

文章编号:1001-9081(2005)07-1595-03

## 基于 VPRS 的信息系统风险分析

肖 龙<sup>1</sup>,戴宗坤<sup>1</sup>,杨 炜<sup>2</sup>

(1. 四川大学 信息安全研究所,四川 成都 610064; 2. 成都市 456 信箱,四川 成都 610061)

(wwop2002@163.com)

**摘 要:**提出了基于可变精度粗糙集(VPRS)的风险分析方法。利用 VPRS 模型中的  $\beta$  约简有效地克服了传统粗糙集处理噪声数据的不足;同时提出了通过定性和定量相结合的方法过滤风险规则,不但约简了风险规则的数量,而且提高了风险规则的有效性。最后,运用 VPRS 模型对文献[4]中的风险评估数据进行分析,有效地挖掘出了隐含在其中的风险规则。

**关键词:**信息系统;风险评估;粗糙集;VPRS

**中图分类号:** TP309.2 **文献标识码:** A

## Risk analysis of information system based on variable precision rough set

XIAO Long<sup>1</sup>, DAI Zong-kun<sup>1</sup>, YANG Wei<sup>2</sup>

(1. Institute of Information Security, Sichuan University, Chengdu Sichuan 610064, China;

2. 456 Mail Box of Chengdu, Chengdu Sichuan 610061, China)

**Abstract:** The risk analysis method based on variable precision rough set (VPRS) was proposed. It used reduct functions and filtered risk regulations and combined quantitative measurement with qualitative analysis. So it not only reduced greatly the data, but also increased the validity of risk regulations. At last, we analyzed the data in the reference [4] and mined the risk regulations effectively among them by using this method.

**Key words:** information system; risk evaluation; rough set; VPRS (Variable Precision Rough Set)

### 0 引言

在当今大规模的开放系统互联的网络环境下,无论采用多么完善的安全保护措施,风险总是存在的。因而适当的方法是在整个网络通信过程中应用智能化的方法进行风险管理,而周密的信息安全风险分析是可靠、有效的风险管理的必要前提,因此对信息系统进行风险分析和评估已越来越受到人们的重视。而对于像信息系统这样的复杂巨系统存在着大量的风险评估数据,如何从中找出隐含的内在规律是本文研究的重点。

粗糙集(rough set)的一个重要特点是具有很强的定性分析能力,即不需要预先给定某些特征或属性的数量描述便可直接从给定问题的描述集合出发,通过不可分辨关系确定给定问题的近似域,从而找出该问题的内在规律<sup>[1]</sup>。但在信息系统的风险分析中,大量的定量数据存在着分散性、噪声干扰、不完整等各种因素,单单基于 Pawlak 粗糙集理论难以准确地刻画规律的重要性,利用可变精度粗糙集(VPRS)模型却有效地改进 Pawlak 模型的不足,挖掘出隐含在数据背后的内在规律。

### 1 Pawlak 的粗糙集模型

根据 Pawlak 的粗糙集模型<sup>[2]</sup>,有如下定义和描述。

**定义 1** 决策系统(Decision System, DS)

称  $S = (U, A, V, f)$  为决策系统,其中:  $U$  为非空有限集,称为论域;  $A$  为有限个属性的集合;  $V = \bigcup_{a \in A} V_a$ ,  $V_a$  为属性  $a \in A$

的值域;  $f: U \rightarrow V_a$  为一单射,使论域  $U$  中任一元素取属性  $a$  在  $V_a$  中的某一唯一值。

一般情况下,属性  $A$  由条件属性集合  $C$  和决策属性集合  $D$  组成,即  $C \cup D = A$  且  $C \cap D = \emptyset$ 。

**定义 2** DS 的不可分辨关系(Indiscernibility Relation)

对于  $S = (U, A, V, f)$  的决策系统,给定属性集合  $P \subseteq A$ ,称二元关系  $R(P) = \{(x_i, x_j) \in U \mid \forall a \in P, f(x_i, a) = f(x_j, a)\}$  为  $S$  的不可分辨关系。

显然,不可分辨关系是一个等价关系。通过一个不可分辨关系可以得到决策系统的划分,形成  $n$  个等价类,用  $U/R(P)$  表示。在  $S$  中,若  $A = C \cup D$ ,则  $C$  和  $D$  就可分别构成两个等价关系  $R(C)$  和  $R(D)$ ,分别地把  $U$  划分成  $U/R(C)$  和  $U/R(D)$ 。对于  $x_i \in U$ ,它确定了对于等价关系  $R(C)$  的一个等价类,用  $[x_i]_R$  来表示满足:

$$[x_i]_R = \{x_j \in U \mid \forall a \in C, f(x_i, a) = f(x_j, a)\}$$

**定义 3** 粗糙隶属度函数(Rough Membership Function)

设  $Y$  是  $U/R(D)$  中的一个等价类,对于  $[x_i]_R$  的粗糙隶属度函数  $\text{Pr}(Y \mid [x_i]_R)$  定义为:

$$\text{Pr}(Y \mid [x_i]_R) = \frac{\text{card}(Y \cap [x_i]_R)}{\text{card}[x_i]_R}$$

其中  $\text{card}(\cdot)$  表示集合的基数。

粗糙隶属度函数  $\text{Pr}(Y \mid [x_i]_R)$  解释了  $Y$  和  $[x_i]_R$  之间的相互依赖程度。

**定义 4** 上近似和下近似(Lower Approximation and Upper

收稿日期:2005-01-28;修订日期:2005-04-07 基金项目:国家 863 计划项目(2001AA142171)

作者简介:肖龙(1977-),男,重庆人,博士研究生,主要研究方向:信息安全、模式识别、人工智能;戴宗坤(1945-),男,重庆人,研究员,主要研究方向:开放系统互连的安全体系和安全工程方法;杨炜(1972-),男,四川彭州人,工程师,主要研究方向:信息安全。

Approximation)

称  $R(C)_Y$  为  $Y$  的  $R(C)$  下近似, 称  $R(C)^-Y$  为  $Y$  的  $R(C)$  上近似。定义如下:

$$R(C)_Y = \bigcup_{Pr(Y|X_i)=1} \{X_i \in U/R(C)\}$$

$$R(C)^-Y = \bigcup_{Pr(Y|X_i)>0} \{X_i \in U/R(C)\}$$

定义 5 正域、负域和边界域 (Positive Region, Negative Region and Borderline Region)

称  $POS_c(Y) = R(C)_Y$  为  $Y$  的  $R(C)$  正域, 称  $NEG_c(Y) = U - POS_c(Y)$  为  $Y$  的  $R(C)$  负域, 称  $BND_c(Y) = R(C)^-Y - R(C)_Y$  为  $Y$  的  $R(C)$  边界域。

由此通过集合  $Y$  的上近似和下近似将论域  $U$  划分成三个不相交的区域。直观地说,  $POS_c(Y)$  是对于等价关系  $R(C)$  能确定地划入  $Y$  的元素集合,  $NEG_c(Y)$  是对于等价关系  $R(C)$  肯定不能划入  $Y$  的元素的集合,  $BND_c(Y)$  是对于等价关系  $R(C)$  不能确定是否能划入  $Y$  的元素集合。若  $BND_c(Y) \neq \emptyset$ , 也就是说  $Y$  不可以由对于等价关系  $R(C)$  相对应的等价类精确定义, 则称  $Y$  为  $R(C)$  粗糙集, 若  $BND_c(Y) = \emptyset$ , 则称  $Y$  为  $R(C)$  精确集。

## 2 可变精度粗糙集模型 (VPRS)

Pawlak 粗糙集模型的核心问题是分类分析, 而且这种分析必须是确定的。但在实际应用中数据包含噪音是在所难免的。为了增强粗糙集模型的抗干扰能力, 文献[3]提出了可变精度粗糙集模型 (Variable Precision Rough Set, VPRS)。

在 VPRS 中允许一定的误分类率  $\beta \in [0, 0.5]$ 。与 Pawlak 粗糙集模型相类似, VPRS 中也定义了  $Y$  的  $R(C)$   $\beta$  正域  $POS_c^\beta(Y)$ ,  $Y$  的  $R(C)$   $\beta$  负域  $NEG_c^\beta(Y)$  以及  $Y$  的  $R(C)$   $\beta$  边界域  $BND_c^\beta(Y)$ , 定义如下:

$$POS_c^\beta(Y) = \bigcup_{Pr(Y|X_i) \geq 1-\beta} \{X_i \in U/R(C)\}$$

$$NEG_c^\beta(Y) = \bigcup_{Pr(Y|X_i) \leq \beta} \{X_i \in U/R(C)\}$$

$$BND_c^\beta(Y) = \bigcup_{\beta \leq Pr(Y|X_i) < 1-\beta} \{X_i \in U/R(C)\}$$

$POS_c^\beta(Y)$  表示根据等价关系  $R(C)$  将  $U$  中元素误分类到  $Y$  中的概率不超过  $\beta$  的等价类的集合,  $NEG_c^\beta(Y)$  表示根据等价关系  $R(C)$  将  $U$  中元素误分类到  $Y$  中的概率超过  $1-\beta$  的等价类的集合,  $BND_c^\beta(Y)$  则表示两者之差。

在 VPRS 中还引进了属性的  $\beta$  约简。  $\forall a \in C$ , 如果  $POS_c^\beta(Y) = POS_{C-\{a\}}^\beta(Y)$ , 则认为  $a$  是冗余属性, 称  $C' = C - \{a\}$  为  $C$  的一个  $\beta$  约简。显然当  $\beta = 0$  时, VPRS 模型和 Pawlak 粗糙集模型是一致的。

## 3 基于 VPRS 的信息系统风险分析

信息系统的风险是不确定性事件发生的概率及其造成的可能损失的概率函数。若已知不确定事件发生概率  $P$  及其损失函数  $C$ , 则关于风险  $x$  的度量为:  $R(x) = R(C, P)$ 。

在信息系统中, 不确定事件发生的概率与系统及其资源的脆弱性  $V$  (或漏洞), 以及针对信息系统及其资源的威胁性  $T$  有关。信息系统的漏洞或缺陷主要来自于硬件、软件、网络和通信协议等方面; 而针对系统的威胁则分为主动或被动两类, 主要包括对通信或网络资源的破坏, 对信息的滥用、讹用或篡改, 信息或网络资源的被窃、删除或丢失, 信息的泄漏, 服务的中断和禁止等。

对于不确定事件发生的后果包括对资产的直接和间接影响  $F_1$ , 能力影响  $F_2$ , 系统恢复费用  $E$  三方面。对资产影响包括环境恶化、数据泄漏、通信被干扰和信息丢失等; 对能力的影响主要有中断、延迟和削弱所带来的后果; 系统恢复费用包括信息恢复费与服务恢复费<sup>[1]</sup>。

根据上述分析, 构建层次结构如图 1 和图 2。

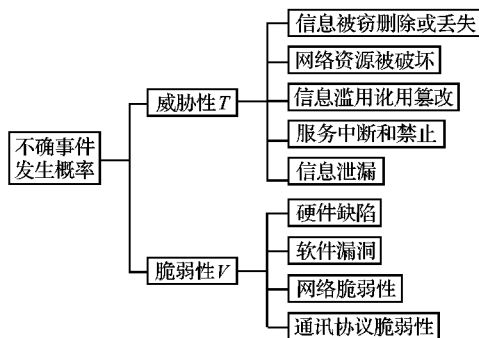


图1 不确定事件发生概率层次结构

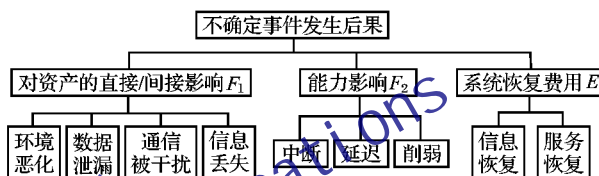


图2 不确定事件发生后后果层次结构

在风险量化过程中, 由于上述诸多因素存在着不确定性, 使得很难通过有效的数据累计确定风险中不确定事件发生的概率, 也很难准确地判断其发生后的严重程度。为了较为准确地描述各风险因素对系统风险的影响强度及范围, 运用模糊综合评判 (Fuzzy Comprehensive Evaluation, FCE) 的方法<sup>[4]</sup>得到系统风险量化值  $R(x)$ , 威胁性  $T$  和资源的脆弱性  $V$  以及发生后造成的影响  $F_1, f_2$  和系统恢复费用  $E$ 。

利用可变精度粗糙模型 (VPRS) 对所得到的风险评估数据进行挖掘, 提取系统风险规则, 主要步骤如下:

步骤1 将信息系统的风险评估表示为一决策系统  $DS$ , 其中  $A = \{T, V, f_1, f_2, E, R(x)\}$ ,  $V_a = \{\text{通过 FCE 得到的属性值}\}$ ,  $U = \{\text{系统风险}\}$ ,  $f: U \rightarrow V_a$  为一单射,  $C = \{T, V, f_1, f_2, E\}$ ,  $D = \{R(x)\}$ ,  $C \cup D = A$ , 从而形成二维属性表。

步骤2 对二维属性表进行预处理, 包括离散化数据值, 删除重复元组, 以均值填补缺省值。为了离散化, 根据数据值将威胁性、脆弱性以及发生后的影响和系统恢复费用分为  $(0, 0.3], (0.3, 0.6], (0.6, 1]$  三类, 将风险划分为 {高, 中, 低} 三级。

步骤3 给定误分类率  $\beta$  值, 对预处理后的决策表进行属性的  $\beta$  约简。

步骤4 在  $\beta$  约简的基础上过滤并选取风险规则。对  $\bigwedge_{a_i \in RED} a_i = \gamma_i \rightarrow \bigwedge_{b_i \in D} b_i = \tau_i$  这一条风险规则 (简记为  $\alpha \rightarrow \beta$ ), 依据支持数  $support(\alpha \rightarrow \beta) = support(\alpha \cdot \beta)$ , 可信度  $accuracy(\alpha \rightarrow \beta) = \frac{support(\alpha \cdot \beta)}{support(\alpha)}$ , 覆盖率  $coverage(\alpha \rightarrow \beta) = \frac{support(\alpha \cdot \beta)}{support(\beta)}$  来产生过滤规则<sup>[5]</sup>。其中:

$$support(\alpha) = card(\{x_i \in U \mid \bigwedge_{a_i \in RED} f(x_i, a_i) = \gamma_i\})$$

$$support(\beta) = card(\{x_i \in U \mid \bigwedge_{b_i \in D} f(x_i, b_i) = \tau_i\})$$

$$support(\alpha \cdot \beta) = card(\{x_i \in U \mid (\bigwedge_{a_i \in RED} f(x_i, a_i) = \gamma_i) \wedge$$

$$(\bigwedge_{b_i \in D} f(x_i, b_i) = \tau_i) \}$$

$\wedge$  表示并算子,  $RED$  为 VPRS 中的  $\beta$  约简。直观地说, 支持数表示在整个论域  $U$  中支持该规则的元素个数, 可信度表示运用该风险规则进行推理正确的概率, 覆盖率表示该规则的支持数在相应的决策类中的比重。在规则过滤中一般选取可信度和覆盖率都相对较高的有效规则。同时结合实际情况和专家经验选取风险规则。

#### 4 应用实例

利用文献[4]中建立的专家数据库, 提取100例系统风险评估数据, 离散化处理后如表1所示。

表1 风险评估数据

风险事件	威胁性 $T$	脆弱性 $V$	影响 $F_1$	影响 $F_2$	恢复费用 $E$	风险等级
1	(0, 0.3]	(0, 0.3]	(0, 0.3]	(0.3, 0.6]	(0, 0.3]	低
2	(0.3, 0.6]	(0.3, 0.6]	(0.3, 0.6]	(0.6, 1]	(0.3, 0.6]	中
3	(0.6, 1]	(0.3, 0.6]	(0.6, 1]	(0, 0.3]	(0.3, 0.6]	高
4	(0, 0.3]	(0.3, 0.6]	(0.3, 0.6]	(0, 0.3]	(0, 0.3]	低
5	(0.6, 1]	(0, 0.3]	(0.6, 1]	(0.3, 0.6]	(0, 0.3]	中
6	(0.6, 1]	(0.6, 1]	(0, 0.3]	(0.6, 1]	(0.3, 0.6]	高
7	(0.6, 1]	(0, 0.3]	(0, 0.3]	(0.3, 0.6]	(0.6, 1]	高
8	(0.6, 1]	(0, 0.3]	(0.6, 1]	(0.3, 0.6]	(0, 0.3]	高
9	(0.6, 1]	(0.6, 1]	(0, 0.3]	(0.6, 1]	(0.3, 0.6]	中
10	(0.6, 1]	(0.3, 0.6]	(0.3, 0.6]	(0.6, 1]	(0, 0.3]	中
...	...	...	...	...	...	...

采用 VPRS 模型, 选取误分类  $\beta = 0.3$  对表1数据进行  $\beta$  约简。在此基础上定义风险过滤规则:  $accuracy(\alpha \rightarrow \beta) \geq 0.75$ ,  $coverage(\alpha \rightarrow \beta) \geq 0.2$ ,  $support(\alpha \rightarrow \beta) \geq 3$ , 对提取出来的风险规则再进一步通过专家筛选, 过滤出15条强有效的风险规则, 如表2所示。

表2 风险规则

规则	威胁性 $T$	脆弱性 $V$	影响 $F_1$	影响 $F_2$	恢复费用 $E$	风险等级
1	—	—	(0, 0.3]	(0, 0.3]	(0, 0.3)	低
2	—	—	(0.3, 0.6]	(0.6, 1]	(0, 0.3)	中
3	(0.6, 1]	(0.3, 0.6]	(0.6, 1]	—	—	高
...	...	...	...	...	...	...

表2中每一行表示一条风险规则, “—”表示运用此风险规则时无需考虑此属性。例如, 规则1表示为: 当风险事件发生后对系统的影响及其恢复费用都较低时, 无论威胁性和脆弱性有多大, 系统所受到的风险等级低。

#### 5 结语

通过运用 VPRS 模型对信息系统风险评估数据的挖掘得到了较为有效的风险评估规则, 借助这些风险规则不但可以帮助一般的管理人员做出专家级的风险决策, 而且也大大简化了进一步数据分析的数量。但在实际运用中, 对于风险规则的过滤和提取还需反复做一些调整, 以期得到最符合实际最有效的风险规则。例如 VPRS 中的  $\beta$  值取多少最好, 可信度、覆盖率和支持数怎么选取等。

#### 参考文献:

- [1] 刘浩. Rough 集及 Rough 推理[M]. 北京: 科学出版社, 2001.
- [2] PAWLAK Z. Rough sets[J]. International journal of information and computer science, 1982, 11(5): 314–316.
- [3] ZIARKO W. Variable precision rough set model[J]. Journal of computer and system sciences, 1993, 46(1): 39–59.
- [4] 肖龙, 戴宗坤. 信息系统风险的多级模糊综合评判模型[J]. 四川大学学报(工程科学版), 2004, 36(3): 98–102.
- [5] QHRN A. Discernibility and rough sets in medicine: tools and application[D]. Norwegian University of science and Technology, 1999.

(上接第1591页)

XPath 查询可直接在 URL 中指定, 也可在 URL 指定的模板中指定, 参数可传递到直接在 URL 指定的 XPath 查询, 也可传递到使用 XPath 变量的模板中指定的 XPath 查询。例如下面是查询“地址”的模板文件(a.xml):

```
<?xml version="1.0" encoding="GB2312" ?>
<ROOT xmlns:sql="urn:schemas-microsoft-com:xml-sql">
  <sql: xpath-query mapping-schema="GeoMetadataXDR.xml">
    /地址
  </sql: xpath-query>
</ROOT>
```



图4 执行查询模板的结果集

#### 5 结语

基于 XDR 纲要的空间元数据存储策略较完整地保存了 XML 空间元数据文档的信息, 它不仅利用了关系数据库数据存储完整性、存取的高效性、强大的安全机制、并发访问控制机制等优点, 而且在查询方面既可以使用 SQL 语句进行内容查询, 也可以使用 XPath 进行结构查询, 为促进信息的交换与共享提供了一条新思路。

#### 参考文献:

- [1] BOURRET R.. XML Database Products middleware[EB/OL]. <http://www.rpbouret.com/xml/XMLDatabaseProds.htm#middleware>, 2004–10.
- [2] FLORESCU D, KOSSMANN D. Storing and Querying XML Data using an RDBMS[J]. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering Special Issue on XML, 1999, 22(3): 29–34.
- [3] 邓芳. XML 文档到数据库数据转换研究[J]. 北京邮电大学学报, 2004, 27(1): 84–88.
- [4] 沈兆阳, 李劲. SQL Server 2000 与 XML 整合应用[M]. 北京: 清华大学出版社, 2001.
- [5] 地理信息共享领域元数据专用标准[EB/OL]. <http://nfgis.nsdi.gov.cn/>, 2005–02.