# 一种广义综合知识表示方法及其应用*

吴正龙
(炮兵学院 合肥 230031)
(中国科学院合肥智能机械研究所 合肥 230031)
王儒敬
(中国科学院合肥智能机械研究所 合肥 230031)
邱超凡
(炮兵学院 合肥 230031)

**摘 要** 知识发现技术能发现数据中有用的模式和知识,但如何和专家系统有效集成仍然尚未得到解决.本文从知识表示的角度探讨二者集成,提出一种面向知识发现的广义综合知识表示方法.该方法能有效表示包括关联、分类、序贯、神经网络、基于案例推理等在内的多种知识类型,同时该方法将知识发现方法表示在其内部,从而有利于自动知识获取的实现.在此基础上提出一种新型专家系统原型.该原型从语义、机制和接口三个层次集成知识发现技术,可以有效进行自动知识获取.

**关键词** 专家系统,知识表示,知识发现,集成
**中图法分类号** TP182

## 1 引 言

专家系统是人工智能应用研究最活跃和最广泛的课题之一.自从 1965 年第一个专家系统 DENDRAL 在美国斯坦福大学诞生以来,仅仅经过 20 多年的研究发展,到二十世纪八十年代中期,各种专家系统就已遍布各个领域,取得很大成功.

目前知识获取仍然是开发专家系统的一个瓶颈.在早期,知识获取被视为一个从人类知识到特定知识库的转换过程.这种转换基于以下假设:知识已经清晰存在,只需要搜集并加以表示.这些知识一般是通过对特定领域专家进行咨询得到,并被表示为产生式规则.这种基于知识转换的知识获取具有知识转换困难、表示形式有限等缺点.近来一些研究认为,知识库系统的开发过程更应被视为一个建模过程[1],知识获取不再被视为对知识进行转换,而是成为建模过程的一部分.传统知识转换得到的规则仅仅是知识模型的一种.

数据挖掘技术能从数据库中有效发现包括关联、分类、序贯等多种模式的新知识,近年来,在商业[2]、工业[3]、军事[4]、农业[5]等方面得到迅猛发展,但多数研究偏重于具体挖掘算法,缺乏与智能系统的有效结合.

在数据日益丰富的今天,如何将知识发现技术与专家系统有效集成,从数据中获取知识,实现专家系统的自动知识获取,具有特别重要的理论和实际意义.本文提出一种基于知识建模的广义综合知识表示方法(Knowledge Discovery-Oriented Knowledge Representation, KDOKR),并在此基础上提出一种新型专家系统(Knowledge Discovery-Based Expert System, KDBES),从语义、机制和接口三个层次集成知识发现技术,可以有效进行自动知识获取.

## 2 知识表示方法 KDOKR

### 2.1 面向知识发现的广义综合知识表示方法 KDOKR

文献[6]中提出"知识体·对象块·构件"的知识表示方法.在此基础上,扩展该知识表示方法,提出一种面向知识发现的广义综合知识表示方法

* 国家自然科学基金(No. 69835001)、国家 863 高科技重点(No. 2001AA115170)资助项目
收稿日期:2003-08-08;修回日期:2004-03-29

KDOKR.下面给出其主要部分 BNF 范式,其他部分可参考文献[7].

KDOKR 知识表示方法 BNF 范式如下:

〈知识对象〉::=〈知识对象名〉〈知识类型〉〈推理方法〉〈知识发现方法〉

〈知识类型〉::=〈模型〉|〈广义规则系统〉|

〈模型〉::=〈函数〉|〈CBR〉|〈神经网络〉|〈贝叶斯网络〉|〈回归模型〉

//以下为广义规则系统内容

〈广义规则系统〉::=RS〈广义规则系统名〉+〈广义规则对象块〉+END

〈广义规则系统名〉::=〈字符串〉

〈广义规则对象块〉::=〈规则架〉〈规则体〉

〈规则架〉::=RULE〈规则架名〉IF〈语言变量〉[,〈语言变量〉]+THEN〈语言变量〉[,〈语言变量〉]+

〈规则架名〉::=〈字符串〉|〈整数〉

〈语言变量〉::=〈语言变量名〉〈模糊集表〉//模糊集表为空时,表示该语言变量为清晰变量;

〈模糊集表〉::=〈模糊集〉[,〈模糊集〉]+

〈模糊集〉::=〈模糊集名〉〈隶属函数类型〉〈隶属函数左界〉〈隶属函数右界〉

〈模糊集名〉::=〈字符串〉

〈隶属函数类型〉::=〈三角函数〉|〈左三角函数〉|〈右三角函数〉|〈高斯函数〉|〈Sigmoid 函数〉|〈反 Sigmoid 函数〉|

〈隶属函数左界〉::=〈实数〉

〈隶属函数右界〉::=〈实数〉

〈规则体〉::=RB{〈规则组〉|〈计算对象块〉|〈外部知识对象块〉}

〈规则组〉::=〈广义规则〉[,〈广义规则〉]+

〈广义规则〉::=IF〈前提规则件〉[,〈前提规则件〉]+THEN〈结论规则件〉[,〈结论规则件〉]+WITH〈支持度〉〈置信度〉〈权重〉

……

//广义规则系统内容结束

〈推理方法〉::=〈广义规则系统推理〉|〈模型计算〉|

〈知识发现方法〉::=〈规则挖掘〉|〈数据建模〉|

〈规则挖掘〉::=〈规则挖掘算法〉〈算法参数表〉〈输入〉〈输出〉

〈规则挖掘算法〉::=〈关联规则挖掘算法〉|〈分类规则挖掘算法〉|〈序贯模式挖掘算法〉|

〈数据建模〉::=〈模型结构〉〈输入〉〈输出〉

……

可以看出,知识发现方法被表达在知识表示方法中,广义规则可以表示知识发现得到的多种知识类型,如关联规则、分类规则、序贯模式等.挖掘出的规则需要被转换成广义规则的形式.关联规则和序贯模式通常都带有支持度和置信度,而分类规则的支持度和置信度将被赋予 1.规则的权重由用户确定.

### 2.2　知识推理方法

采用广义规则之后,KDOKR 中主要有两种知识类型,一种是模型,一种是广义规则系统.模型的推理方法基本上是模型计算.下面重点介绍广义规则系统的推理方法.

#### 2.2.1　规则激活程度

设广义规则系统有 $m$ 条广义规则,对于第 $i$ 条规则中第 $j$ 个前提规则件 $RF_{ij}$(Rule Factor)中语言变量值为 $LV$(Language Variable),则

(1) 其模糊集表 $FSL$(Fuzzy Set List)为空,语言变量为清晰变量时,如虫害名称.该前提规则件的语言变量值为 $LVV$(Language Variable Value),$LVV$ 为语言值,如三化螟等.此时前提规则件中的关系符 $RO$(Relation Operand)只能为 = 或是 ≠.设对应的输入 $Input_{ij}$,则该前提规则的激活度 $AD_{ij}$(Activate Degree)计算为

$$AD_{ij}(Input_{ij}) = \begin{cases} 1, & Input_{ij} = LVV \\ 0, & Input_{ij} \neq LVV \end{cases}$$

*when* $RO =$ "=",

或

$$AD_{ij}(Input_{ij}) = \begin{cases} 1, & Input_{ij} \neq LVV \\ 0, & Input_{ij} = LVV \end{cases}$$

*when* $RO =$ "≠".

(2) 其模糊集表 $FSL$ 不为空,语言变量为模糊变量时,如气温.

① 当 $LVV$ 为数值时,此时前提规则件是一种比较式,如气温 ≥ 20℃.

a.当输入 $Input_{ij}$ 为数值时,前提规则件激活度计算为

$$AD_{ij}(Input_{ij}) = \begin{cases} 1, & \text{比较式成立,} \\ 0, & \text{比较式不成立.} \end{cases}$$

b.当输入 $Input_{ij}$ 为语言值时,如气温为高.

设输入语言值的隶属度函数为 $\mu_{Input}$,其左界为 $x_L$,右界为 $x_R$,则前提规则件的激活程度为比较式区间和左右界区间交集长度与左右界区间长度的比值.

如对于前提规则件"气温 >= 20℃",输入为"气温为高",模糊集"气温高"的左界为 15℃,右界为 35℃,则该前提规则件的激活程度为区间[20℃,+∞]和区间[15℃,35℃]的交集长度与区间[15℃,35℃]长度比值为(35 − 20)/(35 − 15) = 0.75.

② 当 $LVV$ 为语言值时,如高、中、低等,此时 $RO_{ij}$ 只能为 = 或是 ≠.

a.当输入 $Input_{ij}$ 为数值时,如 20℃,则该前提

规则件的隶属度计算为

$AD_{ij}(Input_{ij}) = \mu_{FS(k)}(Input_{ij})$,

where $RO$ = "=",

或

$AD_{ij}(Input_{ij}) = 1 - \mu_{FS(k)}(Input_{ij})$,

where $RO$ = "≠",

where $\mu_{FS(k)} \in FSL$ and $FSN_k = LVV$,

其中 $\mu_{FS(k)}$ 表示模糊集 $FS(k)$ 的隶属函数,$FSN_k$ 表示第 $k$ 个模糊集的模糊集名,如高、中、低等.

b. 当输入 $Input_{ij}$ 为语言值时,前提规则件的隶属度计算同(1).

对于输入,所有规则都被激活,其中有 $n$ 个前提规则件的第 $i$ 条规则被激活的程度为

$$AD_i = \prod_{j=1}^{n} AD_{ij}(Input_{ij}).$$

### 2.2.2 计算广义规则系统输出

在计算广义规则系统输出时,可能需要转换结论规则件形式,以便能够进行计算. 如果结论规则件中的语言变量值为语言值,则用顺序整数代替. 如对语言变量"雌虫密度"的结论规则件,其语言变量值分别为低、中、高,则可以用 1、2、3 来代替. 对于有 $m$ 条广义规则的系统,其第 $j$ 个输入 $Output_j$ 计算为

$$Output_j = \frac{\sum_{i=1}^{m} \omega_i * AD_i * LVV_{ij}}{\sum_{i=1}^{m} \omega_i * AD_i},$$

上式中,$LVV_{ij}$ 表示第 $i$ 条广义规则的第 $j$ 个结论规则件的语言变量值. $Sup_i$ 和 $Conf_i$ 分别为第 $i$ 条广义规则的支持度和置信度. 计算得到的输出为实数,可能需要归整转换为语言值. 如对于上述的雌虫密度语言变量来说,计算输出为 1.4,将被转换为低;而 1.6 将被转换为中.

### 2.2.3 输出的不精确性

专家系统需要对非精确的数据和知识进行"非精确处理". 规则的支持度、置信度等客观性度量反映了规则的不精确性. 在广义规则系统的推理方法中,采用基于确定性理论的非精确推理方法,并采用规则支持度和置信度来综合度量规则的不精确性.

(1) 不精确性描述

定义广义规则不精确性的度量 $CF(R)$ 为

$CF(R) = Sup(R) * Conf(R)$.

(2) 证据(输入)的不精确性

用户提供的证据可能是不精确的,如 $CF$(气温为高) = 0.60,同时由于一条广义规则的结论作为另一条广义规则的前提,可能带来了不精确性. 记证据(输入) $Input$ 的不精确性度量为 $CF(Input)$,且有 $CF(\sim Input) = \sim CF(Input)$.

对于由多个证据 $Input_1$、$Input_2$、…、$Input_k$ 逻辑与而成的输入 $Input$,其 $CF(Input)$ 为

$$CF(Input) = \min\{CF(Input_1), CF(Input_2), \cdots, CF(Input_k)\}.$$

而对于由多个证据 $Input_1$、$Input_2$、…、$Input_k$ 逻辑或而成的输入 $Input$,其 $CF(Input)$ 为

$$CF(Input) = \max\{CF(Input_1), CF(Input_2), \cdots, CF(Input_k)\}.$$

(3) 不精确推理

已知广义规则

$R$: IF $Input$ THEN $Output$ $CF(R)$ $CF(Input)$,则

$CF(Output) = CF(R) * \max\{0, CF(Input)\}$.

而对于多条广义规则推理,已知广义规则 $R1$: IF $Input1$ THEN $Output$ $CF(R1)CF(Input1)$ 和广义规则 $R2$: IF $Input2$ THEN $Output$ $CF(R2)CF(Input2)$,则有

$CF_{R1}(Output) = CF(R1) * \max\{0, CF(Input1)\}$,

$CF_{R2}(Output) = CF(R2) * \max\{0, CF(Input2)\}$,

然后计算如下 $CF(Output)$:

① $CF(Output) = CF_{R1}(Output) + CF_{R2}(Output) - CF_{R1}(Output) * CF_{R2}(Output)$, 当 $CF_{R1}(Output) \geqslant 0$, $CF_{R2}(Output) \geqslant 0$;

② $CF(Output) = CF_{R1}(Output) + CF_{R2}(Output) + CF_{R1}(Output) * CF_{R2}(Output)$, 当 $CF_{R1}(Output) < 0$, $CF_{R2}(Output) < 0$;

③ $CF(Output) = CF_{R1}(Output) + CF_{R2}(Output)$, 其他情况.

KDOKR 知识表示方法有以下优点:

(1)知识对象不仅包括知识类型、推理方法,而且指定了知识发现方法. 与传统基于知识转换的规则获取不同,KDOKR 可以从数据中获取知识.

(2)广义规则有效表示了关联规则、分类规则、序贯模式、模糊规则等多种规则型知识,这使得由知识发现得到的规则可以直接以广义规则的形式存入到规则组中,能够很好支持知识发现与专家系统的集成.

(3)广义规则可以描述模糊信息,广义规则系统也是一种模糊规则系统. 如果所有语言变量为清晰变量,则广义系统简化为产生式规则系统.

(4)知识推理中采用知识发现里定义的规则支持度和置信度来表示不精确推理中的规则可信度度量,从而使得 KDOKR 能够较好地支持不精确推理.

# 3　一类新型专家系统原型 KDBES

## 3.1　KDBES 体系结构

基于 KDOKR 知识表示方法,我们提出一类新型专家系统原型——KDBES. KDBES 是在传统专家系统的基础上集成了知识发现而成的,其系统框架如图 1 所示.

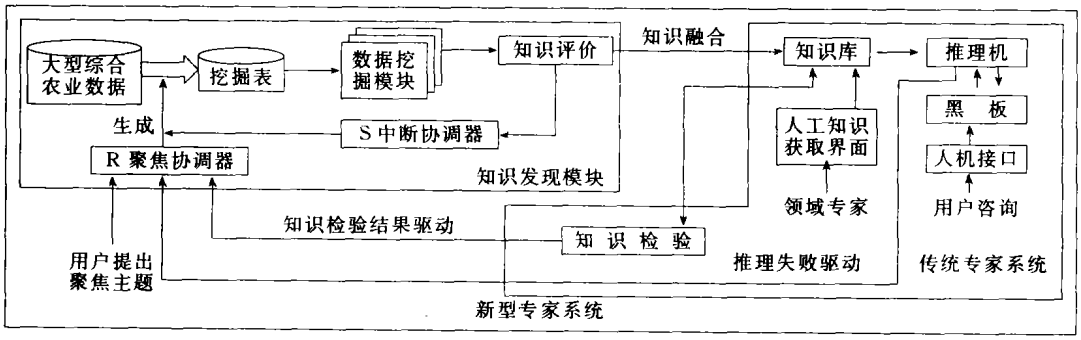


图 1　新型专家系统结构框图

## 3.2　知识发现时机

知识获取目前不可能(即使可能也不应该)完全依赖于知识发现. 领域专家的重要性在任何时候都不应该弱化. 但是在以下时机,可能需要知识发现.

(1) 构建知识库的开始. 在建造专家系统知识库时,没有领域专家,只有知识工程师和面向该领域的大量原始数据,这时知识工程师可以通过知识发现过程来得到该领域知识的体系结构.

(2) 某些特殊应用领域. 在这些领域中,难以找到能够概括归纳知识的领域专家,同时已有大量的原始数据可供使用,这时知识获取只能借助于知识发现.

(3) 知识检验过程之后. 知识检测过程发现知识库中知识的矛盾、冗余、遗漏和蕴涵等不一致现象,由不一致现象聚焦主题,进行知识发现.

(4) 推理决策过程中. 推理过程因为知识的短缺而失败,但领域专家也无法提供更加详细和充分的知识,这时可以由短缺知识(根据推理中断处得到)聚焦主题,进行知识发现.

## 3.3　知识发现的集成方式

集成的基本目标是要使专家系统和知识发现成为一个整体. 可以从三个层面来理解二者之间的集成:(1)语义集成. 从最深的层次来说,二者应该在语义上成为一个整体;(2)机制集成. 在较高层次上,系统应该以某种软件内在机制上的一致性达成集成,并且,这种一致性程度越高,系统集成就越平滑,如专家系统和知识发现系统之间的控制机制和通信机制;(3)接口集成. 最高层的集成含义是在专家系统与知识发现系统之间制定某种协议,建立接口,这种集成可以不用考虑各自的内部特点.

由以上讨论,我们通过广义综合知识表示方法 KDOKR,在语义上将知识发现的各种方法集成到知识表示中. 机制集成中采用基于三种驱动的集成机制,即用户提出聚焦主题驱动、知识检验结果驱动、以及专家系统推理运行失败驱动,有针对性地进行主题式数据挖掘. 接口集成中采用软插件技术(OLE/ActiveX 技术),将各种数据挖掘(建模)算法封装成符合 COM 规范的、具有一组接口(包括功能描述、外接消息和相应说明信息)的软构件,达到知识发现在专家系统中即插即用的目的.

## 3.4　R、S 协调器

R 聚焦协调器根据领域专家或用户提供的元知识、专家系统知识库的一致性和有效性检验结果、专家系统推理运行失败的断点对数据源进行聚焦,从而得到挖掘数据表. 例如元知识"发现平均气温和日照时数与雌虫密度之间的关系",则 R 协调器将从虫害发生数据库中抽取平均气温、日照时数、雌虫密度三个属性的数据,生成挖掘表. 元知识类似于一个指定架构的挖掘任务,而对于"平均气温和雌虫密度之间缺乏关联"的知识库检验结果,R 协调器将生成由平均气温和雌虫密度构成的挖掘数据表. 对于"知识对象[病害诊断]推理失败"的推理运行失败断点,R 协调器将抽取所有在知识对象[病害诊断]中出现的属性数据,构成挖掘表. R 协调器使得数据挖掘过程更具针对性.

数据挖掘得到的规则性知识首先进行知识评价,如果知识已经存在于知识库中(随着系统运行时间增加,这种情况最终会发生),则 S 中断协调器发生作用,中断后续的知识融合过程,回到数据挖掘的开始,此时可能需要选择新的数据进行挖掘.

### 3.5 知识检验

经过知识评价的知识加入到专家系统知识库中，之后要进行知识检验，包括规则的重复、冗余、从属检验.

记规则为 $R$，规则件为 $RF$，规则件集为 $RFS$，前提规则件集与结论规则件集为 $RFS\_P$，$RFS\_C$. 对于两个规则件集 $RFS1$，$RFS2$，存在下列关系：

(1) 如果 $\forall RF \in RFS1$，有 $RF \in RFS2$，则称 $RFS1 \subseteq RFS2$.

(2) 如果 $\forall RF \in RFS2$，有 $RF \in RFS1$，则称 $RFS1 \supseteq RFS2$.

(3) 如果 $RFS1 \subseteq RFS2$ 且 $RFS1 \supseteq RFS2$ 同时成立，则称 $RFS1 = RFS2$.

(4) 如果 $\exists RF \in RFS1$，有 $RF \notin RFS2$，且 $\exists RF \in RFS2$，有 $RF \notin RFS1$，则称 $RFS1 \neq RFS2$.

对于规则 $R1$，$R2$，$R3$ 有

(1) 重复. 如果有 $RFS\_P(R1) = RFS\_P(R2)$，$RFS\_C(R1) = RFS\_C(R2)$，且有 $Sup(R1) \geqslant Sup(R2)$，$Conf(R1) \geqslant Conf(R2)$，则称规则 $R1$ 和 $R2$ 存在重复，且 $R1$ 包含 $R2$.

(2) 冗余. 如果 $RFS\_P(R1) = RFS\_P(R2)$，$RFS\_C(R2) = RFS\_P(R3)$，$FRS\_C(R3) = RFS\_C(R1)$，且有 $Sup(R1) \geqslant Max(Sup(R2), Sup(R3))$，$Conf(R1) \geqslant Max(Conf(R2), Conf(R3))$，则称规则 $R1$、$R2$、$R3$ 之间存在冗余，且 $R2$、$R3$ 为冗余规则. 去除冗余规则可以简化知识，但有可能降低规则的可理解性.

(3) 从属. 如果 $RFS\_P(R1) \subseteq RFS\_P(R2)$，$RFS\_C(R1) = RFS\_C(R2)$，且有 $Sup(R1) \geqslant Sup(R2)$，$Conf(R1) \geqslant Conf(R2)$，则称规则 $R1$、$R2$ 之间存在从属，$R2$ 从属于 $R1$.

(4) 矛盾. 如果 $RFS\_P(R1) = RFS\_C(R2)$，$RFS\_P(R2) = RFS\_C(R1)$，则称 $R1$、$R2$ 之间存在矛盾. 矛盾的知识由用户决定处理，也可以作为知识检验聚焦主题.

规则型的知识可以以广义规则的形式直接放进规则组中，进行知识融合，而对于模型来说，融合意味着新模型取代旧模型. 用户在融合过程中发挥主观作用，他们可以在客观准则之外，最终决定哪些知识将被保留，哪些知识将被删除. 基于超图的技术被用来发现知识的重复、冗余和从属.

## 4 结论与展望

本文提出面向知识发现的广义综合知识表示方法 KD()KR，能够有效表示多种知识，是一种复合型知识表示方法，同时将知识发现方法表示在知识表示方法中，能够有效支持知识的自动获取. 在此基础上，本文系统研究了知识发现与专家系统的语义、机制和接口集成方法，提出了一种新型专家系统 KD-BES. 它基于三种驱动调用的知识发现进程，有针对性的(基于主题的)进行知识发现，并将发现的知识融合到专家系统知识库中，一定程度上实现了自动知识获取.

知识决定着专家系统的能力，而知识获取却历来是建造专家系统的瓶颈. 知识发现给专家系统的自动知识获取带来希望，二者之间有效集成最终决定知识发现是否能够真正改善专家系统自动知识获取的能力. 知识发现自身也强调所发现知识的价值存在于它的适当使用中. 将知识发现应用到专家系统中，将大大改善专家系统的知识获取能力，从而提高专家系统的决策能力，同时也将促进知识发现的更广阔应用.

## 参 考 文 献


[1] Su Limin, Zhang Hong, Hou Chaozhen, Pan Xiuqin. Research on an Improved Genetic Algorithm Based Knowledge Acquisition. In: Proc of the 2002 International Conference on Machine Learning and Cybernetics. Beijing, China, 2002, Ⅰ: 455－458

[2] Fong S, Chan S. Mining Online Users' Access Records for Web Business Intelligence. In: Proc of the IEEE International Conference on Data Mining. Maebashi City, Japan, 2002, 759－762

[3] Braha D, Shmilovici A. Data Mining for Improving a Cleaning Process in the Semiconductor Industry. IEEE Trans on Semiconductor Manufacturing, 2002, 15(1): 91－101

[4] Ceruti M G, McCarthy S J. Establishing a Data-Mining Environment for Wartime Event Prediction with an Object-Oriented Command and Control Database. In: Proc of the 3rd IEEE International Symposium on Object-Oriented Real-Time Distributed Computing. California, USA, 2000, 174－179

[5] Seek W L, Kerschberg L. A Methodology and Life Cycle Model for Data Mining and Knowledge Discovery in Precision Agriculture. In: Proc of the IEEE International Conference on Systems, Man, and Cybernetics. San Diego, California, USA, 1998, Ⅲ: 2882－2887

[6] 淮晓永，熊范纶，张 琳，王儒敬，赵 星. 一个智能型智能信息系统开发环境——Visual XF7.0. 见：863 计算智能计算机主题学术会议. 北京，2001，34－41

[7] 吴正龙. 基于数据库和知识库的知识发现及其应用研究. 博士学位论文，中国科学技术大学，合肥，2003

# A GENERALIZED KNOWLEDGE REPRESENTATION METHOD AND ITS APPLICATION

Wu Zhenglong
(*Artillery Academy*, *Hefei* 230031)
(*Institute of Intelligent Machines*, *Chinese Academy of Sciences*, *Hefei* 230031)
Wang Rujing
(*Institute of Intelligent Machines*, *Chinese Academy of Sciences*, *Hefei* 230031)
Qiu Chaofan
(*Artillery Academy*, *Hefei* 230031)

## ABSTRACT

Knowledge discovery from database (KDD) technology can be used to uncover important patterns and knowledge embedded in data. How to use KDD technology to facilitate knowledge acquisition of expert systems still remains difficult. From the point of view of knowledge representation, the integration between KDD technology and expert system is discussed in this paper. A generalized knowledge representation method is presented. Based on knowledge modeling, the method is effective to represent a variety of knowledge including associations, classification, sequential pattern, neural network, case-based reasoning, etc. KDD technology is integrated into this method so that automatic knowledge acquisition becomes possible. A new framework of expert system is proposed then. Integrating KDD technology into expert system at three levels: semantics level, mechanism level and interface level, this new expert system can acquire knowledge from data effectively.

**Key Words** Expert System, Knowledge Representation, Knowledge Discovery from Database, Integration