一种基于粗集理论的遗传分类算法*)

姚 前[1] 陈 舜[1] 谢 立[1] 张 军[2]
(南京大学计算机科学与技术系 南京 210008)[1] (首都经济贸易大学信息学院 北京 100026)[2]

摘 要 本文提出了一种基于粗集理论的遗传分类算法,该算法可以无需任何辅助信息,只根据数据自身提供的信息对数据进行简化,提取有用的特征,并求得相应的规则。同时,还提出了一种基于属性重要度的分辨矩阵简化方法,该方法可提高对条件属性的约简效率。

关键词 粗集理论,遗传分类算法,网络告警信息处理

A Rough Set Theory Based Genetic Algorithm

YAO Qian[1] CHEN Sun[1] XIE Li[1] ZHANG Jun[2]
(Department of Computer and Technology, Nanjing University, Nanjing 210008)[1]
(Capital University of Economics and Business, Information College, Beijing 100026)[2]

Abstract A Rough Set Theory based genetic algorithm is proposed in this paper, this algorithm can make reduction on data according to the data itself, gain useful information and get the right rules without any assistant information. A method to get the discernibility matrix is also proposed, it can increase the efficiency of reduction of knowledge. Applying the proposed algorithm in network security alarm processing, the system can handle large scale alarm data and get concise rules.

Keywords Rough set theory, Genetic algorithm, Network security alarm process

1 引言

在各类网络安全设备所给出的网络告警数据中,常常含有大量无关或冗余的信息,告警综合处理就是一个通过对这些信息进行关联与综合分析,实现对告警的压缩的过程。数据挖掘算法是告警综合处理的一种典型算法。通过属性约简,选出与任务相关的属性,同时去除无关冗余的属性,对于减少数据挖掘算法的执行时间、提高数据挖掘系统的效能起着重要的作用。

粗集理论和遗传算法均为近年来快速发展的智能算法。粗集理论可以无需任何辅助信息,只根据数据自身提供的信息对数据进行简化,提取有用的特征,并求得知识的最小表达。遗传算法是一种模拟自然界优胜劣汰的原则的随机搜索算法,具有并行性、全局性、收敛快等特点,它已经在机器学习系统中表现出良好的学习能力。

粗集理论提出了研究不完整数据及不精确知识的表达、学习、归纳的一套方法,能有效地进行数据约简,提取有用信息,简化决策规则,提高决策效率。粗集理论使用"下近似"和"上近似"的概念来处理不确定性问题,不需要任何关于数据的先验信息,而是直接从给定问题的描述集合出发,通过不可分辨关系和不可分辨类确定规定问题的近似域,从而从数据中获取规则知识。采用粗集理论进行数据约简将大大减少机器学习中涉及的对象和属性的数量,算法简单,但是,粗集挖掘方法获取的规则分散重复,且较复杂,不利于进行规则匹配。

遗传算法是一种很好的具有学习功能的算法,它不仅可以对规则进行优化和求精,还能从现有的规则中获取未发现的新知识。也就是说,它可以产生使系统更适合于环境的新规则,并能对以前的规则给予新的可信度,甚至可以剔除不再适用的旧规则,对规则进行求精,以防止知识库的不断扩大而影响知识处理的速度。

基于这两种算法各自的优点,本文提出了一种有效的知识获取方法——基于粗集理论的遗传分类算法,该算法能处理大规模告(预)警数据及其不一致性,获得简洁的告(预)警综合规则。为提高对条件属性的约简效率,本文还提出了一种基于属性重要度的分辨矩阵简化方法。本文所提出的基于粗集理论的遗传分类方法的主要思想是:在对数据进行处理时,首先采用粗集理论计算出属性核值及各个属性的重要度,对原始数据进行属性约简和属性值约简;然后用遗传算法对约简后的数据进行知识挖掘,得到最优规则。由于约简不仅可以减少待处理的数据,而且使得整个挖掘工作是在有意义的数据上进行的,从而提高了算法的效率,加快了学习速度。

2 粗集理论与遗传算法的基本概念与原理

2.1 粗集理论的基本概念

粗集理论(Rough Set Theory,简称 RST)是在 20 世纪 80 年代由波兰学者 Pawlak Z. 提出的一种研究不完整性和不确定性的数学工具,能有效地分析和处理不精确、不一致、不完整等各种不完备信息,并从中发现隐含的知识,揭示潜在的规律。粗集方法在数据挖掘和知识发现方面的应用已经越来越引起人们的关注,主要用于从知识库中发现新知识或挖掘出潜在的新知识,找出其内部数据的关联关系和特征。

2.1.1 知识表达系统

基于粗集理论的观点,可把客观对象世界抽象为一个知识表达系统,也称属性-值系统:

$$S=\langle U,A,V,f\rangle \tag{1}$$

*)基金项目:国家十五科技攻关项目(金融示范工程),课题编号:2001BA102A04。姚 前、陈 舜 博士生,主要研究方向为分布式系统和计算机安全;谢 立 教授,博士生导师,主要研究方向为分布式计算、并行处理、先进操作系统等;张 军 博士,教授。

其中：$U=\{x_1,x_2,\cdots,x_n\}$为论域；

$A=C\cup D$为属性集合(等价关系集合)；

子集$C=\{a_1,a_2,\cdots,a_m\}$和$D=\{d\}$分别称为条件属性集和决策属性集，故也称S为决策系统，V是属性值的集合；

$f:U\times A\to V$为信息函数，它指定U中的每一对象x的属性值。知识表达系统可以用表格表示，表格的行表示对象(样本)，列表示属性，对象与属性的交会点就是对象在该属性下的值，故知识表达系统的表格表达也称为决策表，知识表达系统也称为决策信息系统。

2.1.2 不可分辨等价关系与分辨矩阵

对于知识表达系统S，设$B\subseteq A$，不可分辨等价关系IND(B)定义为：

$$IND(B)=\{(x,y)\in U\times U\mid \forall b\in B, f(x,b)=f(y,b)\} \tag{2}$$

$IND(B)$将U划分为k个类$X_1,X_2,\cdots,X_k$，即

$$U/IND(B)=[X]IND(B)=\{X_1,X_2,\cdots,X_k\} \tag{3}$$

用$des\{X_i\}$表示U上的一个等价类描述。

分辨矩阵(Discernibility Matrix)是一个$n\times n$的方阵$M_D(B)$：

$$M_D(B)=\{M_D(i,j)\}_{n\times n},\quad 1\leqslant i,j\leqslant n=|U/IND(B)| \tag{4}$$

其中，$M_D(i,j)$表示分辨矩阵中第i行第j列的元素：

$$M_D(i,j)=\begin{cases}\{a_k\mid a_k\in B\wedge a_k(x_i)\neq a_k(x_j)\}, & d(x_i)\neq d(x_j)\\ 0, & d(x_i)=d(x_j)\\ \phi, & a_k(x_i)=a_k(x_j)\wedge d(x_i)\neq d(x_j)\end{cases} \tag{5}$$

其中，$a_k(x_j)$是样本x_j在属性a_k上的取值

$d(x_i)$是样本x_i的决策属性值，$i,j=1,2,\cdots,n$。$M_D(i,j)$是属性集，它表明了两个类在该属性集上的值不同。

任意分辨矩阵$M_D(B)$都可由其元素$M_D(i,j)$唯一地确定一个分辨函数$F_{M_D(B)}$：

$$F_{M_D(B)}=\bigwedge_{M_D(i,j)\neq 0,M_D(i,j)\neq\phi}\ \bigvee_{a_i\in M_D(i,j)} a_i \quad \text{其中 } i,j=1,2,\cdots,n \tag{6}$$

2.1.3 下近似与上近似

对于知识表达系统S，每个子集$X\subseteq U$，B为属性子集，集合X关于B的下近似和上近似分别定义为：

$$B_*(X)=\bigcup\{Y_i\mid Y_i\in U/IND(B)\wedge Y_i\in X\} \tag{7}$$

$$B^*(X)=\bigcup\{Y_i\mid Y_i\in U/IND(B)\wedge Y_i\cap X\neq\phi\} \tag{8}$$

对下近似$B_*(X)$应建立确定性规则，对上近似$B^*(X)$应建立可能性规则(含可信度)。$B_*(X)$也称为X的B正域，记作$POS_B(X)$：

$$POS_B(X)=B_*(X) \tag{9}$$

X的B负域是指由根据已有知识判断肯定不属于X的对象组成的集合，记作$NEG_B(X)$：

$$NEG_B(X)=U-B^*(X) \tag{10}$$

集合X关于B的边界区定义为：

$$BN_B(X)=B^*(X)-B_*(X) \tag{11}$$

$BN_B(X)$是根据B、U中既不能肯定属于X、又不能肯定不属于X的对象组成的集合。若$BN_B(X)$是空集，则称X关于B是清晰的(crisp)；若$BN_B(X)$不是空集，则称X为关于B的粗集。

集合X关于B的近似精度$\alpha_B(X)$定义为：

$$\alpha_B(X)=\frac{|B_*(X)|}{|B^*(X)|} \tag{12}$$

式中：$|B_*(X)|$表示集合$B_*(X)$的基数或势card，对有限集合来说表示集合中所包含元素的个数。

2.1.4 约简与核

属性的约简与核是粗集理论的两个重要的基本概念。约简是指可以根据原始信息系统的数据对可分辨对象进行分辨的最小的属性集合，核指的是一个信息系统中所有约简的交集。在决策系统中，约简可以最简单地表示决策系统的结论属性D对条件属性集合C的依赖，D对于C的依赖度定义为：

$$k(C,D)=\frac{\text{card}(POS_C(D))}{\text{card}(U)}=\frac{|POS_C(D)|}{|U|} \tag{13}$$

其中card()和||表示集合的基数。

对于属性$p\in C$，如果$k(C-\{p\},D)=k(C,D)$，则p在决策中相对于D是冗余的，否则是不可缺少的。而属性p在C和D中的重要度定义为：

$$SGF(p,C,D)=k(C,D)-k(C-\{p\},D) \tag{14}$$

如果C中的任意一个属性关于D是不可缺少的，则C关于D是正交的；对于$B\in C$，如果$k(B,D)=k(C,D)$，而且B是正交的，则B是C的D约简，记为：$RED_D(C)$。一般情况下，C的D约简有多个。C的所有属性约简的交集称为核，记作：$CORE_D(C)$，$CORE_D(C)=\cap RED_D(C)$。

2.2 遗传算法的基本原理

遗传算法是一种概率搜索算法，通过模拟自然界优胜劣汰的生物进化机制，在较大的参数空间中快速搜索问题的最优解，尤其适用于处理传统搜索方法中难以解决的复杂和非线性问题。遗传算法不仅避免了局部优化算法的缺陷，而且可以利用固有知识缩小搜索空间，避免其它全局优化算法产生搜索的组合爆炸。基于遗传算法的机器学习把遗传算法从历来离散的搜索空间的优化搜索算法扩展到具有独特的规则生成功能的崭新机器学习算法，可解决人工智能中知识获取和知识优化精练的“瓶颈”。

2.3 分类学习的基本要求

机器学习主要关心分类问题，它是许多其它问题的基础和核心。分类学习的过程就是通过对具有类别标记的实例数据进行训练，得出一个能够预测新实例类别的模型。其目的是在不同的实体中寻找共同特征，并按照特征将实体分成不同的类别。

分类学习需要有一个训练样本数据集作为输入。训练集由一组数据库记录或元组构成，每个元组是一个由有关字段值组成的特征向量，除此之外，训练样本还有一个类别标记。一个具体分类器样本的形式为：

$$(v_1,v_2,\cdots,v_n,c)$$

其中：v_i表示字段值，

c表示类别。

不同的分类器有不同的特点，评价一个分类器的尺度有三种：

① 预测准确度；

② 计算复杂度；

③ 模型描述的简洁度。

其中预测准确度是用得最多的一种比较尺度，特别是对于分类任务。计算复杂度依赖于具体的实现细节和硬件环境，操作对象是海量数据库，因此空间和时间复杂度问题是一个非常重要的环节。对于描述型的分类任务，模型描述越简洁越好。

3 基于粗集的遗传分类算法

3.1 基于粗集的遗传分类算法的结构

基于粗集理论的遗传分类算法利用粗集理论在处理大数据量、消除冗余信息等方面的优势，约简训练数据，寻找最小属性集，从中获取粗规则，然后应用遗传算法对获取的粗规则进行进一步的挖掘，得到最优规则。其框架结构如图1所示，由三部分组成：预处理、基于粗集理论的遗传分类算法、后处理。

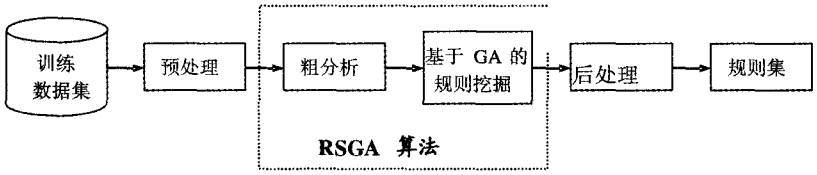

图1 粗集遗传算法的框架结构图

预处理部分主要是将训练数据集中的初始数据信息转换为决策表形式，明确条件属性和决策属性，并将属性的语言描述转换为实数。如果条件属性中包含连续值属性，需对其进行离散化。预处理还需对遗传算法的有关参数（如种群的大小、迭代次数、遗传算子的选择、停机条件等）进行初始化等。基于粗集理论的遗传分类算法由粗分析和基于遗传算法的规则挖掘两部分组成。后处理的作用是对产生的规则进行检查去除冗余规则，生成通用而简洁的规则。

3.2 连续属性值离散化

基于粗集理论的遗传分类算法要求属性的特征具有逻辑型或离散型，因此需要对连续型属性进行离散化。

根据离散化过程中是否考虑类别属性（即决策属性），可以将离散化算法分为有监督算法和无监督算法两种。有监督算法在离散化时利用类别属性的值作为参考，其输入样本集中除了待离散化的数值属性外，还具有一个或多个离散型的类别属性；而无监督算法在离散化过程中不考虑类别属性，其输入样本集中仅含带离散化属性的值，可将该样本集看成一维数组。

基于粗集理论的遗传分类算法的目标是基于训练样本集产生分类规则，以此预测新样本的类别，故采用有监督算法—Naive Scaler，在保证决策一致的条件下，将连续属性离散化。Naive Scaler 算法不需要额外的参数，能够根据决策表数据库本身进行离散化处理。

3.3 基于粗集的遗传分类算法

基于粗集理论的遗传分类算法由粗分析和基于遗传算法的规则挖掘两部分组成。粗分析在数据挖掘中的主要功能是用粗集方法实现数据约简和粗规则获取，基于遗传算法的规则挖掘对获取的粗规则进行优化。其基本框架可归纳为：

(1)以不可分辨关系对所研究领域的知识进行约简，生成约简属性集；

(2)进行属性值约简，得到最小决策规则；

(3)进行一致性判断，利用上、下近似集逼近描述对象，从而获得粗规则；

(4)用遗传算法进行规则挖掘；

(5)利用粗集理论提取规则。

基于粗集理论的规则获取主要是通过对原始决策表约简，在保持决策表中决策属性和条件属性之间的依赖关系不发生变化的前提下对决策表进行简化，包括属性约简和值约简。

3.3.1 属性约简

决策表属性约简的过程，就是从决策表系统的条件属性中去掉不必要的属性，从而分析所得约简中的条件属性对于决策属性的决策规则。

分辨矩阵的元素取值为条件属性的合取项，分辨函数的最小简化的析取范式对应于决策表 S 的全体约简。如果矩阵的元素取值仅含单个属性，则该属性一定是决策表的核值，故分辨矩阵中所有只包含单属性的元素组成的单属性的集合就构成了决策表的相对属性核，即条件属性 C 的 D 核 $\mathrm{CORE}_D(C)$。

利用分辨矩阵计算约简的方法，其时间和空间复杂度是随着决策表 S 的大小成指数变化的，而且要求生成中间环节——分辨矩阵，必然造成时空上不必要的开销。基于分辨矩阵的简化方法就是一边从信息系统中提取关于属性值是可分辨的属性并构成分辨合取范式，一边作这种逻辑公式的等价变换，直接得到最小析取范式，它对应于该信息系统的诸多约简，这样就不用生成可分辨矩阵的中间环节，从而节省了空间和时间，提高了计算机执行约简算法的运行效率。

利用分辨矩阵的简化方法求解决策表的约简时，如果出现单个属性为一合取项，则这些属性构成的集合是核集，约简后得到的分辨合取范式都可等价地变换为最小分辨析取范式，其中每个析取项是对应于该信息系统的一个约简。

条件属性集 C 一般有多个 D 简化，$\mathrm{RED}_D(C)$ 为 C 的 D 简化族，选取属性个数最少并且在这个条件下等价类个数最少的简化 S 作为 C 的 D 最小简化，记为 $\mathrm{MinRED}_D(C)$。由于计算约简是 NP-hard，本文提出一种计算约简的启发式约简方法——基于属性重要度的分辨矩阵简化法。其主要思想：首先利用分辨矩阵的简化方法求出条件属性 C 的 D 核 $\mathrm{CORE}_D(C)$ 和 C 的 D 简化族 $\mathrm{RED}_D(C)$，核包含在任一简化之中，因此求取最小简化可以从核开始，在 $C'=\mathrm{RED}_D(C)-\mathrm{CORE}_D(C)$ 中利用属性重要度进行启发式约简。它与其它的启发式约简方法的差别在于：它的搜索范围 C' 要比 C 小得多，提高了算法的效率。

基于属性重要度的分辨矩阵简化算法的实现步骤如下：

(1)利用分辨矩阵的简化方法求条件属性 C 的 D 核 $\mathrm{CORE}_D(C)$ 和 C 的 D 简化族 $\mathrm{RED}_D(C)$；

(2)令 $\mathrm{MinRED}_D(C)=\mathrm{CORE}_D(C)$；

(3)如果 $\mathrm{POS}_{\mathrm{MinRED}_D(C)}(D)=\mathrm{POS}_{\mathrm{CORE}_D(C)}(D)=\mathrm{POS}_C(D)$，转向(7)；

(4)否则，令 $C'=\mathrm{RED}_D(C)-\mathrm{CORE}_D(C)=\{r_1,r_2,\cdots,r_k\}$，计算 C' 中每个属性的重要度 $SGF(r_k,C,D)=k(C,D)-k(C-\{r_k\},D)$；

(5)将 C' 中的属性按其重要度由高到低排列为：$C'=\{r'_1,r'_2,\cdots,r'_k\}$；

(6)for($i=1;i<=k;i++$)

{

$MinRED_D(C)=MinRED_D(C)+\{r'_i\}$;

if $POS_{MinRED_D(C)}(D)=POS_C(D)$

break;

}

(7)$MinRED_D(C)$即条件属性集 C 的 D 最小简化，结束。

3.3.2 属性值约简

属性值约简是在属性约简的基础上对决策表的进一步简化，去掉决策规则中条件属性的冗余值。属性约简只是在一定程度上去掉了决策表中的冗余属性，但还没有充分去掉决策表中的冗余信息，还需要进一步对决策表进行处理，得到更加简化的决策表。

化简决策表中条件属性的冗余值，必须先计算决策表中条件属性的核值，然后再计算属性值的简化。

3.3.3 一致性判断

根据分辨矩阵的定义，当两个样本发生冲突时，即所有的条件属性取值相同而决策属性值不同，它们所对应的分辨矩阵中元素的取值为空集。因此，在利用简化的分辨矩阵方法求属性的核值过程中，便可知样本数据是否一致。若其中出现空集，则计算决策类的下近似和上近似，然后对下近似集建立确定性粗规则，对上近似集建立可能性粗规则。否则，对一致性数据集只存在确定性规则。

3.3.4 基于遗传算法的规则挖掘

通过粗分析已经为遗传算法提供了一组较为理想的基因，利用遗传算法直接对其进行编码，形成初始种群，从而比较容易得到全局最优的规则。基于遗传算法的规则挖掘一方面通过遗传操作从粗分析获取的确定性和可能性粗规则中提取出确定性规则和可能性规则；另一方面，通过遗传变异可以产生先前没有被发现的新规则。

基于遗传算法的规则挖掘过程：

① 从粗分析的结果中产生初始种群；

② 对初始种群进行编码，形成染色体；

③ 根据评价函数计算各染色体的适应值；

④ 进行遗传操作：

选择运算采用比例选择法；

交叉运算采用单点交叉法；

变异运算采用自适应变异算子；

⑤ 如果满足停机条件，解码退出；否则，采用精英策略进行种群替换，然后转向③。

3.3.5 告警关联规则的提取

染色体解码后仍以决策表的形式存放，表中的每一行即代表一条规则，然后进行规则提取。决策规则的表示形式为：

if X_1 && X_2 &&…&& X_n then Class C (CF) (19)

其中：X_1、X_2、…、X_n——规则前提部分，分别表示非空的条件属性及其值；

Class C——规则的结论部分，表示规则所属的类别(可疑或攻击)；

CF——规则的可信度，对于可能性规则，$0<CF<1$，对于确定性规则，$CF=1$。

3.3.6 后处理

规则后处理主要是利用与、或等布尔算子对产生的确定性规则和可能性规则进行检查、修剪与精简，去除冗余规则，生成通用而简洁的规则，在此不再赘述。

4 基于粗集的遗传分类算法的描述

基于粗集理论的遗传分类算法的数据挖掘过程实现步骤如下：

①将样本数据表达成决策表，确定条件属性和决策属性；

②对样本数据进行选择与转换

③对连续属性值进行离散化处理；

④利用分辨矩阵的简化方法求取条件属性的合取范式；

⑤求得条件属性的核和简化族；

⑥利用基于属性重要度的分辨矩阵简化方法求出条件属性的最小约简；

⑦计算各条件属性的核值，得到决策表的条件属性核值表；

⑧求取各个决策类的最小决策算法，得到决策规则属性值的简化；

⑨对决策表进行一致性检查：若包含冲突的数据样本，用粗分析分别计算决策类的上、下近似；

⑩获取用决策表表示的粗规则；

⑪从粗规则中产生初始种群；

⑫对初始种群进行编码，形成染色体；

⑬计算对各染色体的适应值，对其进行评价；

⑭根据各染色体的适应值大小进行遗传操作，产生新种群；

⑮计算新种群中各染色体的适应值，评价新种群；

⑯若新种群满足终止条件，则转向⑰；否则，采用排序法进行种群替换，然后转向⑬；

⑰染色体解码，形成规则；

⑱规则修剪后输出。

结论 本文简要分析了粗集理论与遗传算法各自的优点，综合它们的各自的优点，提出了一种有效的知识获取方法——基于粗集理论的遗传分类算法。该算法适用于网络安全综合监控系统的告警信息处理，能处理大规模告(预)警数据及其不一致性，获得简洁的告(预)警综合规则。

参 考 文 献

1 Pawlak Z. Rough sets. International Journal of Information and Computer Science, 1982, 11(5): 341～356

2 Pawlak Z. Rough set: theoretical aspects of reasoning about data. Dordrecht, Poland: Kluwer Academic Publishers, 1991

3 Pawlak Z. Rough sets approach to knowledge-based decision support. European Journal of Operational Research, 1997. 48～57

4 Mrozek A. Rough sets and dependency analysis among attributes in computer implementations of expert's inference models. International Journal of Man Machine Studies, 1989, 30: 457～473

5 Ziarko W. Introduction to the special issue on rough sets and knowledge discovery. Computational Intelligence, 1995, 11(2): 223～226

6 Hu X. Knowledge discovery in database: An attribute-oriented rough set approach: [Dissertation]. Canada: University of Regina, 1995

7 Duntsch I, Gediga G. Simple data filtering in rough set systems. International Journal of Approximate Reasoning, 1998, 18: 93～10

8 Wang Jue, Miao Duoqian. Analysis on attribute reduction strategies of Rough set. Journal of Computer Science and Technology, 1998, 13(2): 189～192

9 ANNA M R, ETIENNE E K. A comparative stud)r offuzzy rough setsD]. Fuzzy Sets and System, 2002(126): 137～155

10 Miao Duoqian, Wang Jue. An information-based algorithm for reduction of knowledge. IEEE ICIPS '97, 1997. 1155～1158

11 Wrobleski J. Finding minimal reducts using genetic algorithm. Technical report, Warsaw University, 1995

12 Khoo L P, Zhai Lian-Yin. A prototype genetic algorithm-enhanced rough set-based rule induction system. Computer in Industry, 2001, 4(6): 95 ～106

13 Yeung D S, Shiu S C K, Tsang E C C. Modeling flexible manufacturing systems using weighted fuzzy colored Petri nets. Journal of Intelligent and Fuzzy Systems, 1999, 7(2): 137～149