Computer Center, Institute of High Energy Physics, CAS, Beijing 100039;
2. Information Engineering College, Capital Normal University, Beijing 100037)

【Abstract】 This paper proposes an association rule-mining algorithm based on rough set theory, pre-processing of data is done with rough set theory. At the same time attribute restraints have been applied to association rule mining for fear that useless rules are produced. The rules are classification rules that can be used to classify data whose class is unknown. Rule filtering is used to delete redundant rule and only the most general, and essential rules are kept. Tests in intrusion detection show that the algorithm is efficient and applicable.

【Key words】 Data mining; Rough set theory; Association rule; Intrusion detection

关联分析（Association analysis）是指在数据记录的数据项之间挖掘关联关系，某些数据项的出现预示着该记录中其它一些数据项出现的可能。Agrawal 等在 1993 年设计了 Apriori 算法[1]，首先提出了挖掘顾客交易数据库中项集之间的关联规则问题。近年来发展了很多挖掘算法[2~4]，关联规则除了应用在购物篮分析外，还应用到入侵检测模型的建立、建立分类器、电信数据分析、人口普查数据分析等方面。

20 世纪 80 年代初，波兰的 Pawlak 针对 G.Frege 的边界线区域思想提出了粗糙集（Rough Set），知识约简、离散化问题和不完全知识的补齐是粗糙集理论的主要研究内容，这些问题的研究在一定程度上（甚至是很好地）解决了传统数据挖掘中存在的超大数据、噪音数据、空值、不完整数据和冗余数据问题。

本文的主要工作就是将以上两种技术进行结合，提高关联规则挖掘的效率和实用性，并在入侵检测系统中进行应用。

1 相关工作背景

1.1 粗糙集理论

粗糙集理论假定知识是一种对对象进行分类的能力。而知识必须与具体或抽象世界的特定部分相关的各种分类模式联系在一起，这种特定部分称之为所讨论的全域或论域(universe)。

定义 1 T={U，A，V，f}是一个决策系统，其中 U 是一个非空集合，表示数据库中的所有记录（Record）；$V=\bigcup_{a\in A}V_a$ 是属性值组成的集合；A=C∪D，C, D 是 A 的两个属性子集，分别称为条件属性和决策属性且 C∩D=Ø。

(1)离散化问题

粗糙集理论是基于集合论的基础上提出来的，而实际处理的数据中连续属性很常见，因此连续属性的离散化问题是粗糙集理论的主要研究内容之一。离散化也是降低超大数据量的有效方法之一。

(2)不完整数据问题

从不完整的数据集中学习规则比在完整的数据集中学习规则困难。一种简单的解决途径是把存在空缺（遗漏）属性值的实例记录删除，从而得到一个完备的信息表。但是当存在遗漏信息的实例相对较多时，这种方法就会严重影响信息表中的信息量。

1.2 关联规则

设 $I=\{i_1,i_2,...,i_m\}$ 是项的集合。包含 k 个项的项集称作 k-项集（k-itemsets）。设 D 是数据库记录的集合，其中每个事务 T 是项的集合，使得 $T\subseteq I$。设 A 是一个项集，事务 T 包含 A 当且仅当 $A\subseteq T$。

定义 2 关联规则是形如 $A\Rightarrow B$ 的蕴涵式，其中 $A\subset I$, $B\subset I$, $A\cap B=\varnothing$。A 称为规则的左部或规则的前提（简记 LHS），B 称为规则的右部或规则的结论（简记 RHS）。

对关联规则 $A\Rightarrow B$ 的度量标准很多：支持度（Support），置信度（Confidence），关联度（correlation），定义如下：

$$\text{support}(A\Rightarrow B)=P(A\cup B)$$

$$\text{confidence}(A\Rightarrow B)=P(B\mid A)$$

基金项目：国家“973”计划基金资助项目(G1999035806)
作者简介：王旭仁(1972—)，女，讲师、博士，主研方向：信息网络安全，数据挖掘；许榕生，研究员、博导
收稿日期：2004-09-07 **E-mail**：wangxr@mail.ihep.ac.cn

$$correlation(A \Rightarrow B) = P(B \mid A) / P(B)$$

其中 P(A)是指 A 在数据集 D 中出现的概率，其余雷同。$support(A \Rightarrow B)$ 指 A、B 在 D 中同时出现的概率；$confidence$ $(A \Rightarrow B)$ 表示在 A 出现的前提下，B 出现的条件概率；$correlation(A \Rightarrow B)$ 表示 A、B 同时出现的相关性，如果值大于 1，则 A、B 是正相关的，意味着前提的出现蕴含着结论的出现，如果值等于 1，则 A、B 是独立的，它们之间没有相关性，如果值小于 1，则 A、B 的出现是负相关的，这样的关联规则即使置信度或支持度的值很高，也可能会产生误导。

除了上述度量定义外，对关联规则的度量还有很多[5]，如收益(gain)、conviction、方差(variance)、基尼平均差(gini)等。文献[5]中提出了基于规则支持度和置信度的半序定义 SC-Optimality，并证明，对于给定的最小置信度和最小支持度，在 SC-Optimality 下最优的规则包含了在其他任何度量意义下所有最好的规则。这样对关联规则的度量优化进行了形式化的描述，也把关量规则的各种度量进行了统一的讨论，说明 Agrawal 提出的置信度和支持度仍旧是关联规则最基本最本质的度量。

2 基于粗糙集理论的关联规则挖掘研究

本文在 Apriori 的算法的基础上进行了改进，主要工作有：(1)提出了基于粗糙集理论的关联规则挖掘算法，使用粗糙集理论对数据进行预处理，包括不完全数据的补齐、数值数据的离散化处理。(2)针对在关联规则挖掘中，经常会产生很多无用的规则，提出了属性限制对关联规则中出现的属性进行限制，使得挖掘出来的规则是分类规则，就可以对未知数据进行分类；(3) 针对关联规则挖掘总是会挖掘出大量规则，大量的规则集不好使用，也不容易理解。文中对关联规则采取了过滤措施，去掉重复、冗余、没有意义的规则。只保留一般的、本质的关联规则。

2.1 利用粗糙集理论对数据进行预处理

首先使用粗糙集理论对条件属性数据的进行预处理，即进行数据补齐和数值属性的离散化。由于在实验数据中不完整数据所占比例很小，因此对不完整的数据记录直接删除。在选择离散化算法时，要考虑离散化后的数据保持原有的不可分辨关系，在本文中对离散数据采用了 Semi Naïve Scaler 离散化算法。

2.2 属性限制

在关联规则的实际应用中，有时挖掘出来的规则中包含一些并不希望出现的属性。由于待处理的数据是一个二维关系表，很容易转换成定义 1 所描述的决策系统，属性受限的关联规则的定义和粗糙集理论中决策规则的定义是完全一致的[6]：在规则的结论中只允许出现分类（决策）属性及其对应值，其他属性（条件属性集）只允许出现在规则的前提中。这样得到的关联规则就是一个分类的规则，同时避免无用规则的出现。

2.3 关联规则的过滤

在关联规则的算法中，通过限定规则结论中出现的属性，从而减少了产生无用的关联规则，通过支持度或置信度阈值的限制，也可以减少一部分规则，但是这样得到的关联规则数目还是十分的庞大，里面包含了很多冗余规则、重复规则、矛盾的规则，针对这些情况，加以过滤处理：

(1)$\beta \rightarrow \gamma_1$ 和 $\beta \rightarrow \gamma_2$ 是前提重复的两条规则，但是结论却不一致，这时可以把置信度低的规则去掉，避免重复和矛盾的规则；

(2)规则 $\beta_1 \rightarrow \gamma$ 和规则 $\beta_2 \rightarrow \gamma$ 且有 $\beta_1 \subset \beta_2$，则 $\beta_2 \rightarrow \gamma$ 是冗余规则，因为规则 $\beta_1 \rightarrow \gamma$ 更具有概括性，更简单，把规则 $\beta_2 \rightarrow \gamma$ 去除掉；

(3)Chi^2 是统计学中一种用来度量独立性或相关性的常用方法，被用来对关联规则进行过滤[8]；同时根据定义，可以计算关联规则 $\beta \rightarrow \gamma$ 的相关程度。只有同时满足条件 Chi^2 $(\beta \rightarrow \gamma)$≥指定阈值并且 $correlation$ $(\beta \rightarrow \gamma)$>1 的规则予以保留；否则删除。

3 试验和讨论

传统的网络入侵检测技术的局限性越来越明显，使用传统手工分析和编码的方式已经不能适应网络新攻击层出不穷和数据量日益增大的趋势，也不便于分布式分析和协同工作。将数据挖掘的方法应用到入侵检测系统中来，是近年来的一个研究热点[9]。下面的实验就是对网络安全审计数据进行的。

3.1 数据源

在下面实验中使用的数据来自 KDD CUP 99 数据集。训练数据是从 KDD CUP99 的 10%的训练数据中随机抽取的，共 49 402 条记录；测试数据使用的是 KDD CUP99 的 10%的测试数据集，共 311 030 条数据记录。训练数据和测试数据的攻击类型分布如表 1 所示。

表 1 试验数据分布

	训练数据		测试数据	
Type	Count	percent	Count	percent
Normal	9 768	19.77	60 592	19.48
Probing	435	0.88	4 166	1.339
DoS	39 085	79.12	231 453	74.415
U2R	3	0.006	79	0.025
R2L	111	0.224	14 740	4.739

在训练数据中，除了正常连接外，共包含分属于 Probing（刺探攻击）、DoS（拒绝服务攻击）、U2R（非法获得根权限）、R2L（远程登录攻击）4 种攻击类型的 18 种具体攻击。在测试数据中，也包含分属于 4 种攻击类型的 37 种数据，在这 37 种攻击中，有 21 种攻击是在训练数据中没有出现的，以此检测推导的模式能否发现未知攻击，在这 21 种攻击中：属于 Probing 类型的攻击有 2 种，共 1 789 条连接；属于 DoS 的攻击有 5 种，共 6 564 条连接；属于 U2R 的攻击有 4 种，共 33 条连接；属于 R2L 的攻击有 10 种，共有 10 374 条连接。而在训练数据中有两种攻击在测试数据中没有出现，它们是 warezclient 和 spy，分别属于 DoS 攻击和 R2L 攻击。

3.2 试验结果

使用粗糙集理论对数据进行预处理，例如对属性 Duration（网络连接持续时间）进行离散化后得到的断点集为

{0.5, 57, 62, 65, 72.5, 328.5, 337.5, 5007, 5064, 5065.5, 5195, 10017.5, 10228, 12503, 12553.5, 13901.5, 14191.5, 21204.5, 22625, 28847.5}。

试验得到的关联规则形如

src_bytes = [1 031.5 , 1 033.5]
protocol_type = icmp ->class = DoS
(SupCount=22 838, conf=100.00%, sup=46.229%);

在上述规则中，规则左部是条件属性，右部是结论（决策）属性，表示当客户端发送的字节数在 1 031.5~1 033.5 之间并且协议是网间控制报文协议（Internet Control Message Protocal，ICMP）时，得到结论是该连接是 DoS 攻击。在规则后面的是对规则的度量，包含规则的纪录有 SupCount=22 838 条，规则的置信度为 conf=100%，规则的支持度为 sup=46.229%。

在不同的支持度阈值下的分类结果如表 2 所示。可以看出，如果 minsup 的值设置得过高，则一些支持度较低但是置

信度很高的规则可能没被包含进来，分类的效果将受到影响；规则对正常连接和 DoS 攻击的检测准确率较高，对 Probing 的检测准确率最高为 74.8%，这些结果和 KDD CUP 99 优胜者的分析结果[7]基本一致；在支持度阈值降到比训练数据中所有连接类型的分布比例还低的情况下，即 minsup=0.005%，对 U2R 攻击和 R2L 攻击的分类检测还不理想，说明 U2R 攻击和 R2L 攻击分布比例过低、收集的属性不充足对分类的效果造成影响。从表中可以看出，随着支持度阈值的降低，规则的数目也随之增加，但是规则的数目很少，是入侵检测系统可以接受的。

表 2 不同支持度阈值的实验结果比较

Value	Minsup=1%	Minsup=0.05%	Minsup=0.005%
Nomal	99.50%	99.47%	99.47%
Probing	50.77%	74.8%	74.8%
DoS	96.39%	96.6%	96.6%
U2R	0%	3.8%	2.53%
R2L	0.97%	1.21%	0.73%
全部错误率	8.16%	7.68%	7.7%
规则数	74	104	158

4 结论

数据预处理是数据挖掘中重要一步，本文提出了基于粗糙集理论的关联规则挖掘算法，使用粗糙集理论对数据进行预处理，然后再进行关联规则挖掘，提高了关联规则挖掘的效率；将粗糙集理论中决策系统和决策规则的概念应用到关联规则挖掘中，对规则进行属性限制；同时使用规则过滤的方法对大量关联规则进行减少，得到实际可用的关联规则集；最后进行了详细的试验，并对分类结果进行了比较，分类效果较好。

参考文献

1 Agrawal R, Imielinski T, Swami A. Mining Associations Between Sets of Items in Massive Databases. In: Proc. of the 1993 ACM-SIGMOD Int'l Conf. on Management of Data, 1993:207-216

2 Agrawal R, Shafer J. Parallel Mining of Association Rules. IEEE Transaction on Knowledge and Data Engineering,1996, 8(6)

3 Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Database. In: Proceedings of 20th International Conference on Very large Data Bases ,1994-09:478-499

4 Srikant R, Agrawal R. Mining Generalized Association Rules.In: Proceedings of 20th International Conference on Very large Data Bases, 1995-09:407-419

5 Bayardo R J, Agrawal R. Mining the Most Interesting Rules. In: Proc. of the Fifth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 1999: 145-154

6 王国胤. Rough 集理论与知识获取. 西安: 西安交通大学出版社, 2003

7 Results of the KDD'99 Classifier Learning Contest. http://www.cs.ucsd.edu/users/elkan/clresults.html, 1999

8 Borgelt C, Kruse R. Induction of Association Rules: Apriori Implementation. In: 15th Conference on Computational Statistics (Compstat 2002), Heidelberg, Germany, Physica Verlag, 2002

9 王旭仁,毕学尧,许榕生. 对IDS审计数据的关联分析. 计算机工程, 2004,30(6):34-35

(上接第 2 页)

和它的环境通信。另一方面 W 还负责将包转化成区域特定的格式。区域在 NOC 中的作用主要体现在以下 4 个方面：

(1)区域专指一个资源集合或是一个网络片断来执行特殊的任务。如处理流数据区域，处理块数据区域。

(2)每个区域可以自定义自己的通信机制。

(3)区域可将本区域的资源和其它不属于本区域的资源隔离开。

(4)区域将一个特定的技术封装进 NOC 里。

区域的尺寸可以随需求定制，但是它们的边界必须是凸出的，与其他组件隔绝。由于区域的封装常常会给通信能力以及通信响应速度带来不利影响，因此在设计中常常将具有高通信需求的资源配置在同一个区域中。从网络层的角度来看，区域和封装不是将整个网络分割成一个个子网，而是以一种更有效、更合理的方式来组织网络通信。

4 结论

半导体以及集成电路技术的飞速发展，在使得片上系统功能越来越完善的同时也带来了许多挑战性问题。原有的设计模式和方法在面对设计复杂性、功能可靠性、能源有效性等问题时显得力不从心。NOC 基于网络设计的技术来分析和设计已有的 SoCs，在不可靠的信号传输和线路延迟的限制下，实现以最小能量消耗充分满足服务质量需求，并提供高性能和高可靠性保证。可以想象，片上网络（NOC）技术的发展将引发微电子领域的又一次革命。

参考文献

1 Kumar S, Jantsch A, Soininen J P, et al. A Network on Chip Architecture and Design Methodology[J]. In:Proceedings of the IEEE Computer Society Annual Symposium, 2002

2 Benini L, Micheli D G. Networks on Chip: A New Paradigm for Systems on Chip Design. In: Proceedings of the 2002 Design, Automation and Test in Europe Conference and Exhibition, 2002

3 Siegmund R, Miiller D. Efficient Modeling and Synthesis of On-chip Communication Protocols for Network-on-Chip Design [J]. In: Proceedings of the IEEE Computer Society, 2003

4 Lei T, Kumar S. Algorithms and Tools for Network on Chip Based System Design. In: Proceedings of the 16th Symposium on Integrated Circuits and Systems Design, 2003

5 Pinto A, Carkoni L P, Sangiovanni-Vincentelli A L. Efficent Synthesis of Networks on Chip[J]. In: Proceedings of the 21st International Conference on Computer Design, 2003

6 Soininen J P, Jantsch A. Extending Platform-based Design to Network on Chip Systems, In: Proceedings of the 16th International Conference on VLSI Design, 2003

7 Simunic T, Boyd S P, Glynn P. Managing Power Consumption in Networks on Chip. IEEE Transactions on Very Large Scale Integration(VLSI) Systems, 2004,12(1)

8 Marculescu R. Networks-on-Chip: The Quest for On-chip Fault-tolerant Communication[J]. In: Proceedings of the IEEE Computer Society Annual Symposium on VLSI, 2003

9 Zeferino C A, Kreutz M E, Susin A A. RASoC: A Router Soft-core for Networks-on-Chip[J]. In: Proceedings of the Design, Automation and Test in Europe Conference and Exhibition Designers'Forum, 2004

10 Zeferino C A, Susin A A. SoCIN: A Parametric and Scalable Network-on-Chip. In: Proceedings of the 16th Symposium on Integrated Circuits and Systems Design, 2003