计算机科学1999 Vol 26№ 12

农业数据库中知识发现的研究*)

Study on Knowledge Discovery in Database of Agriculture

余建桥 梁 颖

(西南农业大学计算机系 重庆400716)

Abstract KDD in agriculture database application was carefully studied in this article. According to the agriculture knowledge attributes, we raised the sorting rules based on information's entropy, predicting methods based on genetic algorithms, and mining plan of association rules.

Keywords KDD, Sorting rules, Association rules, Genetic algorithms

1. 引言

数据库中的知识发现 KDD(Knowledge Discovery in Database)是近年来随着人工智能和数据库技术的发展而兴起的研究领域。它是从大量的数据中提取出隐含的、先前未知的、平凡的、具有潜在应用价值的信息或模式，如知识规则、约束、规律等。KDD 主要采用机器学习算法、统计方法进行知识学习，一般将 KDD 中进行知识学习的阶段称为数据挖掘。KDD 作为数据库和信息建设领域的前沿研究领域，已成为各重要科研机构和各主要工业研究实验室的研究目标，在决策支持、市场策略、生物医学等应用领域取得了令人兴奋的研究成果。

随着农业现代化与农业信息化进程的加快，农业信息数据库规模日益增大，农业领域 KDD 的研究悄然兴起。本文结合重庆市水稻品种区域布局专家系统的研究，对农业 KDD 作了研究，提出了基于信息论的农业系统分类规划，基于遗传算法的预测方法及农业系统关联规则的采掘。

2. 农业知识发现的任务与方法

2.1 知识发现的任务

KDD 主要是利用各种知识发现算法从数据库数据中发现有关的知识，根据发现的知识的不同种类，可以将 KDD 的任务分为以下几类：

特征 从与学习任务相关的一组数据中提取出关于这些数据的特征式，这些特征式表达了该数据集的总体特征。例如：通过对某种水稻生长特性的分析，提取出该品种水稻生长的一组特征式，利用这些特征式帮助农业生产者更好地掌握某水稻生长状况。

区分 通过对学习数据和对比数据的处理，提取出关于学习数据的主要特征，这些特征可以将学习数据与对比数据区分开来。例如：通过对不同水稻品种形态特征的分析比较，采掘出不同水稻品种间的区分规则。

分类规则 根据数据的不同特征，将其划归为不同的类，这些类建立在一定量的数据训练基础上。例如：通过对水稻籽粒中蛋白质、碳水化合物、脂肪、维生素和氨基酸等营养成分的分析，采掘出稻谷质量的分类规则。

关联规则 通过对数据的分析处理，发现不同属性的数据间的相互依赖关系。例如：通过对影响水稻产量的气候、地貌、土壤等众多因子的数据分析，采掘出因子间相互作用的关联规则。

预测分析 通过对数据的分析处理，估计数据库中某些缺失数据的可能值或一个数据集中某种属性值的分布情况，例如：根据完善的某种水稻的生长特征、形态特征等数据去填补另一品种水稻的缺空数据。

2.2 知识发现的方法

发现算法是从大量的数据中抽取知识的过程，鉴于数据本身的性质，符合推理、统计原理、信息论、遗传算法等知识被运用到知识发现的算法设计中，下面扼要介绍用于知识发现的几种方法。

基于符号推理法是建立于层次概念上的符号推理，以概念包容关系的分类树为基础，逐层分析归纳出相关规则。

基于统计分析方法主要有两种。一种是启发式搜索与统计方法相结合；另一种是基于回归分析的搜索，

*)重庆市攻关项目资助。

这是一种统计推理，主要用于对假设函数关系进行评价。

基于信息论方法是利用信息论中信息增益，寻找数据库中具有最大信息量的属性，建立决策树的一个结点，再根据属性的不同取值建立树的分支，最后把决策树转化为规则，利用这些规则对新事例进行分类

基于遗传算法的方法是运用遗传算法的自适应寻优及智能搜索技术，获取与客观事实最相容的问题解。本方法需要把 KDD 的一个具体任务表达为一种搜索问题，建立"染色体"及各基因。利用基因组合、交叉、变异的自然选择，优胜劣汰，得到问题的解。

3. KDD 在农业上的应用

3.1 基于信息论的分类规则采掘

水稻品种优良与否涉及到水稻全生育，抽穗期、成熟期的长短及这些时期获得的积温，同时与土壤类型，土质状况密不可分。把上述因素作为实例的属性，确定各属性的离散量值（如表1），引入信息熵的概念采掘出分类规则。

表1 属性离散量的确定例

属性：全生育期	<145天	<155天，>145天	>155天
属性取值	1	2	3

（1）信息熵的定义

定义1 某实例的第 i 个属性的信息熵定义为：

$$\text{entropy}(i) = -p \log_2 P$$

其中：P 表示个体属性 i 出现的概率。

定义2 某实例集的整体信息熵定义为：

$$\text{entropy} = \frac{1}{n}\sum -(n_i * \log_2(n_i/n))$$

其中：n 为整个实例集的大小，n_i 为属于第 i 个类别的实例的个数。

（2）建立决策树 计算整个实例集的信息熵 entropy 及个体属性的信息熵 entropy(i)，将个体属性的信息熵从大到小排序，选择具有最大信息量的个体属性对实例集进行分类形成树干，再选择第二大信息量的个体属性对实例集进一步分类形成树枝，…，最后丰满树叶形成一个基于信息熵的决策树。

（3）从决策树中得到下列形式的分类规则

IF（属性 i=…）OR（属性 j=…AND 属性 K=…）

THEN 类别=…

例如：IF（全生育期="2"）OR（抽穗期="3"AND 土壤="2"AND 土质肥力="1"）THEN 类别="高产类"

3.2 基于遗传算法的预测分析

在收集专家知识时，极有可能存在数据缺空、不完善，通过预测分析，估计出某些丢失数据的可能值，例如，水稻合理布局与土壤、地形、气候等因子有关，在收集某位领域专家知识时，缺少个别因子的数据，对专家知识库的建立带来困难。本系统利用遗传算法的搜索功能找到最匹配的缺空数据，完善领域专家知识，算法如下：

（1）确定两个父辈染色体

P1：无空缺数据的某位领域专家知识染色体。

P2：空缺数据的领域专家知识染色体。对应于空缺数据的某段基因为空白（假若基因5、6空缺），如图1。

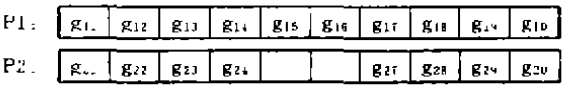

图1 染色体表达

（2）填补空白基因 用染色体 P1中基因5、6填补到染色体 P2的相同基因段中，得到图2。

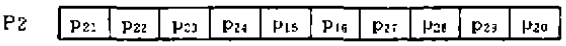

图2 填补基因

（3）交换操作 随机交换染色体 P1、P2的某段基因（如基因3、4），得到子染色体 C1、C2，如图3。

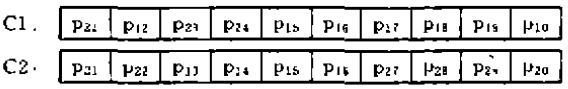

图3 交换操作

（4）变异操作 随机变异染色体 C1、C2的某基因，得到更新后的子染色体 C1、C2，如图4

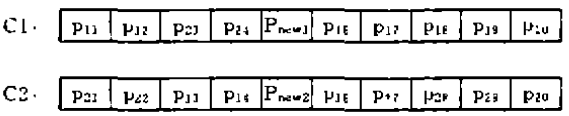

图4 变异操作

（5）收敛性判别 与事先确定的误差精度及适应值增长率进行比较。不满足条件重复步骤（3）、（4）、（5）；满足条件时遗传算法结束。

迭代收敛后的 C2，即为本系统预测分析后，得到的完善的专家知识。

3.3 关联规则的采掘

关联规则的采掘的作用是从指定的数据库中采掘出满足一定条件的依赖性关系。问题是这样描述的：设 $I=\{i_1, i_2, \cdots, i_n\}$ 是 n 个不同项目的集合，D 是针对 I 的事件（如：交易事件）的集合，D 中每一项事件包含若干项目 I'，且 $I' \subset I$。则关联规则表示为 $X \Rightarrow Y$，其中 $X, Y \subset I$，并且 $X \cap Y = \Phi$。X 称作规则的前提，Y 是结果。

定义3 对于X⊂I，如果事件集合D中包含X的事件个数为S，则X的支持度等于s，即Support(X)=s。

定义4 X⇒Y的置信度为：

Confidence(X⇒Y)=Support(X∪Y)/Support(X)

采掘关联规则的问题就是提出这样一些规则，它们的支持度和置信度分别大于用户指定支持度和置信度的阈值。因此该问题可以分解成以下两个子问题：产生所有支持度大于阈值的项；对于每个大于支持度阈值的项产生出所有比置信度阈值大的规则。以下是本系统中关联规则采掘实例。

土壤属性是决定水稻产量的重要因素之一。在研究水稻产量与众多土壤因子关系时，本课题调查了8位农技员。表2列出了他们认为对水稻产量起主要作用的因子。

表2 对水稻产量有影响的主要土壤因子

记录号	土壤因子				
1	有机质	粘粒	PH	全P	全N
2	有机质	CEC	PH	全P	全K
3	有机质	粘粒	CEC	PH	全P
4	有机质	粘粒	CEC	PH	全N
5	有机质	CEC	粘粒	有效K	有效P
6	有机质	CEC	粘粒	有效K	碱解N
7	PH	全P	全N	有效K	有效P
8	PH	全P	全K	碱解N	有效P

表中注释：N、P、K为氮，磷，钾；CEC为阳离子交换量；PH为酸碱度。

为了从表2的数据中了解各因子间是否存在关联，以采掘出有效的关联规则。经本专家系统的算法计算后，得到支持度大于阈值0.5的单项因子，见表3。把表3中单项因子进行两两组合，表4为支持度大于阈值0.5的双项因子组合。

表3 单项因子的支持度(支持度>0.5)

因子名称	支持度
有机质	6/8
PH	6/8
粘粒	5/8
CEC	5/8
全P	5/8

表4 双项因子的支持度(支持度>0.5)

因子对名称	支持度
有机质、粘粒	5/8
有机质、CEC	5/8
PH、全P	5/8

根据表4，得到以下6条关联规则，且置信度均大于阈值0.5：

R1：有机质⇒粘粒，support=0.625，confidence=0.83
R2：粘粒⇒有机质，support=0.625，confidence=1
R3：有机质⇒CEC，support=0.625，confidence=0.83
R4：CEC⇒有机质，support=0.625，confidence=1
R5：PH⇒全P，support=0.625，confidence=0.83
R6：全P⇒PH，support=0.625，confidence=1

R1与R2表示土壤粘粒可与有机质形成有机无机复合胶体。粘粒含量高时，有利于有机质的保持和积累。

R3与R4表示有机质与阳离子含量密切相关。土壤里的有机质含量越高，阳离子含量就越大。

R5与R6表示土壤里酸碱度是影响固磷作用的重要因素，不论是固磷作用的化学沉淀机制，表面反应机制或闭蓄机制，都不同程度地受到酸碱反应的影响。

以上关联规则的采掘有较高的置信度，与农业实际知识有较好的吻合，基本上反映了土壤中因子之间的关系。规则是合理的，可理解的。

结束语 本文对农业专家系统数据库知识发现进行了研究与探讨，尤其是在数据预测，采掘分类规则和关联规则方面提出了算法。实践证明，该算法基本上是可行的。在这些尝试中，笔者认为以下内容还待进一步地研究。

分类规则采掘中决策树形成算法直接影响着采掘效率。实例的属性及属性离散量值个数越多，形成决策树就越耗时间，这就对算法提高了更高的要求。

数据填补预测采用了遗传算法。确定适应值函数及误差精度直接关系到问题搜索是否收敛。

关联规则采掘中support与confidence的阈值选择至关重要。阈值过大，可能失去应有的关联规则；阈值过小，可能出现客观上并不存在的关联规则；合理的选择阈值，有利于真实知识的采掘。

参考文献

1 Fayyad U M, et al. Advances in Knowledge Discovery and Daga Mining. AAAI/MIT press, 1996

2 Brachman R J, et al. The Process of Knowledge Discovey in Database: A Human-centered Approach. In: Adavance In Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996

3 Weldon Jay-Louise. Data mining and visualization. Database Programming and Design, 1996, 9(5).

4 陈栋，等. KDD研究现状及发展. 计算机科学，1996，23(6)

5 余建桥. 预测模型获取的遗传算法研究. 计算机科学，1998，25(2)

6 李子木，等. 数据仓库技术的研究现状及未来方向. 计算机科学，1998，25(4)