基于多层次概念提升的知识发现方法*)

A Knowledge Discovery Method based on Multi-level Concept Generalization

刘明吉　王秀峰　李宝林

(南开大学计算机与系统科学系　天津300071)

Abstract　Concept generalization, which is also called attribute-oriented induction, is a KDD method widely used. Through concept generalization we can improve the abstract level of attribute. Thus we can get more succinct rule. But with increasing of the number of attribute and the more and more complicated concept levels, the traditional method based on the set theory becomes lower and lower efficient. We propose a new heuristic algorithm based on genetic algorithm in this paper, which seems to work well while dealing with large scale and complex problem

Keywords　Concept generalization, Data mining, Attribute, Concept tree, Genetic algorithm

1　引言

数据库中的知识发现 KDD(Knowledge Discovery in Database)，也称为数据挖掘(Data Mining，简称 DM)，是近几年来随着数据库和人工智能发展起来的一门新兴的数据库技术。它从大量原始数据中挖掘出隐含的、有用的尚未发现的信息和知识(如规则、模型、规律、模式、约束等)，帮助决策者寻找数据间潜在的关联，发现被忽略的因素，因而被认为是解决现代社会"数据爆炸"和"数据丰富，信息贫乏"的一种有效方法[7,8]。

概念是语义描述的基本单位，也是数据库中各个描述对象的基本特征[1]。在数据库中，许多属性都可以按照自己的概念分类标准进行归类，形成概念汇聚点。各属性值以及概念依据抽象程度不同可构成一个层次结构，通常称为概念树。概念树一般由领域专家提供，与数据库中特定的属性有关。它将各个层次的概念按从一般到特殊的顺序分层排列。在概念树中，最一般的概念是没有具体特性的概念，用 Any 表示；最特殊的概念(叶结点)对应数据库中具体的属性值；而处于概念树层次结构中间的概念是对该属性值归纳过程中产生的更宏观的(更广义的)概念。基于概念提升和聚类是知识发现研究的一个重要方面。

2　问题的提出

基于概念树提升的知识发现方法是一种归纳方法，它其实是一个元组合并的处理过程。采用这种方法从数据库中发现规则知识的核心是执行基本的面向各属性的归纳。其基本思路是：首先，一个属性的较具体的值被该属性的概念树中的父概念所替代；然后，对知识基表中出现的相同元组进行合并，构成更宏观的元组，并计算宏元组所覆盖的元组数目；如果数据库记录生成的宏元组数目仍然很大，那么用这个属性的概念树中更一般的父概念去替代，或者根据另一个属性进行概念树的提升操作，最终生成覆盖面更广、数量更少的宏元组；最后，将归纳所得的最后结果转换成逻辑规则[6]。

然而这种在概念层次上进行取值抽象的方法，往往通过实例的合并来约简数据产生规则，完成归纳过程。这种基于集合论的归纳学习方法明显存在着一些不足：

①这种方法只考虑原始数据所提供的简单计数信息，而忽略了抽象的依据应当来源于实例集的共性。化简的规则应该尽量包含整个实例集的共同特征和信息要素。在提升过程中所设定被提升的属性对象的选择(即提升的方向)往往会直接影响最后的结果；另外，提升时所设定的阈值往往会造成概念过于一般或过于特殊，后果就是提升不到恰当的概念层，要么提升过度造成过度覆盖而丢失有用信息，要么提升不足而未能有效地降低数据的复杂度；

②由于目前实际应用的数据集中，数据属性的数目在不断增加，众多的属性又包含了错综复杂的多层

*)本论文的研究工作得到国家自然科学基金项目(79790130)和天津自然科学基金项目(993600811)资助。**刘明吉**　博士研究生，研究领域为数据仓库和数据挖掘。**王秀峰**　博导，研究领域为遗传算法、计算智能技术。**李宝林**　硕士研究生，研究领域为计算智能技术。

次概念，存在着数据的组合爆炸问题，从而带来了大量的集合运算。以传统的集合论归纳方法来处理这么大量的信息组合，已经显得力不从心。

从大量数据中寻找满足要求的概念是一种复杂的搜索过程，而且随着属性的增多以及属性值域的复杂化，计算量将呈指数增长，鉴于此，我们需要一种能从整体上提高概念学习效率的方法，它应该具有对空间存在较强的搜索能力，又要有较高的执行效率。作为拟自然的一种通用搜索方法，遗传算法在复杂空间问题领域中表现出来的良好性能使它很自然地被用到了概念学习领域。本文提出了一种基于遗传算法的概念提升学习方法，在遗传算法和知识发现相结合的研究方面做了一些有益的探索。

3 基于遗传算法的概念提升策略

遗传算法(Genetic Algorithm，GA)是一种集效率与效果于一身的优化搜索方法。它利用结构化的随机信息交换技术组合群体中各个结构中最好的生存因素，从而复制出最佳代码串，并使之一代一代地进化，最终获得满意的优化结果。在这里，我们在概念提升方法中引入遗传算法，对于由各个属性所构成的信息空间，设定信息空间的搜索点(GA 中的基因)，由这些搜索点共同构成一系列的搜索串(染色体)，在遗传操作的引导下，逐步得到较优的能基本体现属性分层概念的属性及其取值组合，从而得到精简的规则。具体地说，其处理方法主要包括以下几个步骤：

Step1 去掉没有意义的属性

在数据库中，尽管某一属性具有很多不同的属性值，但若没有更具一般意义的高层概念来归纳它们(即没有对应的概念树)，那么该属性在概念提升过程中被认为是没有意义的属性，如表1中"姓名"这个属性没有对应的概念树，则要去掉这个属性。

Step2 概念层次分类

其次对所处理的数据空间进行概念层次分类和定义。对于数据库中的各个属性，设置其在不同层次上的概念结构，编制其概念树。这部分工作应该在领域专家指导下或根据实际情况完成。

Step3 实施遗传算法

(1)遗传编码 在层次分类的同时为其设定恰当的遗传编码。本算法的遗传编码并不需要采用特殊遗传编码方式，完全可以根据概念树各个结点的代码设置灵活地采用各种染色体编码。编码设定的标准只有一个，就是易于遗传操作和算法实现。

(2)遗传操作

·交叉算子。在遗传算法中非常重要。首先按照一定的方法，随机地从交配池中取出要交配的一对染色体，然后进行交叉，产生一对新的位串。交叉算子可以分为单点交叉和多点交叉，本文采用两点交叉。交叉的方法是先根据位串长度 L，随机产生两个交叉位置，即[1，L-1]上的两个整数，然后进行交叉，例如：

染色体1 $10100_{\triangle}01010_{\triangle}10010$
染色体2 $10101_{\triangle}00010_{\triangle}10100$

交叉后代 $10100_{\triangle}00010_{\triangle}10010$
$10101_{\triangle}01010_{\triangle}10100$

·变异算子。变异算子根据一定的变异率 P_m，在染色体上随机选择一个基因，然后改变该基因的特征。变异不仅可以保证引入有用的遗传物质，保持种群的差异性，而且还能适当地提高 GA 的搜索效率。在本算法中，我们对变异算子做了一定的调整，设定了一个变异规则。我们的变异是概念结点在不同的概念层次树上取值所发生的变化。因此，在变异过程中严格按照概念由低到高的变异原则，基因变化可以由低到高，但不允许反向操作。

(3)评价函数 根据我们设定的遗传操作策略，使得我们的染色体沿着概念树逐步提高，它们所代表的概念也在不断地浓缩和泛化(能包含越来越多的原始数据中的数据记录)。我们直接以覆盖度作为其评价标准，覆盖度越大，表示染色体的概括性越好。

$$\text{覆盖度}=\frac{\text{被包含的元组个数}}{\text{全体元组个数}}\times 100\%$$

(4)算法流程 本算法采用简单遗传算法 SGA，编码十分简单，其主要算法流程如下：

①确定遗传算法的有关参数值，比如种群大小、迭代次数、停机条件；

②初始化种群，随机生成表示属性选择与否的染色体；

③根据评价函数的定义，计算各染色体的适合度函数值：

$eval(x_j^t)\quad j=1,\cdots,p\_size$

④根据各染色体适应值的比例信息

$$p_j^t=eval(x_j^t)/\sum_{i=0}^{p\_size}eval(x_i^t)$$

由轮盘赌的方式产生下一代染色体 $U^t=\{x^t\}$，其中 t 表示代数。

显然，适应值较大的个体参与产生下一代的概率较大，为了加快算法的收敛速度，可以将每代中适应值较大的一些个体强行传到下一代，而不受选择过程的限制；

⑤交叉。在 U^t 中以一定的概率 P_c 随机选择个体 x_i、x_j 进行两点交叉操作；

⑥变异。从交叉操作后得到的个体中，以一定的概率 P_m 随机选取某个个体，按照前面的变异算子定义，进行变异操作。至此我们得到了经过遗传操作后获得

的下一代种群 $U^{t+1}=\{x^{t+1}\}$；

⑦如果满足停机条件，则退出，否则转向③。

在本算法中，选择算子体现了适者生存的原则；交叉算子组合父代种群中有价值的信息，产生新的后代，具有遗传功能；变异算子的作用是保持群体中基因的多样性。

4 实例

我们以北方人才市场的人才招聘和就业信息为实验数据，来发现目前哪些素质或条件的人才的就业机会和工资待遇最好。为了便于处理，我们选取其中一些重要的属性数据，构造了以下这张人才就业信息表：

表1 人才就业信息统计表

姓名	性别	年龄	学历	户口	职称	职务	工作年数	单位性质	工资
李立新	男	25	中专	天津	无	司机	5	合资	1200
高　枫	男	26	大本	上海	工程师	程序员	3	合资	2200
刘　帆	女	24	大本	天津	会计师	财务部	1	机关	1300
……	……	……	……	……	……	……	……	……	……
程红英	女	28	硕士	天津	讲师	教师	2	大学	1100
邓子丹	男	53	大本	德州	高工	厂长	30	国企	2500
李红军	男	35	博士	北京	副教授	教师	7	大学	1700

我们从中随机抽取了1000条记录，设定置信度为25%，并根据实际情况，编制了以下一些概念树。

性别(any)：男(0)　女(1)

职称(any)
- 高级：教授、高工、研究员、会计师……
- 中级：工程师、讲师……
- 初级：助理工程师……

工资：高（2500以上）较高（1500～2500）
中（900～1500）低（500～900）极低（500以下）

学历(any)
- 高级：本科、硕士、博士……
- 中级：中专、高中……
- 低级：小学、文盲……

职务(any)
- 技术类：电子、机械、化工、计算机……
- 管理类：　经理、主管……
- ……

单位性质(any)
- 国企：事业、机关、文教……
- 三资：合资、独资……
- 个体：　私营、个体户……

经过算法运行以后，我们得到了如下一些约简规则("—"表示any)：

性别	年龄	学历	户口	职称	职务	工作年数	单位性质	工资	覆盖度
男	—	高	—	—	电子类	—	—	较高	35%
—	—	高	—	—	—	>5	独资	较高	30%
……	……	……	……	……	……	……	……	……	……

这样，我们通过第一条结果可以得到："学习电子类专业的男性大学毕业生的工资待遇较高"的规则。对于其余的结果，我们也可以得到相应的规则。

结论　本文提出了一种基于遗传算法的多层次概念提升学习方法，是遗传算法的强空间搜索能力在概念学习中的一个有益尝试。它充分利用数据内在的概念层次分类，与传统概念学习相比它具有以下特点：

①适于高维、多概念属性、多层次概念的数据库，适于处理较大规模问题；

②减少概念提升过程中的人为干预，在提升过程中引入启发式搜索思想，进一步提高了搜索学习效率；

③由于算法自动控制了提升的概念层次，避免了提升过度或提升不足；

④可以发现在不同特征属性和不同抽象层次上对原始数据空间的不同规则描述。

参考文献

1 Aetal D K. Using genetic algorithms for concept learning. Machine Learning, 1993, 13(2/3)

2 李敏强，寇纪淞. 基于数据库的层次概念知识体系的一种获取方法. 控制与决策，1999，14(Suppl.)：541～544

3 孟海军，李德毅. KDD中基于LAM的概念提升. 计算机科学，1996，23(2)：41～44

4 薛瀚宏 知识发现中的数据约简研究：[中国科学技术大学硕士学位论文]. 1998，5

5 陈恩红. 基于遗传算法的机器学习方法及应用研究：[博士论文]. 中国科学技术大学，1995

6 马建军，陈文伟. 数据开采中的集合论方法. 计算机世界，1997(6)

7 刘明吉，王秀峰，饶一梅. 一个混合特征属性选择算法. 计算机科学，2000，27(11)：75～78

8 Liu Mingji, Wang Xiufeng. A Knowledge Discovery Algorithm Based on Genetic Algorithm. The Third World Congress on Intelligent Control and Automation, IEEE. WCICA'2000