基于粗糙集的网络安全评估规则提取

刘晓玲[1] 谢仙斌[2] 张家录[3] 刘灵丽[1]

[1](湘南学院计算机科学系 郴州 423000)
[2](高斯贝尔数码科技股份有限公司 郴州 423000)
[3](湘南学院数学系 郴州 423000)

摘 要 给出了基于粗糙集的网络安全评估指标的约简方法。提出了一种利用粗糙集理论提取网络安全评估规则，进而利用评估规则构建网络安全评估决策系统的算法。利用一个简化的网络安全评估数据集，验证了本文提出的决策规则提取方法是有效的。

关键词 网络安全；模糊粗糙集；属性约简；评估规则

Extraction of Evaluation Rules of Network Security Based on Rough Set

LIU Xiaoling[1] XIE Xianbin[2] ZHANG Jialu[3] LIU Lingli[1]

[1](*Department of Computer Science*, *Xiangnan University*, *Chenzhou* 423000, *China*)
[2](*Gospell Digital Technology Co*., *Ltd*., *Chenzhou* 423000, *China*)
[3](*Department of Mathematics*, *Xiangnan University*, *Chenzhou* 423000, *China*)

Abstract We present a network security evaluation index reduction method based on rough set theory. An algorithm with rough set theory to mine the rules of computer network security evaluation is proposed. The evaluation rules mining method presented in this paper is validated with a simplified network security evaluation data set.

Keywords network security; fuzzy rough set; attribute reduction; evaluation rule

1 引 言

随着信息技术的快速发展，人们对网络的依赖程度越来越高，因此，网络的安全问题也变得越来越重要，许多网络系统都由于安全问题而造成大量的经济损失。对网络系统的安全性进行先期评估是防范网络出现安全问题的一种有效手段。市场上现有许多的网络安全评估产品，对于加强网络系统的安全性起到了一定的作用，但是这些网络安全评估产品大都局限于对网络系统存在安全漏洞的探测与分析方面，缺乏对网络安全评估数据的深层分析，难以形成对计算机网络系统安全评估规则的整体描述。

粗糙集理论作为一种刻画不完整性和不确定性的数学工具，能有效地分析不精确、不一致、不完整等各种不完备的信息，还可以对数据进行分析和推理，从中发现隐含的知识、揭示潜在的规律。该理论的主要特点是：它无需提供问题所需处理的数据集合之外的任何先验信息，这是其与证据理论和模糊集合理论的最主要区别，也是最重要的优点。目前粗糙集理论已经在模式识别、决策支持与分析及智能控制等许多领域得到了广泛的应用[1-7]。本文应用粗糙集理论构

基金项目：湖南省教育厅科学研究项目(11C1176)。
作者简介：刘晓玲(通讯作者)，硕士，讲师，研究方向为计算机应用、网络安全，E-mail：lxlwork@126.com；谢仙斌，本科，工程师，研究方向为网络通信、数字电视；张家录，硕士，教授，研究方向为非经典数理逻辑与近似推理、粗糙集理论与应用；刘灵丽，硕士，副教授，研究方向图像处理。

建网络安全评估决策支持系统，基本思想是：将网络安全评估指标集作为决策信息系统 $S=(U,C\cup D,V,f)$ 的条件属性集 C，将评估结果作为决策属性集 D，通过对其决策表的约简产生决策规则集，然后以决策规则集为基础建立网络安全评估的决策支持系统。

已有一些文献利用粗糙集理论对网络安全评估进行了探讨[8-10]，但大部分只就网络安全评估指标集(条件属性集)的离散值进行讨论。而一个明显的事实是：网络安全评估指标数据项的类型较为复杂，既有离散型数据(如整型、字符串型、枚举型)又有连续值或定性值，为此需要对网络安全评估的样本数据进行离散化处理，但离散化处理也会带来一些信息上的损失[11]。本文结合网络安全评估指标数据的特点，将各种数据类型统一处理，运用模糊粗糙集理论对网络安全进行评估。

2　预备知识

设 U 是非空有限集，称为论域。$\forall X\subseteq U$，称为 U 中的一个概念。若干个概念 $F=\{X_1,\cdots,X_l\}$称为关于 U 的一个知识，其中 $X_i\subseteq U, X_i\neq\Phi$, $X_i\cap X_j\neq\Phi$, $i,j=1,2,l$ 且有 $\cup X_i=U$。为规范起见，空集 Φ 也被认为是一个概念。又设 $S=(U,C\cup D,V,f\}$是一个决策信息系统。由 $A\subseteq C\cup D$ 可以确定一个不可分辨关系 $ind(A)=\{(u_i,u_j)\mid u_i,u_j\in U, f(u_i,a)=f(u_j,a),a\in A\}$ 不可分辨关系 $ind(A)$显然是一个等价关系。

设 R 是 U 上的一个等价关系，$U/R=\{X_1,\cdots,X_n\}$ 表示 R 产生的分类，U/R 显然是 U 的一个知识。$[x]_R=\{y\in U\mid xRy\}$表示由元素 x 产生的 R 等价类。(U,R)称为近似空间。设 P 是 U 上的等价关系族，则知识库可表示为 $K=(U,P)$。

设 R 是 U 上的一个等价关系，$X\subseteq U$。当 X 为 R 的某些等价类的并时，称 X 是 R 可定义的，否则称 X 为不可定义的，R 可定义集称作为 R 精确集，R 不可定义集称为 R 粗糙集。粗糙集可以用两个精确集(下近似和上近似)来描述。

包含在 X 中的最大可定义集称为 X 的 R 下近似：

$$\underline{R}(X)=\{x\in U\mid [x]_R\subseteq X\}$$

包含 X 的最小可定义集称为 X 的 R 上近似：

$$\overline{R}(X)=\{x\in U\mid [x]_R\cap X\neq\Phi\}$$

$\underline{R}(X)$表示在知识 R 下 U 中所有一定能归入 X 的元素的集合，$\overline{R}(X)$表示在知识 R 下 U 中可能归入 X 的元素的集合。

设有决策信息系统 $S=(U,C\cup D,V,f)$(有时也将之简记为 $S=(U,C\cup D)$)，其中 C,D 分别表示条件属性和决策属性，则决策属性在条件属性下的正域定义为：

$$POS_C(D)=\bigcup_{X\in U/D}\underline{R_C}(X)$$

$POS_C(D)$表明根据 C 所进行的划分 U/C，能够确切地划入 U/D 类的对象集合。对一个属性 $a\in C$，如果 $POS_C(D)=POS_{C-\{a\}}(D)$，则称 a 是独立的，否则称 a 不是独立的，并称

$$\delta(a)=\frac{(\mid POS_C(D)-POS_{C-\{a\}}(D)\mid)}{\mid U\mid}$$

为属性 a 的重要性。如果 $POS_C(D)=U$，则称决策信息系统是相容的，否则称之为不相容的。

决策属性 D 对条件属性 C 的依赖度定义为：

$$\gamma=\gamma_C(D)=\frac{\mid POS_C(D)\mid}{\mid U\mid}$$

显然，一个决策信息系统是相容的当且仅当$\gamma=1$。

对决策系统 $S=(U,C\cup D)$，每一个对象(亦即每一行)都可确定一个决策规则。例如，对象 $u\in U$ 可确定决策规则 $d_u:d_u\mid C\rightarrow d_u\mid D$，这里 $d_u\mid C,d_u\mid D$ 分别表示对象 u 的条件属性值和决策属性值。如果当 $d_{u_i}\mid C=d_{u_j}\mid C$ 时，有 $d_{u_i}\mid D=d_{u_j}\mid D$，则称 d_u 是相容的决策规则，否则称 d_u 为不相容的决策规则。

3　基于粗糙集理论的网络安全评估模型

3.1　网络安全评估指标约简

由于粗糙集是基于符号运算的离散知识推理系统，要求决策表中的值必须是离散数据。网络安全评估指标数据项通常为连续值或定性值，为此需要对网络安全评估的样本数据进行预处理。数据预处理需要将定性值进行定量化，通常采用等距离划分、等频率划分、Boolean 逻辑与 Rough 集理论相结合的离散化算法及基于属性重要性的离散化算法等[3]，但这样处理也会带来信息上的一些损失。结合网络安全评估指标数据的特点，本文采取如下方法进行样本数据预处理。

定义　3.1 设 $S=(U,C\cup\{d\})$是一个网络安全决策系统，$V_a=\{v_1,v_2,\cdots,v_l\}$是网络安全评估指标集(条件属性集)，$d$ 是评估结果(决策属性)。如果 a 是定性属性(如字符串型、枚举型)，则属性值 v_i,v_j 的相似度定义为：

$$\mu_a(v_i,v_j)=\begin{cases}1, v_i=v_j\\ 0, v_i\neq v_j\end{cases}$$

如果 a 是离散数值型属性(如整型)，则属性值 v_i,v_j 的

相似度定义为：

$$\mu_a(v_i,v_j)=1-\frac{|v_i-v_j|}{M-1}$$

其中 M 是离散属性值个数。如果 a 是连续值属性(如实型)，则属性值 v_i,v_j 的相似度定义为：

$$\mu_a(v_i,v_j)=1-\frac{|v_i-v_j|}{a_{\max}-a_{\min}}$$

其中 $a_{\max}$ 和 a_{mix} 分别是属性 a 的最大值和最小值。

定义 3.2 设 $S=(U,C\cup\{d\})$是一网络安全评估系统，$C=\{a_1,a_2,\cdots,a_l\}$是评估指标集，d 是评估结果，则对象 x_i, x_j 的相似度定义为：

$$\mu_C(x_i,x_j)=\frac{1}{l}\sum_{k=1}^{l}\mu_{a_k}(v_{i_k},v_{j_k})$$

其中 v_{i_k},v_{j_k} 分别是对象 x_i, x_j 关于评估指标 a_k 的值。

有些属性虽然是离散枚举型，但如果其值可以排序，则可以将其转化为离散数值型。如，网络安全评估中的指标安全漏洞通常取三个枚举型值：高危、危险、一般，它们可以按照危害程度对它们排序，从而在计算元素之间的相似度时可以将它们转化为离散数值型。

决策信息系统属性约简算法：

输入：一个决策信息系统 $S=(U,C\cup\{d\})$，这里 $C=\{a_1,a_2,\cdots,a_l\}$是取连续值或离散值的条件属性，d 是取离散值的决策属性；

步骤 1. 计算对象之间的两两相似度(计算方法如定义 3.2)，得到模糊相似矩阵 R；

步骤 2. 计算模糊矩阵 R 的传递闭包 $\hat{R}$($\hat{R}$ 是一个模糊等价关系)，并按如下方法确定阈值 λ：(1)将模糊矩阵 $\hat{R}$ 各元素值从大到小排序，如 $\alpha_1>\alpha_2>\cdots>\alpha_k$($k$ 是 $\hat{R}$ 中不同值的个数)；(2)令 $i=1$；(3)对 $\lambda_i\in(\alpha_{i+1},\alpha_i]$，由 $\hat{R}$ 对 U 进行 λ_i 水平上的聚类(由条件属性确定的等价类)，即若 $\mu_{\hat{R}}(x_i,x_j)>\lambda_i$，则 x_i, x_j 在同一类，并计算 λ_i 水平上的相容度。若其相容度小于 1，则转(5)；(4)若 $i<k-1$，则 $i=i+1$，转(3)；(5)确定模糊聚类的阈值是 α_i；

步骤 3. 初始化属性约简集：$B\leftarrow C$；

步骤 4. 令 $j=1$；

步骤 5. 对 $a_j\in B$，在阈值 α_i 下，针对条件属性集 $B-\{a_j\}$，采用与定义 3.2 和 步骤 2 的方式对 U 进行分类。若 $POS_{B-\{a_j\}}(D)=POS_B(D)$，则 $B\leftarrow B-\{a_j\}$；

步骤 6. 若 $j<n$，则 $j=j+1$，转 步骤 5；

步骤 7. 输出属性约简集 B 。

对决策信息系统 $S=(U,C\cup\{d\})$，如果 C 中所有的属性都是不可分辨的并且 U 中没有属性值完全相同的重复对象，则称 $S=(U,C\cup\{d\})$是完全约简的。

3.2 网络安全评估规则提取

根据上面的分析，计算机网络安全评估可以看作一个属性值连续的决策系统 $S=(U,C\cup\{d\}V,f)$，其中非空有限集 $X=\{x_1,x_2,\cdots,x_n\}$表示网络安全评估样本集；条件属性集 $C=\{a_1,a_2,\cdots,a_m\}$表示网络安全评估指标集；决策属性 d 表示网络安全的评估结果；$V=V_C\cup V_d$ 是属性值集合，$V_C=\{V_a\,|a\in C\}$是条件属性值集，V_d 是决策属性值集，第 i 个对象在第 j 个条件属性上的取值 $v_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,m)$是连续变化的；信息函数 $f:U\times(C\cup\{d\})\rightarrow V$ 表示对 $\forall a\in C,x\in U$ 有 $f(x,a)\in V_a$。

根据以上讨论，本文提出基于混合属性值(连续值、定性离散值、定量离散值)网络安全评估规则。为了消除不同指标的度量单位不同所带来的影响，在计算相似度矩阵之前，首先对连续值数据进行规范化。

步骤 1. 规范化连续属性值，其公式如下：

$$x'=\frac{x-a_{\min}}{a_{\max}-a_{\min}}$$

其中 $a_{\max}$ 和 a_{mix} 分别是属性 a 的最大值和最小值。

步骤 2. 按照以上提出的属性约简算法对网络安全评估系统 $S=(U,C\cup\{d\})$进行约简，得到约简后的决策信息系统 $S=(U,C'\cup\{d\})$，$C'=\{a_1,\cdots,a_s\}$。

步骤 3. 由 $S=(U,C'\cup\{d\})$和属性约简时所确定的阈值 λ，得到根据安全评估指标 C' 和评估结果的等价分类 $U/ind(C')=\{X_1,\cdots,X_n\}$，$U/ind(d)=\{Y_1,\cdots,Y_m\}$。

步骤 4. 对每一个 $u\in X_i$，$u=(a_1^i,\cdots,a_m^i)$，这里如果 a_j 是离散属性值，则 $a_j^i=a_j$；如果 a_j 是连续属性值，则：

$$a_j^i=\frac{\sum_{u_k\in X_i}f(u_k,a_j)}{|X_i|}$$

步骤 5. 构造决策规则：$\wedge_{k=1}^{m}(a_k,a_k^i)\rightarrow Y_{ki}$，这里 $X_i\in Y_{ki},i=1,2,\cdots,n,k_i=1,2,\cdots,s$。

步骤 6. 将[0,1]区间值转化为原来的值，公式为：$x=(a_{\max}-a_{\min})x'+a_{\min}$。

4 基于粗糙集理论的网络安全评估实例

为了说明方便，我们用一个简化的网络安全数据集来验证本节所提出的基于粗糙集理论的网络安全评估规则提取模型。

假设数据集样本包含 6 个条件属性(评估指标)、一个决策属性(评估结果)，即决策表的条件属性集为：a_1 安全漏洞，a_2 物理设备故障，a_3 操作系统类型，a_4 应用软件的软件缺陷，a_5 网络管理缺陷，其中 a_2 是离散数值型属性，a_1、a_3 是离散枚举型属性，a_4、a_5 采用连续的百分制来描述。决策属性 d 为相应的安全评估结果。

假设某网络安全评估数据集样本如表 1 所示。

首先，在计算元素之间的相似度时，将连续属性值规范化，对可以排序的离散枚举型属性值 a_1，将高危、危险、一般分别离散数值化 2、1、0，得到如下的表 2。

其次，按照定义 3.1 和 3.2 计算对象之间的相似度，得到相似度矩阵 R 如下：

$$\begin{pmatrix} 1 & 0.35 & 0.40 & 0.20 & 0.55 & 0.40 & 0.60 & 0.35 & 0.35 \\ 0.35 & 1 & 0.35 & 0.55 & 0.70 & 0.45 & 0.55 & 0.70 & 0.60 \\ 0.40 & 0.35 & 1 & 0.40 & 0.35 & 0.80 & 0.60 & 0.65 & 0.25 \\ 0.20 & 0.55 & 0.40 & 1 & 0.55 & 0.20 & 0.60 & 0.45 & 0.55 \\ 0.55 & 0.70 & 0.35 & 0.55 & 1 & 0.45 & 0.55 & 0.60 & 0.90 \\ 0.40 & 0.45 & 0.80 & 0.20 & 0.45 & 1 & 0.60 & 0.45 & 0.25 \\ 0.60 & 0.55 & 0.60 & 0.60 & 0.55 & 0.60 & 1 & 0.55 & 0.55 \\ 0.35 & 0.70 & 0.65 & 0.45 & 0.60 & 0.45 & 0.55 & 1 & 0.50 \\ 0.35 & 0.60 & 0.25 & 0.55 & 0.90 & 0.25 & 0.55 & 0.50 & 1 \end{pmatrix}$$

再次，计算 R 的传递闭包 $\overline{R}$ 为

$$\begin{pmatrix} 1 & 0.55 & 0.60 & 0.60 & 0.55 & 0.60 & 0.60 & 0.60 & 0.60 \\ 0.55 & 1 & 0.65 & 0.55 & 0.70 & 0.55 & 0.70 & 0.70 & 0.70 \\ 0.60 & 0.65 & 1 & 0.60 & 0.60 & 0.80 & 0.60 & 0.65 & 0.55 \\ 0.60 & 0.55 & 0.60 & 1 & 0.55 & 0.60 & 0.60 & 0.60 & 0.65 \\ 0.55 & 0.60 & 0.60 & 0.55 & 1 & 0.55 & 0.70 & 0.70 & 0.90 \\ 0.60 & 0.55 & 0.80 & 0.60 & 0.55 & 1 & 0.60 & 0.65 & 0.55 \\ 0.60 & 0.70 & 0.60 & 0.60 & 0.70 & 0.60 & 1 & 0.70 & 0.70 \\ 0.60 & 0.70 & 0.65 & 0.60 & 0.70 & 0.65 & 0.70 & 1 & 0.70 \\ 0.60 & 0.70 & 0.55 & 0.65 & 0.90 & 0.55 & 0.70 & 0.70 & 1 \end{pmatrix}$$

表 1　假设某网络安全评估数据集样本

U	a_1	a_2	a_3	a_4	a_5	d
u_1	高危	3	Windows Server 2003	80	80	较差
u_2	一般	1	Linux	80	85	安全
u_3	危险	2	Unix	70	70	较差
u_4	一般	1	Vista	90	90	安全
u_5	一般	2	Windows Server 2008	80	85	一般
u_6	危险	2	Linux	70	70	较差
u_7	危险	2	Vista	80	80	一般
u_8	危险	1	Unix	80	85	一般
u_9	一般	2	Windows Server 2003	90	85	安全

表 2　对 a_1 值离散数值化后得到的结果

U	a_1	a_2	a_3	a_4	a_5	d
u_1	2	3	Windows Server 2003	0.5	0.5	较差
u_2	0	1	Linux	0.5	0.75	安全
u_3	1	2	Unix	0	0	较差
u_4	0	1	Vista	1	1	安全
u_5	0	2	Windows Server 2008	0.5	0.75	一般
u_6	1	2	Linux	0	0	较差
u_7	1	2	Vista	0.5	0.5	一般
u_8	1	1	Unix	0.5	0.75	一般
u_9	0	2	Windows Server 2008	1	0.75	安全

表 3 约简后的网络安全评估表

U	a_1	a_2	a_4	a_5	d
u_1	2	3	0.5	0.5	较差
u_2	0	1	0.5	0.75	安全
u_3	1	2	0	0	较差
u_4	0	1	1	1	安全
u_5	0	2	0.5	0.75	一般
u_7	1	2	0.5	0.5	一般
u_8	1	1	0.5	0.75	一般
u_9	0	2	1	0.75	安全

由属性约简算法，经过计算得到网络安全评估表的属性约简集 $B=\{a_1,a_2,a_4,a_5\}$，约简后的网络安全评估表如表 3。

最后，根据约简后的网络安全评估表得到网络安全评估规则如下：

$(a_1,2)\wedge(a_2,3)\wedge(a_4,0.5)\wedge(a_5,0.5)\rightarrow(d,\text{较差})$

$(a_1,0)\wedge(a_2,1)\wedge(a_4,0.5)\wedge(a_5,0.75)\rightarrow(d,\text{安全})$

$(a_1,1)\wedge(a_2,2)\wedge(a_4,0)\wedge(a_5,0)\rightarrow(d,\text{较差})$

$(a_1,0)\wedge(a_2,1)\wedge(a_4,1)\wedge(a_5,1)\rightarrow(d,\text{安全})$

$(a_1,0)\wedge(a_2,2)\wedge(a_4,0.5)\wedge(a_5,0.75)\rightarrow(d,\text{安全})$

$(a_1,1)\wedge(a_2,2)\wedge(a_4,0.5)\wedge(a_5,0.5)\rightarrow(d,\text{一般})$

$(a_1,1)\wedge(a_2,1)\wedge(a_4,0.5)\wedge(a_5,0.75)\rightarrow(d,\text{一般})$

$(a_1,0)\wedge(a_2,2)\wedge(a_4,1)\wedge(a_5,0.75)\rightarrow(d,\text{安全})$

将[0, 1]区间值转化为原来的值，得到最终的评估规则：

$(a_1,\text{高危})\wedge(a_2,3)\wedge(a_4,80)\wedge(a_5,80)\rightarrow(d,\text{较差})$

$(a_1,\text{一般})\wedge(a_2,1)\wedge(a_4,80)\wedge(a_5,85)\rightarrow(d,\text{安全})$

$(a_1,\text{危险})\wedge(a_2,2)\wedge(a_4,70)\wedge(a_5,70)\rightarrow(d,\text{较差})$

$(a_1,\text{一般})\wedge(a_2,1)\wedge(a_4,90)\wedge(a_5,90)\rightarrow(d,\text{安全})$

$(a_1,\text{一般})\wedge(a_2,2)\wedge(a_4,80)\wedge(a_5,85)\rightarrow(d,\text{安全})$

$(a_1,\text{危险})\wedge(a_2,2)\wedge(a_4,80)\wedge(a_5,80)\rightarrow(d,\text{一般})$

$(a_1,\text{危险})\wedge(a_2,1)\wedge(a_4,80)\wedge(a_5,85)\rightarrow(d,\text{一般})$

$(a_1,\text{一般})\wedge(a_2,2)\wedge(a_4,90)\wedge(a_5,85)\rightarrow(d,\text{安全})$

5 结果分析

由于使用的网络安全样本数量较少，得到的评估规则还不能全面反映计算机网络安全的评估过程，但所得结果与实际基本相符。在实际应用中，为了使得规则的描述更为合理，可将规则条件中的连续值必为区间值。如果有足够多的网络攻击评估样本，就可以得到详细的决策规则，然后以此为基础构建计算机网络安全评估的决策支持系统。

参 考 文 献

[1] 熊丽君, 许龙飞. Rough set 理论及其应用研究进展综述 [J]. 暨南大学学报(自然科学版), 2003, 24(3): 70-75.

[2] Jensen R, Shen Q. Fuzzy-rough sets for descriptive dimensionality reduction [C] // Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, 2002.

[3] 王国胤. Rough 集理论与知识获取 [M]. 西安: 西安交通大学出版社, 2001.

[4] 何亚群. 基于粗糙集的智能决策理论与应用研究 [D]. 南京: 南京航空航天大学, 2004.

[5] Anna MR, Etienne EK. A comparative study of fuzzy rough set [J]. Fuzzy Sets and System, 2002,126(2): 137-155.

[6] Kankana C, Ranjit B, Sudarsan N. Fuzziness in rough set [J]. Fuzzy Sets and System, 2000, 110(2): 247-251.

[7] 魏大宽, 黄兵, 周献中. 不完备模糊目标信息系统粗集模型与知识约简 [J]. 计算机工程, 2006, 32(8): 48-51.

[8] 陈志杰, 王永杰. 一种基于粗糙集的网络安全评估模型 [J]. 计算机科学, 2007, 34(8): 98-100.

[9] 梁颖, 王慧强, 赖积保. 一种基于粗糙集理论的网络安全态势感知方法 [J]. 计算机科学, 2007, 34(8): 95-97.

[10] Wang XR, He FM, Wang YL, et al. Network security evaluation research based on rough set theory [C] // International Conference on Circuit and Signal, 2010: 304-306.

[11] 张家录. 随机模糊信息系统的知识约简 [J]. 工程数学学报, 2006, 23(3): 450-454.