一种综合范例推理和规则推理的发现技术*

邢乃宁[1], 高红梅[2], 孙志挥[2]

(1. 南京师范大学数学与计算机科学学院, 南京, 210097)
(2. 东南大学计算机科学工程系, 南京, 210096)

[摘要]　提出一种新的范例推理(CBR)和规则推理(RI)集成的混合构造方法, 该方法利用 RI 的结果为新的查询范例的各属性设置相关权重, 提高 CBR 中搜索源范例的效率.

[关键词]　范例推理; 规则推理; 范例检索; 相关权重; 相似性估计
[中图分类号]TP311;　[文献标识码]A;　[文章编号]1008-1925(2001)01-0036-04

随着数据挖掘研究与应用的不断深入, 人们正在将范例推理和规则推理等方法引入数据库的知识发现(KDD)研究. 众所周知, 范例推理 CBR(Case Based Reasoning)属于机器学习领域的类比学习, 即通过目标对象与源对象的相似性, 从而运用源对象的求解方法来解决目标对象的问题. CBR 的本质是增量的, 能够较好地学习非线性函数, 不像大多数归纳学习方法那样不易于在问题求解过程中进行规则集的扩展和修改. 但 CBR 具有一定局限: 不能产生便于人们理解的简单概念描述, 同时它对噪音很敏感. 而规则推理 RI(Rule Induction)则属于机器学习领域中的归纳学习, 它是以具体实例作为学习对象, 通过归纳推理, 获得概念的一般描述. 该方法可有效地用于数据分类与预测, 并能利用统计的方法有效地除去数据中的噪音. 但该方法同样存在一些缺陷, 如: 只能形成关于样例空间的超矩形区域, 而不能识别数据中小的、低频的变化, 同时规则不能很好地表达连续函数.

从具体推理过程可以看出 CBR 是从目标范例出发搜索源范例库, 因而具有确定性, 而 RI 是从已有的源范例归纳出规则而并不依赖于目标范例, 具有一定的预测性、一般性, 这两者具有一定的互补性. 本文提出一种新的 CBR 与 RI 集成的混合构造方法, 该方法利用 RI 的结果为新的查询范例的各属性设置相关权重, 提高 CBR 中搜索源范例的效率, 并利用相关权重来估算各源范例与新范例间的相似性, 然后将其按相关概率排序以获得最佳范例.

1　特征属性加权与参数估计

范例检索的目标是在源范例中找出某些能解决新范例中提出的问题, 即与新范例相关的范例, 并将它们以一定的相关概率排列. 这里相关概率应根据各范例解决新范例中所提出问题的能力加以精确估计. 在进行相关概率计算时, 需区分范例中各属性所占的比重, 也即为各特征属性设定权重, 这是范例检索的关键.

由于范例检索与数据库的信息检索有类似的目标, 因此可将数据库信息检索的原则用于范例检索. 在一个数据库信息检索系统中, 若对每个查询语句的响应是按文件被使用的概率进

* 收稿日期: 2000-07-02
基金项目: 国家自然科学基金资助(79970092)
作者简介: 邢乃宁, 女, 1966-, 硕士, 南京师范大学数科院讲师, 主要从事数据库系统及理论的教学与研究.

行排列, 那么该系统对数据利用的整体效率是最高的. 设p 表示相关文件中查询t 出现的概率, 用 q 表示相应的非相关文件中查询t 出现的概率, 若将该查询 t 的权值设为:

$$\log \frac{p(1-q)}{q(1-p)}$$

则将获得一个按文件进行排序的最优文件检索系统[1].

假定 CBR 中的某个新范例(查询范例) qc 包含一组属性$\{a_1, a_2, \ldots, a_n\}$, 范例检索与数据库文件检索中的查询语句一样, 可利用上述公式为 a_i 设定如下权重:

$$w(a_i) = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

其中p_i 是与新范例相关的某源范例中属性 a_i 出现的概率, q_i 是与新范例无关的某源范例中属性a_i 出现的概率. 这里源范例是指范例训练集中的各范例. 在此我们假定所有属性都为符号属性(非连续属性) , 对于连续属性可采用一种平均信息增量最小化启发式的离散化方法[2] 将连续属性值进行分段处理.

当源范例与新范例是否相关为已知时, 上述加权公式可写为:

$$\omega(a_i) = \log \frac{r(N-n-R+r)}{(n-r)(R-r)} \quad (1)$$

其中: N —— 源范例数; R —— 与新范例相关的源范例数; n —— 源范例中出现属性 a_i 的范例数; r —— 源范例中与新范例相关且出现属性a_i 的范例数.

尽管N 和n 非常容易得知, 但R 和r 通常事先无法知道, 因此需要有一种方法来估算这两个参数, 即判断出源范例中哪些与新范例相关, 从而决定属性的相关权重.

对于分类问题, 假设每个范例属于一个概念, 一个概念即为一个类别, 对于与新范例相关的源范例, 即能解决新范例所提出问题的源范例应和新范例属于同一概念. 因此, 如果我们能够估计出新范例属于哪一概念, 那么 CBR 中属于同一概念的源范例可被认为与新范例相关.

RI 系统能够分析数据并从中产生分类规则, 同时 RI 系统能够有效地使用统计方法监测噪音和非相关特征. 可以采用 RI 和演绎推理来估算 R 和 r 这两个参数. 在使用 RI 系统时, 可采用产生规则的后向剪枝和概率分类等手段以消去训练集中的噪音.

应用上述方法从训练集中产生规则, 当某一新范例提出时, 将它与规则进行匹配, 如果只有一条规则与新范例相匹配, 或者有多条规则与之匹配, 但这多条规则属于同一概念, 那么属于这一概念的源范例被认为与新范例相关. 如果是多规则匹配, 但所匹配的规则指示不同概念时, 这说明新范例处在所示各概念的边界, 在这种情况下, 所有属于被指示概念的范例均被认为是相关的. 如果没有规则可与之匹配, 则执行部分匹配, 以决定是否在某一规则中有与新范例相匹配的属性, 并在新范例与部分匹配的规则间计算出一个部分匹配分数. 这些部分匹配规则所表示的概念在这一分数的基础上进行竞争, 对于在竞争中取胜的概念, 那么属于这一概念的源范例均被认为是相关范例. 当一相关范例的集合 S 被确定后, R 被设定为 S 中范例的个数, r 被设定为 S 中含属性 a_i 的范例的个数.

由此我们希望对范例训练集中的各范例加权从而按相关概率将它们进行排列, 以确定最佳检索范例. 仍假定所有的属性为符号属性, 给出一个新范例 q, 对 q 中每一属性按公式(1) 计算其加权值. 对于训练集中的每个范例 x, 其所得分数为那些在 q 中的并同时在 x 中出现的属性的权值总和. 如果范例是根据这一分数按降序排列, 那么这些范例实际是按与新范例的相关概率以降序排列. 对于包含连续属性的范例, 可对其权值做出适当的调整, 即乘上某一属性在

q 中的值与在 x 中的值的差的绝对值. 因此为每个范例 x 打分变为估计 x 与 q 的相似度, 可用公式 $\text{Similarity}(x, q) = \sum_{i=1}^{n} W_i \times \text{Simil}(x_i, q_i)$.

其中 $\{a_1, a_2, \ldots, a_n\}$ 表示 n 个属性, 其值域分别为 $\{D_1, D_2, \ldots, D_n\}$. x_i 是第 i 个属性 a_i 在 x 中的值, q_i 是 a_i 在 q 中的值. W_i 是采用相关权重的方法所得的新范例中第 i 个属性的权值. 并且

$$\text{Simil}(x_i, q_i) = \begin{cases} 0 & \text{如果 } a_i \text{ 是符号属性, 且 } x_i \neq q_i; \\ 1 & \text{如果 } a_i \text{ 是符号属性, 且 } x_i = q_i; \\ 1 - d(x_i, q_i) & \text{如果 } a_i \text{ 是连续属性.} \end{cases}$$

其中 $d(x_i, q_i)$ 表示 x_i 与 q_i 间规格化距离, 定义为:

$$d(x_i, q_i) = |x_i - q_i| / D_i$$

2 算法描述

该混合方法用于分类时的算法如下:

(1) 将新范例与产生的规则进行匹配;

(2) 如果只存在唯一的匹配, 即只有一条规则与新范例匹配, 那么该新范例被分在由该规则所表示的那一类中;

(3) 如果存在多规则与之匹配, 但这些规则表示同一类 C, 那么新范例被归入 C 类中;

(4) 如果存在多规则匹配, 但这些规则表示了不同的类, 或者不存在与之匹配的规则, 但部分匹配存在, 那么属于匹配规则(或部分匹配的规则) 所表示的那些类的范例被认为是相关范例, 然后根据上节所述加权法和相似性估算将范例进行排序, 加权方法中的参数由相关范例确定. 转向(6);

(5) 如果部分匹配也不存在, 仍用上节的方法为源范例排序, 此时加权方法中的参数设为 $R = r = 0$;

(6) 从排序后的范例中选出 K 个最相关范例, 组成集合 S, K 是一个由用户定义的参数;

(7) 如果 S 中所有的相关范例属于 C 类, 则将新范例归入 C 类;

(8) 否则, 对于 S 中的每个类 Y_i, 为其计算一个决定分数, 定义如下:

$$DS(Y_i) = \sum_{j=1}^{m} \text{Similarity}(c_i, q)$$

其中, c_i 表示 S 中属于 Y_i 类的 m 个范例中的一个, q 是新范例;

(9) 将新范例归入得分高的那一类.

在上述过程中, (1) ~(3) 执行演绎推理对新范例分类. CBR 被用来解决与规则匹配时的矛盾及部分匹配问题. (4) ~(6) 执行范例检索, 通过相关加权方法决定一组相关范例. (7) ~(9) 执行范例求解, 根据被检索到的范例决定新范例的所属类别.

3 实例分析

由于需对 RI 只能形成关于样例空间的超矩形区域, 而不能识别数据中小的、低频的变化等不足进行检测. 为此, 我们将该方法与 C4.5 这样一种能够处理连续属性的决策树方法进行比较. 现将该混合方法和 C4.5 方法分别用于解决以下两种问题, 每个问题包含有一个目标概

念, 每个问题中的一个实例可能属于该目标概念, 也可能不属于目标概念.

问题 1: 包含 4 个列名条件属性, 每个属性可能有 5 个取值: - 2, - 1, 0, 1, 2, 目标概念是: 当且仅当某实例前 3 个属性中有任意两个或两个以上的属性为负值, 那么该实例属于概念. 从 625 个实例空间中随机抽取 125 个作为训练实例, 剩余的作为测试集.

问题 2: 包含两个连续属性分别代表两维空间中的两个轴(x 和 y), 目标概念是: 如果

$$\frac{x2}{a}+\frac{y2}{b}\leq c$$

那么该实例属于目标概念, 其中 a、b、c 为常数. 从样例空间选择一组实例, 其中 1/ 3 用作训练集, 剩余的用作测试集. 通过对上述问题进行试验, 表 1 列出对测试集进行精确分类的结果.

表 1 的试验结果表明: 该方法与 C4. 5 方法所得正确率均为 100%, 这是因为问题 1 中的概念具有 "矩形" 边界区域, 规则推导算法善于学习和表示这些概念. 而在学习双轴决策规则时, 该方法优于 C4. 5, 这一结果与我们希望的一致, 规则不是特别善于表示非线性边界的概念. 为说明问题起见, 我们不妨来讨论一下只包含单轴决策学习的情况. 设连续属性 x 表示年龄, $x<50$ 为目标概念, 在运用 RI 进行分类时, 对该属性取值进行分段处理, 假定分段区域为[0, 30], [30, 60], [60, 90], 那么实例空间中落在[30, 60] 的实例有 66% 的可能属于目标概念, 而有 33% 的可能不属于目标概念, 对于这种不确定的情况, 若采用 C4. 5 的分类方法对测试集进行分类时, 会产生一定的误差, 而采用 CBR 后, 通过相似性估算, 可以减小分类误差, 提高分类精度, 这就是该新方法的优越性.

表 1 两种方法的比较

问题	混合方法/ %	C4. 5/ %
问题 1	100	100
问题 2	98. 81	98. 6

4 结束语

本文在 CBR 和 RI 方法的基础上, 考虑两者的互补性将两种方法相结合, 利用 RI 的结果为新的查询范例的各属性设置相关权重, 以达到提高 CBR 中搜索源范例效率的目的. 并给出具体的算法, 同时通过实验结果表明混合方法在一般情况下优于多数方法如 C4. 5 方法等. 对于上述方法, 如在连续属性的相似度定义方面作改进, 则其搜索源范例的效率可得到进一步的提高.

[参考文献]

[1] Roberston S E, Spark Jones K. Relevance Weighting of Search Terms. J Am Soc Information Science, 1976, (27): 129～146

[2] Fayyad U M, Irani K B. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. Proc IJCAI-93, 1993: 1022～1027

[3] Nick Cercone. Ruleinduction and casebased reasoning hybrid architectures appear advantageous. IEEE Trans on Knowl and Data Eng, 1999, 11(1): 166～173

(下转第 76 页)

2.6 其它沉淀剂

本文试验了 $Ca(OH)_2$ + H_3PO_4、MgO + H_3PO_4、MgO 作沉淀剂, 在 pH= 8～12 时范围内氨去除率为 21～40%, 效果较差. 这是因为它们与 NH_4^+ 不能形成像 $MgNH_4PO_4$ 一样的溶度积较小的产物.

本文采用 $Mg(OH)_2$ + H_3PO_4 为沉淀剂, 在 pH= 9～11 范围内, 将废水中的氨以复合肥 $MgNH_4PO_4$ 的形式析出, 氨的去除率> 95%, 实际水样的试验结果表明, 处理后的废水中氨的含量达到工业废水氨氮的排放标准.

[参考文献]

[1] Tunay O, *et al*. Ammonia removal by magnesium ammonium phosphate in industrial wastewaters [J]. Wat Sci Technol, 1997, 36(2～3): 399～406

[2] 原丁. 氮肥工业中氨氮废水治理技术进展[J]. 化工环保, 1995, 15(2): 73～77

[3] 赵庆良, 李湘中. 化学沉淀法去除垃圾渗滤液中的氨氮[J]. 环境化学, 1999, 20(5): 90～92

Ammonia Removal from Wastewater by Chemical Precipitation

WANG Yuping, PENG Panying, CHEN Yuchao, CUI Shihai

(College of Chemistry and Environment Science, Nanjing Normal University, Nanjing, 210097, PRC)

Abstract: Ammonia in wastewater can be removed by using chemical precipitants $Mg(OH)_2$ and H_3PO_4 to form a composite fertilizer $MgNH_4PO_4$, with the removal ratio of ammonia above 95% under the optimal conditions. The effects of pH value, precipitant constitutions and the initial concentration of ammonia in the wastewater on ammonia removal efficiency are disscussed.

Key words: ammonia; chemical precipitation; wastewater treatment

[责任编辑: 严海琳]

(上接第 39 页)

Integrated Discovering Technology of Case-Based Reasoning and Rule Induction

Xing Naining[1], Gao Hongmei[2], Sun Zhihui[2]

(1. College of Mathematics and Computer Science, Nanjing Normal University, Nanjing, 210097, PRC)

(2. Department of Computer Science and Engineering, Southeast University, Nanjing, 210096, PRC)

Abstract: In this articlea new discovering technology is put forward that integrates case-based reasoning and rule-induction. The result of RI is used to assign the weight for each attribute of a new case in order to improve the retrieve efficiency in CBR.

Key words: case-based reasoning; rule induction; case retrieve; relevent weight; similarity measurement

[责任编辑: 刘健]