第19卷 第6期
1993 年11 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 19, No. 6
Nov., 1993

知识发现的理论及其实现[1)]

洪 家 荣

(哈尔滨工业大学计算机系，哈尔滨 150006)

摘 要

本文提出知识发现的一种理论，该理论基于对人类知识发现认识过程的模拟，包括经验数据分类、各类数据的概念抽象、及概念间蕴涵关系的发现等步骤. 文中介绍了实现这个理论的一个集成化学习系统 KD_3，以及它在自动建立专家系统知识库等方面的应用.

关键词：机器学习，知识发现，概念聚类，示例学习，概念蕴涵，集成化学习.

一、引 言

机器发现（machine discovery）是在没有外界帮助下从已知经验数据中自动发现蕴含在其中的关系和规律. 随着科学技术的发展，积累的数据越来越多，机器发现的作用也就越来越显得重要. 为了适应专家系统知识获取的需要，近年来机器发现的一个新分支——知识发现（knowledge discovery）迅速发展起来. 知识发现是指从符号的经验数据中发现有用知识，一般是产生式规则. 当前国际上已出现一些知识发现系统，如 CHARADE[1], IRULE[2] 等. 然而，当前的知识发现方法尚缺乏完整的理论. 本文通过对人类知识发现过程的分析，根据认识论的基本原理，提出知识发现的一个理论框架，并用一个集成化学习系统 KD_3 来实现这种理论. 最后，本文给出 KD_3 应用于自动建造专家系统知识库的例子.

二、知识发现的原理

1. 知识发现的计算模型

根据认识论，人类的知识属于理性认识，它是由感性认识(经验)发展而来. 因此，人类要计算机能够发现知识，就必须让它模拟人类感性认识向理性认识不断飞跃的辩证运动[3]. 现在虽然这一辩证运动的生理机制尚不清楚，但认识论与心理学都揭示出人类概念的形成往往经历对经验的分析与分类、抽象与概括的过程. 由于概念是抽象的，因此概念之间的关系只能由它们外延间的关系来决定.

在本文提出的知识发现模型中，经验由已知数据(例子的集合)组成，概念的外延是数

本文于 1991 年 3 月 6 日收到.
1) 国家自然科学基金和国家 863 资助项目.

据的一些子集，相应的内涵是满足它的外延并排除其余概念的外延的描述；而概念的外延是通过对已知数据的聚类得到，相应的内涵则通过示例学习获得.

2. 相关概念

设 E 是一个 n 维有穷离散属性的向量空间，即 $E = D_1 \times \cdots \times D_n$，其中 D_i 是符号或整数的有穷集合.

定义 1. E 中的一个元素 e 叫做例子，即 $e = \langle v_1, v_2, \cdots, v_n \rangle, v_i \in D_i, j = 1, 2, \cdots, n$；形为 $[x_i = W_i]$ 的关系语句叫选择子，$W_i \subseteq D_i$；一个选择子或几个选择子的合取($\wedge$，有时省略)式叫做公式；一个公式或几个公式的析取($\vee$)式叫做概念描述，简称描述.

定义 2. 已知一选择子 $[x_i = W_i]$ 与一个例子 $e = \langle v_1, \cdots, v_n \rangle$，如果 $v_i \in W_i$，则称该选择子覆盖例子 e，或称例子 e 满足该选择子；如果一个公式中的所有选择子都覆盖例子 e，则称该公式覆盖 e；如果一个描述中有一个公式(析取项)覆盖 e，则称该描述覆盖 e.

设 S_1 与 S_2 是两个例子集，$S_1 \subseteq E$，$S_2 \subseteq E$，分别叫做正例集与反例集.

定义 3. 一个选择子(公式或描述)是 S_1 对于 S_2 的一致覆盖，如果它覆盖 S_1 中的每个例子(叫正例)而不覆盖 S_2 中的任何例子(叫反例).

定义 4. 一个示例学习问题是一个五元组 (A, B, S_1, S_2, C)，其中 A 是学习算法，B 是学习偏向 (Bias)，S_1 是正例集，S_2 是反例集，C 是产生的描述. 学习偏向通常取做最优化准则，如最优覆盖或最简公式. 在学习算法与学习偏向确定以后，学习的结果 C 可用函数 $\mathrm{Cover}(S_1, S_2)$ 表示.

注意 1. 如果 $S_2 = E - S_1$，则 $\mathrm{Cover}(S_1, S_2)$ 覆盖的例子集正好是 S_1，因此该学习问题变为精确覆盖问题，即演绎推理；如果 $S_2 \subset E - S_1$，则 $\mathrm{Cover}(S_1, S_2)$ 除了覆盖 S_1 外，还可能覆盖 E 中除在 S_1 及 S_2 中以外的元素，因此该学习问题是一种归纳推理. 归纳推理是学习算法的一种概括能力，正是这种概括能力使学习得到的描述有更普遍的适用性.

定义 5. 一个描述 F 在集合 S 中的外延是 S 中所有满足 F 的元素之集，记为 $\mathrm{Extension}(F, S)$.

引理 1. 设 $F = F_1 \wedge F_2$，$S = S_1 \cup S_2$，则

(1) $\mathrm{Extension}(F, S) = \mathrm{Extension}(F_1, S) \cap \mathrm{Extension}(F_2, S)$;

(2) $\mathrm{Extension}(F, S) = \mathrm{Extension}(F, S_1) \cup \mathrm{Extension}(F, S_2)$.

证明. 由外延的定义直接推出.

设 $S_1 \subseteq S$，记 $F(S_1, S) = \mathrm{Cover}(S_1, S - S_1)$.

引理 2. 如果 $S_1 \subseteq S$，则 $F(S_1, S)$ 在 S 中的外延是 S_1，即 $\mathrm{Extension}(F(S_1, S), S) = S_1$.

证明. 由注意 1 知，如果 $S = E$，则 $F(S_1, S) = \mathrm{Cover}(S_1, S - S_1)$ 恰好覆盖 S_1；如果 $S \subset E$，则 $F(S_1, S)$ 除了覆盖 S_1 外，还可能覆盖 E-S 中的其他元素，但这些元素已不在 S 中.

定义 6. 如果描述 F_1 在 S 中的外延是描述 F_2 在 S 中的外延的子集，即 Extension

$(F_1,S)\subseteq$Extension(F_2,S)，则称 F_1 在 S 中(逻辑地)蕴涵 F_2，记为 $F_1 \underset{S}{\Rightarrow} F_2$，该表达式叫做决策规则，简称规则，$F_1$ 叫前提，F_2 叫结论．如果 $S=E$，则 $\underset{S}{\Rightarrow}$ 的 S 可缺省．F_1 叫做同 F_2 在 S 中(逻辑地)等价，记为 $F_1 \underset{S}{\Longleftrightarrow} F_2$，如果 $(F_1 \underset{S}{\Rightarrow} F_2)\wedge(F_2 \underset{S}{\Rightarrow} F_1)$．

关于满足什么条件两个描述才能构成规则，有如下引理．

引理 3. 如果 $S_2\subseteq S_1\subseteq S$，则

1) $F(S_2,S) \underset{S}{\Rightarrow} F(S_1,S)$；

2) $F(S_2,S) \underset{S}{\Longleftrightarrow} F(S_2,S_1)\wedge F(S_1,S)$．

证明．1) 因为 Extension$(F(S_2,S),S)=S_2$，Extension$(F(S_1,S),S)=S_1$，以及 $S_2\subseteq S_1$，得证．

2) Extension$(F(S_2,S_1)\wedge F(S_1,S),S)$ = Extension$(F(S_2,S_1),S)\cap$Extension$(F(S_1,S),S)$ = {Extension$(F(S_2,S_1),S_1)\cup$Extension$(F(S_2,S_1),S-S_1)\}\cap S_1 = S_2\cap S_1\cup$ Extension$(F(S_2,S_1),S-S_1)\cap S_1 = S_2\cup\varnothing = S_2$ = Extension$(F(S_2,S),S)$．

3. 知识发现算法

本文提出的知识发现算法由三个步骤组成，即例子分类、描述抽象和规则形成．

1) 例子分类．由两种方法对例子集合分类：

层次分类．将已知例子集 S 依次分为一个树状的层次结构，$\{S\}$，$\{S_1,\cdots,S_k\},\cdots,\{S_{11},\cdots,S_{1k}\},\cdots\{S_{k1},\cdots,S_{kk}\},\cdots$．其中 S_{ij} 是 S_i 的子集．

平行分类．令 $k=2,3,\cdots,r$，分别对同一个例子集 S 分类成 k 个子集，形成 r 个族，$\{S\},\{S_1^{(2)},S_2^{(2)}\},\cdots,\{S_1^{(r)},\cdots,S_r^{(r)}\}$．

2) 描述抽象．产生每个子集在整个例子集 S 中的描述．即

对层次分类，产生 $\{F_1,\cdots,F_k\},\cdots,\{F_{11},\cdots,F_{1k}\},\cdots,\{F_{k1},\cdots,F_{kk}\}$．其中 $F_i=$Cover$(S_i,S-S_i),\cdots,F_{ij}=$Cover$(S_{ij},S-S_{ij})$．

对平行分类，产生 $\{F_1^{(2)},F_2^{(2)}\},\cdots,\{F_1^{(r)},\cdots,F_r^{(r)}\}$．其中 $F_i^{(l)}=$Cover$(S_i^{(l)},S-S_i^{(l)})$，$i=1,2,\cdots,l$，及 $l=2,\cdots,r$．

3) 规则形成．先在第 1) 步产生的集合中寻找子集关系，然后在第 2) 步产生的相应的描述间构造决策规则．

下列定理保证知识发现算法的正确性．

定理 1. 1) 假定 S_{ij} 和 S_i 分别由层次分类得到，F_{ij} 和 F_i 是相应的描述．则 $F_{ij} \underset{S}{\Rightarrow} F_i$．

2) 假定 $S_i^{(l)}$ 和 $S_j^{(m)}$ 是由平行分类产生，$F_i^{(l)}$ 和 $F_j^{(m)}$ 是相应的描述．则如果 $S_i^{(l)}\subseteq S_j^{(m)}$ 则 $F_i^{(l)} \underset{S}{\Rightarrow} F_j^{(m)}$．

证明．由引理 3 直接得出．

三、知识发现算法的实现

知识发现由一个集成化学习系统 KD_3 实现．KD_3 主要包括概念聚类系统 LEOBS、

示例学习系统 GS[4] 和集合运算.

1. 概念聚类系统 LEOBS.

LEOBS 是国际上著名的概念聚类系统 CLUSTER/2[5] 的改进版本，主要用于事例分类. 概念聚类较之普通聚类分析更能产生有意义的描述.

2. 示例学习系统 GS.

GS 是一种新的示例学习算法的实现，主要用于描述获取. 当前国际上有名的示例学习系统有 ID_3[6] 和 AQ_{15}[7]. 但由于 ID_3 产生的决策树结构不适用于规则发现，而 AQ_{15} 尚不能处理属性值的缺省，因而 KD_3 采用 GS 算法.

GS 的核心算法是: 选择一个覆盖最多正例的选择子，并测试该选择子是否覆盖了反例，如果没有则一个公式生成；否则将被该选择子覆盖的正、反例子取出做成新的正、反例集，然后生成新的选择子并逻辑乘到先前生成的公式上. 重复这一过程直到一个公式生成为止. 当一个公式生成后，将正例集中被该公式覆盖的所有正例删除，并对剩下的正例集重复上述过程，直到正例集变空为止.

3. 集合运算. 它是寻找各类间的子集关系，然后构造相应描述间的逻辑蕴涵式，即产生式.

KD_3 具有层次法规则发现和平行法规则发现能力，前者使用层次分类，后者使用平行分类.

四、应 用 举 例

知识发现系统 KD_3 可应用于从实际数据库中发现有用知识. 例如，使用 KD_3 已从哈尔滨地区机动车司机体检资料中发现了司机心电图异常的原因[1)]，从数据中发现专家系统知识库中的知识[8]. 现在举例说明 KD_3 是如何发现规则的.

表 1 动物的例子

编号	属性 / 动物	毛发	牙齿	眼睛	羽毛	脚	食物	奶	会飞	产蛋	会游泳
1	虎	有	犬齿	前方	无	有爪	肉	有	不	不	是
2	豹	有	犬齿	前方	无	有爪	肉	有	不	不	是
3	长颈鹿	有	钝	旁边	无	蹄	草	有	不	不	是
4	斑 马	有	钝	旁边	无	蹄	草	有	不	不	是
5	鸵 鸟	无	无	旁边	有	有爪	谷	无	不	是	不
6	企 鹅	无	无	旁边	有	蹼	鱼	无	不	是	是
7	信天翁	无	无	旁边	有	有爪	谷	无	是	是	不
8	鹰	无	无	前边	有	有爪	肉	无	是	是	不
9	毒 蛇	无	犬齿	旁边	无	无	肉	无	不	是	不
10	蜜 蜂	无	无	旁边	无	*	蜜	无	是	*	不

注: 其中*表示缺省.

1) 王岩、洪家荣、孙希文、刘宁、孟淑贤，一个基于属性的学习发现系统 ABLD 及其应用，第三届全国青年计算机学术会议，长沙，1991.

已知一些动物的例子，如表 1 所示。

1. 层次规则发现

图 1 是层次规则发现产生的层次结构——决策树。花括号中的数字是在该节点中的例子编号。

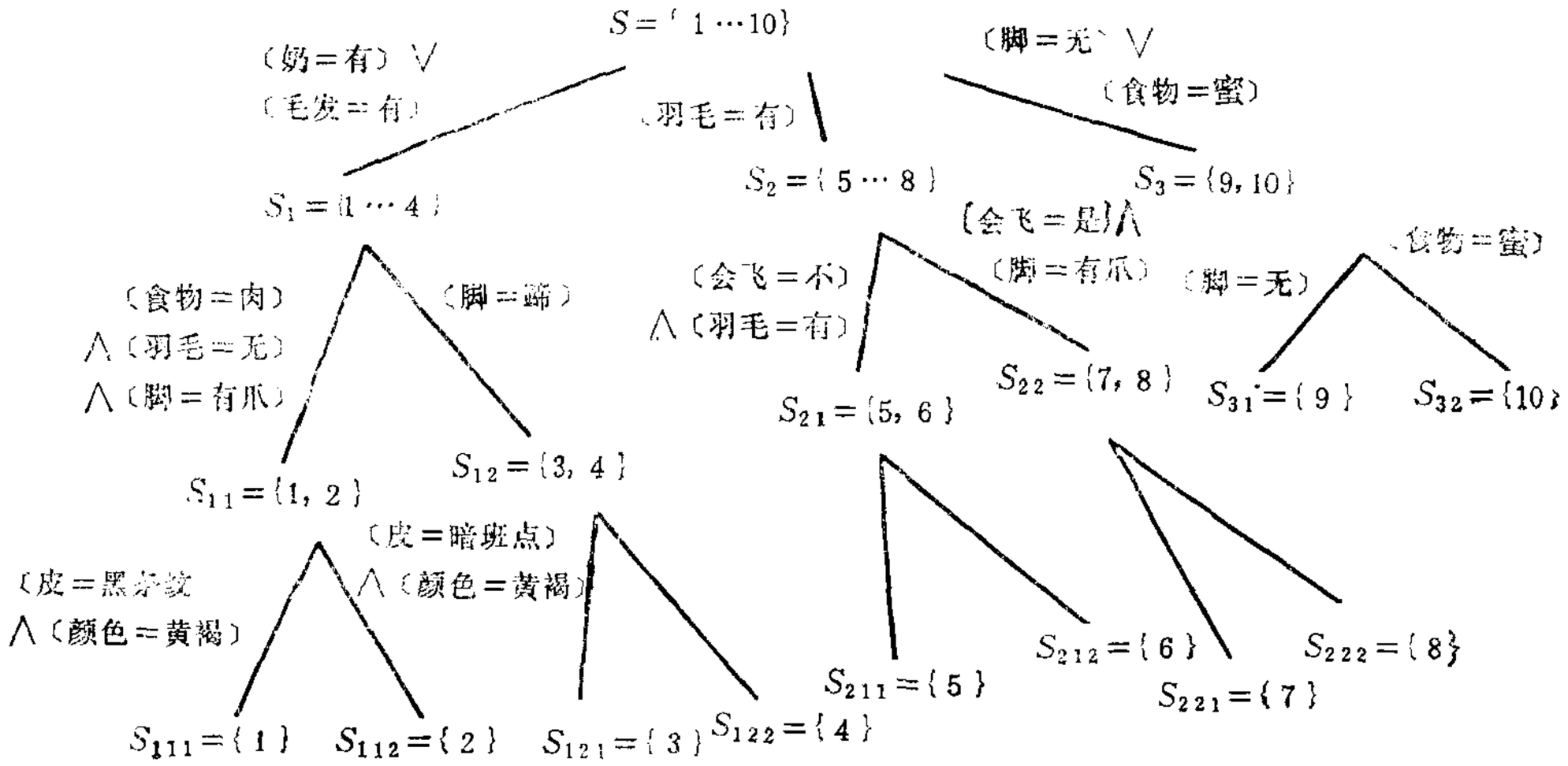

图 1　层次法规则发现产生的层次结构树

下面是层次法发现的规则的一部分：

1）[脚＝蹄] $\underset{s}{\Rightarrow}$ [奶＝有]∨[毛发＝有]，因 $S_{12}\subseteq S_1$.

2）[食物＝肉][羽毛＝无][脚＝有爪] $\underset{s}{\Rightarrow}$ [奶＝有]∨[毛发＝有]，因 $S_{11}\subseteq S_1$.

3）[会飞＝是][脚＝有爪] $\underset{s}{\Rightarrow}$ [羽毛＝有]，因 $S_{22}\subseteq S_2$.

上述规则的意义是自明的。例如规则 1 断言，如果一个动物有蹄，那么它一定有奶或有毛发，这在动物界是属实的。

2. 继承法规则发现

一类事物的概念是该类事物共同本质的抽象，人类认识事物旨在把握事物的概念。因此，概念构成人类知识的核心部分。知识发现只有使用概念才能获得更丰富、更深刻、更有意义的知识。从图 1 知，概念聚类系统 LEOBS 产生的子类一般都具有确定的意义。例如：S_1 是哺乳类的子集，S_2 是鸟类的子集，S_{11} 是食肉目的子集，S_{12} 是有蹄目的子集，等等。因此，如果用哺乳类、鸟类、食肉目和有蹄目等概念对相应的集合命名，就抓住了这些集合的本质特征。应当指出，概念同该概念的描述是有区别的，概念是普遍的和唯一的，而概念描述只能从某些侧面表示该概念。因此，一个概念的描述可能有许多个。例如，“有奶”和有“毛发”都是哺乳类的概念描述。由定义 6，任何概念的描述都逻辑地蕴涵该概念。例如，[奶＝有]⇒[类＝哺乳]，[脚＝蹄]⇒[目＝有蹄]等；另一方面，在产生描述的例子集合中，描述同概念彼此等价，例如在表 1 的例子集合 S 中，[奶＝有] $\underset{s}{\Leftrightarrow}$ [类＝哺乳]，[脚＝蹄] $\underset{s}{\Leftrightarrow}$ [目＝有蹄]等。

在层次法和平行法规则发现中，概念描述是由一个例子集合在整个已知例子集合中产生的。在使用命名概念的规则发现中，概念描述是一个集合在它的一个包集合中产生的，这种方法叫做继承法规则发现。

定理 2. 设 S_i 和 S_{ij} 等依次由层次分类产生，即 S_i 是 S_{ij} 的包集合，设 Label(S_i) 和 Label(S_{ij}) 等代表命名的概念，并设 $F_i = \mathrm{Cover}(S_i, S - S_i), \cdots$; $F_{ij} = \mathrm{Cover}(S_{ij}, S_i - S_{ij}), \cdots$; 则

1) $F_i \wedge F_{ij} \wedge \cdots \wedge F_{ij\cdots kl} \Rightarrow \mathrm{Label}(S_{ij\cdots kl})$,

2) $F_i \wedge F_{ij} \wedge \cdots \wedge F_{ij\cdots kl} \underset{S}{\Longleftrightarrow} \mathrm{Label}(S_{ij\cdots kl})$,

3) $\mathrm{Label}(S_{ij\cdots k}) \wedge F_{ij\cdots kl} \underset{S}{\Longleftrightarrow} \mathrm{Label}(S_{ij\cdots kl})$.

证明. 1)与 2)由引理 3 的 2)可知，$F_i \wedge F_{ij} = F(S_i, S) \wedge F(S_{ij}, S_i) \underset{S}{\Longleftrightarrow} F(S_{ij}, S)$，并由此类推得，$F_i \wedge F_{ij} \wedge \cdots \wedge F_{ij\cdots kl} \underset{S}{\Longleftrightarrow} F(S_{ij\cdots\cdots kl}, S)$，此即集合 $S_{ij\cdots\cdots kl}$ 的描述，因而它蕴涵并在 S 中等价该集合的命名概念 $\mathrm{Label}(S_{ij\cdots\cdots kl})$. 3)由引理 1 的 1)部分，

$$\begin{aligned}&\mathrm{Extension}(\mathrm{Label}(S_{ij\cdots\cdots k}) \wedge F_{ij\cdots\cdots kl}, S)\\ &= \mathrm{Extension}(\mathrm{Label}(S_{ij\cdots\cdots k}), S) \cap \mathrm{Extension}(F_{ij\cdots\cdots kl}, S)\\ &= S_{ij\cdots\cdots k} \cap [S_{ij\cdots\cdots kl} \cup (\text{可能出现在 } S - S_{ij\cdots\cdots k} \text{ 中的元素})]\\ &= S_{ij\cdots\cdots k} \cap S_{ij\cdots\cdots kl} \cup \varnothing = S_{ij\cdots\cdots kl}.\end{aligned}$$

因此，$\mathrm{Label}(S_{ij\cdots\cdots k}) \wedge F_{ij\cdots\cdots kl}$ 是命名概念 $\mathrm{Label}(S_{ij\cdots\cdots kl})$ 的描述，从而在 S 中彼此等价.

将继承法规则发现应用于表 1 的数据，KD_3 发现了十几个规则[8]，下面是其中的三条：1) [毛发＝有]∨[奶＝有]⇒[类＝哺乳]； 2) [类＝哺乳][食物＝肉]$\underset{S}{\Longleftrightarrow}$[目＝食肉]； 3) [目＝食肉][皮＝黑条纹][颜色＝黄褐]$\underset{S}{\Longleftrightarrow}$[动物＝虎]。其中“类”，“目”等是命名的概念.

本文提出了一种基于认识论的知识发现原理，并用一个集成化学习系统 KD_3 来模拟实现这一原理，最后还介绍了 KD_3 应用于数据库知识发现得到的一些有趣结果. KD_3 已在 SUN 工作站上运行，总共有三万多 PASCAL 语句行. 目前 KD_3 正在向实用化发展.

参 考 文 献

[1] Ganascia, J. G., CHARADE: A Rule System Learning System, Proc. IJCAI'87, Milan, Italy, (1987), 345—347.

[2] Goodman, R.M. and Smyth, P., Information-Theoretic Rule Induction, Proceedings of European Conference on Artificial Intelligence'88. Munich, West Germany.

[3] 洪家荣，认识论应当成为人工智能的主要基础 ——兼评当前国际上关于人工智能基础的论争，计算机科学，(1992),(2),1—5.

[4] Hong, J.R., Uhrik, C., A New Similarity-Based Learning Algorithm GS and a Comparison with ID_3, Proc. Int. Comp. Sci. '88, Hong Kong, (1988), 387—392.

[5] Stepp, R.E. and Michalski, R.S., Conceptual Clustering: Inventing Goal-oriented Classification of Structured Objects, Machine Learning: An Artificial Intelligence Approach II, Tioga, Palo Alto, CA: Morgan Kaufmann, (1986), 471—498.

[6] Quinlan, J.R., Learning Efficient Classification Procedures and Their Application to Chess End.

Games, Machine Learning: An Artificial Intelligence Approach, Tioga, Palo Alto, CA: Morgan Kaufmann, (1983), 463—482.

[7] Hong, J.R., Michalski, R.S., and Mozetic, I., AQ_{15}: Incremental Learning of Attribute-Based Descriptions from Examples, The Method and User's Guide, Tech. Rept. UIUCDCS-F-86-949, Dept. of Comp. Sci., Uni. of Illinois, Urbana, 1986.

[8] Hong, J.R. and Mao, C.J., Incremental Discovery of Rule and Structure by Hierarchical and Parallel Clustering, Knowledge Discovery in Databases, Chapter 10, AAAI/MIT Press, (1991), 177—194.

A THEORY OF KNOWLEDGE DISCOVERY AND ITS IMPLEMENTATION

HONG JIARONG

(Dept. of Computer Science. Harbin Institute of Technology, Harbin 150006 *China)*

ABSTRACT

In this paper. a theory of knowledge discovery is presented. This theory is based on the simulation of the process of human knowledge discovery, and comprises classification of data, generalization of data in each class against those in other classes, and discovery of implication relation between generated concepts. An implementation of this theory by using an integrated learning called KD_3 is described, and applications of KD_3 to automated construction of knowledge bases for expert systems are outlined.

Key words: machine learning; knowledge discovery; conceptual clustering; learning from examples; implication between concepts; integrated learning.

洪家荣　52 岁，教授。1964 年毕业于哈尔滨工业大学计算数学专业，1983—1986 年在美国伊利诺大学计算机科学系进修。现任哈尔滨工业大学计算机研究所智能系统研究室主任，中国机器学习学会、中国农业知识工程研究会、黑龙江省思维科学学会副理事长。主要研究方向包括机器学习、专家系统、手写字符识别、机器人路径规划与计算几何。发表论文 80 余篇。在国际上较有影响的工作有示例学习的扩张矩阵理论、通用学习系统 AQ_{15}，及三角形剖分算法等。