第2卷第3期
1999年9月

管 理 科 学 学 报
JOURNAL OF MANAGEMENT SCIENCES IN CHINA

Vol. 2 No. 3
Sep. 1999

研究简报

从数据库中发现知识的方法研究与应用[1]

赛 英 陈文伟[2]
(国防科技大学系统工程与数学系)

【摘要】知识发现与数据开采是90年代中期兴起的决策支持新技术. 本文以近年来获得快速发展的粗集方法为基础,从数据库中自动抽取相关属性,发现知识,提出了最小属性集算法以及规则化简算法,从而改进粗集方法使之得到实用的规则,并应用于一个疾病诊断.

关键词:粗集,最小属性集,数据开采,知识发现
分类号:TP311.13

0 引 言

随着计算机技术的高速发展和应用普及,政府、商业、企业等机构每天都在产生并积累着大批量的数据. 如何从这些大规模的数据库中抽取出有用的信息来辅助决策具有重要意义. 目前,从数据库中发现知识(knowledge discovery in database, KDD)成为一个十分活跃的研究领域,它是人工智能、机器学习与数据库技术相结合的产物,被数据库以及机器学习研究者誉为90年代最有前景的研究课题.

数据开采是KDD过程的一个特定步骤,它用专门算法从数据库中抽取模式. 数据开采的方法和技术有很多,比较有影响的包括:ID3方法,粗集方法,概念树方法,覆盖正例排斥反例方法,神经网络方法,遗传算法,公式发现,统计分析方法以及模糊论方法等. 其中粗集方法近年来在智能信息处理研究领域获得了很大发展,它是波兰华沙理工大学Z. Pawlak等科学家提出的. 虽然目前在我国关于粗集理论的研究成果还比较少,但在国外,粗集方法已应用在很多领域,用来处理不完整数据和不精确、不确定性问题. 但是,粗集方法得到的规则往往过多,包含很多无用规则. 本文以粗集方法为基础,提出了求最小属性集算法和规则化简算法,克服了粗集方法的不足,得出精练实用的规则.

1 粗集理论的主要概念

1) 信息系统

四元组$S=(U,A,V,f)$是一个信息系统,其中

U:非空、有限的对象(元组)集合,

$$U=\{x_1,x_2,\cdots,x_n\}.$$

A:属性的有限集合. A中的属性又分为两个不相交的子集,即条件属性集C和决策属性集D,$A=C\cup D$.

$V=\bigcup V_a, \forall a\in A, V_a$是属性$a$的值域.

$f: U\times A\to V$是一个信息函数,它为每个对象的每个属性赋予一个信息值. 即$\forall a\in A, x_i\in U, f(x_i,a)\in V_a$.

信息系统的数据以关系表的形式表示. 关系表的行对应要研究对象,列对应对象的属性. 对象的信息是通过指定对象的各属性值来表达的.

① 国家自然科学基金资助项目(79670019).
② 赛 英:博士生. 研究方向:智能决策支持系统,数据仓库,数据开采等. 通讯地址:湖南长沙国防科技大学系统工程与数学系,邮编410073.
本文1999年1月26日收到.

2) 不可分辨关系

定义1　令$IND \subset A$，定义关于属性集IND的不可分辨关系，$\widehat{IND}$为

$$\widehat{IND} = \{(x_i,x_j)\} \in U \times U : \forall a \in IND, f(x_i,a) = f(x_j,a)\}$$

如果$(x,y) \in \widehat{IND}$则称x和y关于IND是不可分辨的. 容易证明，$\forall IND \subseteq A$，不可分辨关系$\widehat{IND}$是U上的等价关系.

3) 近似空间

信息系统$S = \langle U,A,V,f\rangle$，$IND \subset A$，$\widehat{IND}$是关于$IND$的不可分辨关系，有序对$AS = \langle U, \widehat{IND}\rangle$称作近似空间. 等价关系将$U$划分为等价类集合，每个等价类中的对象关于$IND$是不可分辨的. 这种划分记作$U/\widehat{IND}$，它表示一个等价类族. AS中任意有限个等价类的并集称为一个可定义的集合.

定义2　令$X \subseteq U$，定义X在AS中的下近似集合($\underline{IND}X$)为

$$\underline{IND}X = \bigcup \{Y \in U/\widehat{IND} : Y \subseteq X\}$$

定义3　令$X \subseteq U$，定义X在AS中的上近似集合($\overline{IND}X$)为：

$$\overline{IND}X = \bigcup \{Y \in U/\widehat{IND} : Y \cap X \neq \varnothing\}$$

由定义，$\forall x_i \in \underline{IND}X$，$x_i$一定属于$X$；$\forall x_i \in \overline{IND}X$，$x_i$可能属于$X$. 也就是说，在近似空间$AS$中，$X$的下近似是指$X$所包含的最大可定义集合；而$X$的上近似是指包含$X$的最小可定义集合.

4) 属性集之间的依赖

定义4　$C,D \subset A$是两个属性集，C与D之间依赖程度$\gamma(C,D)$定义为

$$\gamma(C,D) = \frac{|POS(C,D)|}{|U|}$$

其中$POS(C,D) = \bigcup \underline{C}X, X \in U/\tilde{D}$，即$POS(C,D)$表示划分$U/\tilde{D}$中每个等价类$X$关于不可分辨关系$\tilde{C}$的下近似集合($\underline{C}X$)的并集.

$POS(C,D)$包含了U中所有根据属性集C的信息可以准确无误地划分到关系$\tilde{D}$的等价类(决策类)中去的对象. 用$||$表示集中的元素个数. 因此$\gamma(C,D)$就表示了根据C能被准确分类的对象在系统中所占的比例，或称为属性集C区分决策类($U/\tilde{D}$)的能力. 容易证明如下性质：

(1) $\gamma(C,D) \in [0,1]$. 若$\gamma(C,D) = 0$，表示根据C的信息无法将任何对象准确分类；若$\gamma(C,D) = 1$，表示根据C的信息可对U中所有对象准确分类.

(2) 若$C = \varnothing, D \neq \varnothing$，则$POS(C,D) = \varnothing$，$\gamma(C,D) = 0$.

5) 属性的重要度(significance)

定义5　$C,D \subset A, a \in C$，如下定义属性a关于条件属性集C和决策属性集D的重要度：

$$SGF(a,C,D) = \gamma(C,D) - \gamma(C - \{a\},D)$$

$SGF(a,C,D)$表示在分类$U/\tilde{D}$下，C中属性a的重要程度，即C中由于缺少属性a而导致的不能被准确分类的对象在系统中所占的比例. 容易证明如下性质：

(1) $SGF(a,C,D) \in [0,1]$.

(2) 若$SGF(a,C,D) = 0$，表示属性a关于D是可省的.

因为从属性集C中去除属性a后，根据$C - \{a\}$中的信息，原来可被准确分类所有对象仍能准确划分到各决策类中去.

(3) 若$SGF(a,C,D) \neq 0$，表示属性a关于D是不可省的.

因为从属性集C中去除属性a后，某些对象不再能被准确划分.

2　从数据库中发现知识

2.1　最小属性集及求解算法

在信息系统中，根据决策属性将一组对象划分为各不相交的等价类(决策类)，我们希望能通过条件属性来决定每一个类，从而产生每一类的判定规则. 大多数情况下，数据库中存在一些对某个给定的学习任务而言无关的、不重要的属性. 如果能找到一个最小的相关属性集，并且它具有与全部条件属性同样的区分决策属性所划分的类的能力，那么就能去除不相关属性，从而简化关系表，产生的规则集将更简练、更有意义. 基于粗集理论提出了一种求最小属性集的算法.

2.1.1 **最小属性集**(reduct)

定义 6 设 C,D 分别是信息系统 S 的条件属性集和决策属性集.

属性集 $P(P\subseteq C)$ 是 C 的一个最小属性集,当且仅当 $\gamma(P,D)=\gamma(C,D)$,并且 $\forall P'\subset P$, $\gamma(P',D)\neq\gamma(P,D)$.

这说明若 P 是 C 的最小属性集,则 P 具有与 C 同样的区分决策类的能力.换句话说,对任意对象而言,若根据 C 的信息它能被准确分类,则根据 P 中较少的信息它也能被准确分类.并用 P 是最小的.也就是说任意 $P'\subset P$,都不具有与 P 同样的区分决策类的能力.需要注意的是:C 的最小属性集一般是不唯一的.

2.1.2 核(core)

定义 7 核定义为信息系统 S 中所有最小属性集的交集.

容易证明,核是属性重要度(关于条件属性集 C 和决策属性集 D)不为 0 的所有属性的集合,即:$\text{core}=\{a:\text{SGF}(a,C,D)\neq 0,a\in C\}$.

上面提到最小属性集不是唯一的,而要找到所有的最小属性是一个 NP 问题[6].但在大多数应用中,没有必要找到所有的最小属性集.用户可根据不同的原则来选择一个他认为最好的 reduct.比如选择具有最少属性个数的 reduct;或者根据每个属性的实际含义,为其定义一个费用函数,从而选择具有最小费用 reduct;再或者选择具有最小导出关系的 reduct(即去除冗余属性,相同元组合并后导出的关系最小).

我们提出的求最小属性集算法是一个贪心算法,它利用属性重要度和属性间的依赖关系,可在多项式时间内得到一个基于用户的最小的属性集.说它是基于用户的,因为用户可以强制它包含用户所感兴趣的属性,而不管这些属性是否冗余,这样做将更符合实际情况.

2.1.3 求最小属性集算法

core = ∅;

计算 C 中每个属性 a 的重要度 SGF(a,C,D),若不为 0,则 core = core + {a};

对 core 中属性按属性重要度由大到小排序;

用户指定感兴趣属性集 interest;

reduct = core + interest;

B = C − reduct;

while(γ(reduct,D) ≠ γ(C,D)) do

{

计算 B 中每个属性 b 关于 reduct 和 D 的属性重要度 SGF(b,reduct + {b},D);

选择具有最大重要度的属性 a;

若有多个属性的重要度相同,则选择其中与 reduct 中属性导出关系最小的属性 a;

reduct = reduct + {a};

B = B − {a};

}

N = |reduct|;

For i = 1 to N do

if(a, 不在 core + interest 中)

{

reduct = reduct − {a,};

if(γ(reduct,D)| = γ(C,D))

reduct = reduct + {a,};

}

得到最小属性集后,就可以从原关系表中去除多余的属性,并将相同的元组合并,这样能简化关系表而不丢失任何信息.

2.2 获取规则

设 C,D 分别是简化关系表的条件属性集和决策属性集.不可分辨关系 $\check{C}$ 和 $\check{D}$ 对 U 的划分表示为 $U/\check{C}$ 和 $U/\check{D}$.令 E_i 和 X_j 分别代表 $U/\check{C}$ 和 $U/\check{D}$ 中的各个等价类,$\text{Des}(E_i)$ 表示对等价类 E_i 的描述,即等价类 E_i 对应的各条件属性的特定取值;$\text{Des}(X_j)$ 表示对等价类 X_j 的描述,即等价类 X_j 对应的各决策属性的特定取值.则有如下原则:

1) 当 $E_i\cap X_j\neq\varnothing$,则有

$r_{ij}:\text{Des}(E_i)\rightarrow\text{Des}(X_j)$

1° 当 $E_i\cap X_j=E_i$ 时,建立的规则 r_{ij} 是确定的,规则可信度 $cf=1.0$.

2° 当 $E_i\cap X_j\neq E_i$ 时,建立的规则 r_{ij} 是不确定的,规则可信度为

$$cf=\frac{|E_i\cap X_j|}{|E_i|}$$

2) 当 $E_i\cap X_j=\varnothing$ 时,E_i 和 X_j 不能建立规则.

2.3 简化规则及其算法

简化关系表中的元组分为两类,一类是不协

调元组,即两条或多条元组条件相同而结论不同.根据2.2,由这些元组建立的规则的可信度小于1.另一类是协调元组.对应每一协调元组,可直接得到一个可信度为1的分类规则.这样的规则的条件属性得到了一定程度的简化,它只包含最小集中的属性.但是如此产生的规则仍然很多,规则的概括能力差,不够简练.也就是说,对某些规则而言,它们的条件属性中还存在很多冗余,因此这些规则必须进一步简化.基于粗集理论我们提出了一种简化方法.

设简化关系中某协调元组 t 的条件属性集和决策属性集分别为 C 和 D,$E_i \in U/\tilde{C}$,$t \in E_i$,$X_j \in U/\tilde{D}$,$t \in X$.由于 t 是协调的,显然有 $E_i \cap X_j = E_i$,根据2.2节,对元组 t 可以建立可信度为1的规则 r_{ij}:Des(E_i)→Des(X_j).对此规则而言,如果去除其条件属性集 C 中的某些属性后得到的规则的可信度仍为1,则这些属性是冗余的,去除它们可以得到对应的简化规则.由于可得到下述结论:

设 $a \in C$,$C' = C - \{a\}$.$E_i' \in U/\tilde{C}$,$t \in E_i'$,则元组 t 的属性 a 是可以省略的,当且仅当 $E_i' \cap X_j = E_i'$.否则 a 是不可省略的.

定义8　一个元组的所有不可省略的属性的集合称为该元组的核属性集.

用 W 表示元组 t 的核属性集,设 $E_i \in U/\tilde{W}$,$t \in E_i$,$X_j \in U/\tilde{D}$,$t \in X_j$,若 $E_i \cap X_j = E_i$,则核 W 中属性就构成了对应简化规则的条件.否则,若 $E_i \cap X_j \neq E_i$,说明单靠核 W 中的属性不足以构成规则的条件,需要加入其它属性.此时可根据不同的优化规则加入不同的属性,从而得到不同的决策规则,即决策规则的简化并不唯一.下面给出一种算法.

规则简化算法:

设 reduct 是由上面算法求出的最小属性集.

(1) 对于简化关系表中的不协调元组,按2.2节直接写出规则,不再进行简化.

(2) 对于表中的协调元组进行如下操作:

(3) 求表中每个元组的核属性集.W = 核属性集.

(4) 对表中每个元组,执行(5),直到处理完所有元组.

(5) 若元组的核构成其简化规则的条件,则输出该简化规则,转(4);否则转(6).

(6) 令a是reduct中除W中属性外排在最前面的一个属性.W = W + {A}.转(5).

3　从数据库中发现知识的应用举例

在我们研制开发的智能决策服务器中实现了上述算法,并成功地使用它从一个病例数据库中发现了肺炎和肺结核两种疾病的分类规则.由于篇幅所限,在此只列出部分病例作一简单说明.

该病例数据库包含6个特征属性:发烧、咳嗽、X光所见、血沉、听诊音和诊断结果.其中"诊断结果"是决策属性D,其余的属性构成条件属性集C.

病例号	发烧	咳嗽	X光所见	血沉	听诊音	诊断结果
1	高	剧烈	片状	正常	水泡音	肺炎
2	中度	剧烈	片状	正常	水泡音	肺炎
3	低	轻微	点状	正常	干鸣音	肺炎
4	高	中度	片状	正常	水泡音	肺炎
5	中度	轻微	片状	正常	水泡音	肺炎
6	无	轻微	索条状	正常	正常	肺结核
7	高	剧烈	空洞	快	干鸣音	肺结核
8	低	轻微	索条状	正常	正常	肺结核
9	无	轻微	点状	快	干鸣音	肺结核
10	低	中度	片状	快	正常	肺结核

按照2.1节给出的求最小属性集的算法,由于每个属性 a 关于 C 和 D 的重要度 SGF(a,C,D) = 0,所以此例中的核 core = ∅,若用户不提供感兴趣的属性集,此时 rdeuct = ∅,$B = C$.计算 B 中每个属性 b 关于 reduct 和 D 的重要度 SGF(b, reduct + {b},D),可得发烧、咳嗽、X 光所见、血沉、听诊音的重要度分别为:0.4,0,0.3,0.3,0.7.可见听诊音的重要度最高,根据算法,reduct = {听诊音}.此时 γ(reduct,D) = 0.7,由于 γ(C,D) = 1,即 γ(reduct,D) ≠ γ(C,D) = 1,所以还需再计算发烧、咳嗽、X光所见、血沉关于当前的 reduct 及 D 的重要度,分别为0.3,0.1,0.1,0.3.

发烧和血沉的重要度相等，但血沉与 reduct 中的属性（这里为听诊音）导出的关系（5 条元组）比发烧与 reduct 中的属性导出的关系（7 条元组）小，故 reduct = reduct + {血沉} = {听诊音，血沉}，此时 $\gamma(\text{reduct},D) = \gamma(C,D) = 1$，且没有冗余属性，因此该病例数据库的最小属性集 reduct = {听诊音，血沉}. 由此最小属性集导出的关系为：

元组号	听诊音	血沉	诊断
1	水泡音	正常	肺炎
2	干鸣音	正常	肺炎
3	正常	正常	肺结核
4	干鸣音	快	肺结核
5	正常	快	肺结核

可以看出，导出关系表比原有的关系大大简化了，这使用户更容易获得各种简练的规则集. 比如，根据 2.3 节的规则简化算法，得到元组 1、3 的核属性集为{听诊音}，元组 2 的核属性集为{听诊音，血沉}，元组 4、5 的核属性集为{血沉}. 该例中这些元组的核都恰恰构成其简化规则的条件，从而可获得如下的决策规则集：

（1）听诊音 = 水泡音 → 诊断 = 肺炎

（2）听诊音 = 正常 → 诊断 = 肺结核

（3）听诊音 = 干鸣音 ∧ 血沉 = 正常 → 诊断 = 肺炎

（4）血沉 = 快 → 诊断 = 肺结核

4 结束语

本文以粗集方法为基础，研究属性间的依赖程度和属性重要度，提出了求最小属性集的算法和化简规则的算法，使粗集方法在数据库中发现知识具有实用效果，克服了粗集方法中所挖掘的大量规则（包含无用规则）的不足，得到简练、贴切的规则. 这对于有效地从大型数据库发现知识十分有意义.

另外，粗集方法也可以与其它数据开采方法结合起来应用，能更有效地发现知识. 对这一方面的研究工作也很有意义.

参 考 文 献

1 Hu X, Cercone N. Mining knowledge rules from databases: a rough set approach. 1063-6382/96, IEEE 1996

2 Hu X. Knowledge discovery in databases: an attribute-oriented rough set approach. Ph. D Thesis, Dept. of Computer Science, University of Regina, Canada, June 1995

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

4 曾黄麟. 粗集理论及其应用. 重庆：重庆大学出版社, 1998.3

5 陈文伟. 智能决策技术. 北京：电子工业出版社, 1998.6

6 Wojciech Ziarko. The discovery, analysis, and representation of data dependancies in databases. in Knowledge Discovery in Databases G. Piatetsky-Shapiro and W. J. Frawley, (eds) (Menlo Park, CA: AAAI/MIT, 1991. 213 ~ 228

Research and Application for an Approach of Knowledge Discovery in Database

Sai Ying, Chen Wenwei

Department of System Engineering & Mathematics, National University of Defense Technology

Abstract Knowledge mining and discovery is a new technology for decision support system beginning in thd middle of 1990's. In this paper, based on the rough set theory that has been developing very fast in recent years, we put forward an algorithm for finding a minimal attribute set of the database being mined and also another algorithm for simplifying the discovered rules. We improve the rough set method in such a way that it can generate more appropriate rules. In the end, we give an application in disease diagnosis.

Keyword: rough set, minimal attribute set, data mining, knowledge discovery in database