第 20 卷 第 6 期　　自 动 化 学 报　　Vol. 20, No. 6
1994 年 11 月　　ACTA AUTOMATICA SINICA　　Nov., 1994

研究简报

一种新的基于记忆的归纳学习算法

王正欧　林　燕
（天津大学系统工程研究所　天津　300072）

关键词：基于记忆的推理，归纳学习，不相似量度，分类树.

1　引言

归纳学习是一种研究得较多的机器学习算法，其中 Michalski 的方法[1]是有代表性的. 但正如文[2]中指出的，当描述元数量较大时，这种方法并不实用. 为此文[2]曾提出了一种直观而又简便的归纳学习算法. 本文则提出了另一种简便的基于记忆的归纳学习算法.

2　有关的概念和符号

2.1　数据记录的表示方法

这里采用基于记忆推理[3]的数据记录表示方法，现简述如下：一个记录 R 是由若干特征域和一个目标域组成. 记录 R 的特征域 F 的值记作 $v.F_R$，目标域 G 的值记作 $v.G_R$. 所谓特征就是一个特征域和一个值的结合. 特征域不允许为空，而目标域则允许为空. 一个目标记录 TR 就是包含一个空目标域的记录，其目标域值是待推导的. 一个数据库 D 就是由那些目标域值已充填的记录组成的集合，例如训练事例等.

在归纳学习算法中，代替目标记录 TR 的是要形成其概念判别描述的目标记录类型 TC.

2.2　符号表示

D——所有训练事例的数据集合；

R——数据集合 D 中任一记录；

V_R——目标域可能值的集合；

TC——目标记录类型；

$v.F_{TC}$——TC 关于特征 F 的值的集合. 当 $v.F_{TC}$ 是数值集合时，$v.F_{TC}=\{m, M\}$，其中 $m=\min\{v.F_R:\forall R\in TC\}$，$M=\max\{v.F_R:\forall R\in TC\}$；当 $v.F_{TC}$ 为非数值集合时，取 $v.F_{TC}=\{v.F_R:\forall R\in TC\}$.

本文于 1992 年 2 月24日收到.

3 基于记忆的归纳学习算法

3.1 目标类型记录 TC 特征权的确定

TC 的特征 F 的权重可用该特征对目标域(即对类型判别)影响的大小来衡量. 具体方法是限定数据集合为 $D[v.F_R \in v.F_{TC}]$, 在此集合中找出目标域,取不同值的频率. 基于这些频率 TC 关于特征 F 的权重 $\omega_{TC}(F,G)$ 可表示为

$$\omega_{TC}(F,G)=\left[\sum_{g\in V_G}\left(\frac{|D[(v.F_R\in v.F_{TC})\Lambda(v.G_R=g)]|}{|D[v.F_R\in v.F_{TC}]|}\right)^2\right]^{1/2}. \tag{1}$$

式中右端分子表示所限定的数据集合与具有相同目标域值的元素集合的交集; $|D|$ 表示 D 中所包含的记录的个数.

由式(1)确定的权重表明了特征 F 在决定对 TC 判别描述中作用的大小.

3.2 TC 与所有其他类型的不相似量度

有时一个特征的不同值对目标域的影响是等价的,因此还需要考虑 $v.F_{TC}$ 和 D 中其他类型 C 的 $v.F_C$ 作用于 G 的效果差别 $d_{TC,C}(F,G)$. 通过计算子集合 $D[v.F_R\in v.F_{TC}]$ 及 $D[v.F_R\in v.F_C]$ 中目标域 G 取不同值的频率差为

$$d_{TC,C}(F,G)=\sum_{\substack{g\in V_G\\ C\neq TC}}\left(\frac{|D[v.F_R\in v.F_{TC})\wedge(v.G_R=g)]|}{|D[v.F_R\in v.F_{TC}]|}\right.$$

$$\left.-\frac{|D[v.F_R\in v.F_C)\wedge(v.G_R=g)]|}{|D[v.F_R\in v.F_C]|}\right)^2, \tag{2}$$

$d_{TC,C}(F,G)$ 的大小表明了 F 对 TC 和 C 区分能力的大小.

把 $\omega_{TR}(F,G)$ 和 $d_{TC,C}(F,G)$ 结合,即可得到 TC 与 C 间关于特征 F 的不相似量度:

$$\delta_{TC,C}(F,G)=\omega_{TC}(F,G)\cdot d_{TC,C}(F,G). \tag{3}$$

令 $\Delta_{TC,C}(F,G)$ 表示用特征 F 区分 TC 和 C 的总体能力,即得 TC 和 C 间总的不相似量度:

$$\Delta_{TC}(F,G)=\sum_{C\neq TC}\delta_{TC,C}(F,G)=\omega_{TC}(F,G)\sum_{C\neq TC}d_{TC,C}(F,G). \tag{4}$$

上述计算中 F 可取自 TC 的所有特征,因此上述量度对不同的 F 可平行地计算出来. 此外, D 中所有的类型均可分别视为 TC, 因此对不同的 TC 也可平行地计算上述量度. 由此可以确定 D 中所有类型的总体区分能力

$$\Delta(F,G)=\sum_i\Delta_i(F,G), \tag{5}$$

并提取具有最大的从其他类型中区分出 TC 能力的特征 F^*, 作为分类决策树的结点,即

$$\Delta(F^*,G)=\max_F\{\Delta(F,G)\}. \tag{6}$$

3.3 算法的具体步骤及决策树的形成

(i) 对训练事例,根据各类型全部特征值应用闭区间规则[1]进行预处理;

(ii) 按式(1)—(6)进行计算,确定 F^*;

(iii) 把 F^* 作为决策树的一层，并按其特征值确定树的各个分支。如各分支均能唯一地达到代表目标域中某个值的叶结点,则转 (v)，否则转 (iv);

(iv) 从原数据集合 D 中限定某分支所对应的新的数据集合 D'，重复步骤 (ii)—(iii);

(v) 生成概念的判别描述。概念描述采用产生式规则形式,并通过深度优先搜索决策树来获得。树中不同层次的描述元之间取其合取，同一层内取其析取作为产生式规则的条件;树中的叶结点作为产生式规则的结论。

用上述算法做了几个实例,验证了其有效性。有关内容可见作者硕士学位论文[1)]。

4 结束语

本算法从计算特征的权重出发,计算相似性量度,并形成决策树，最后生成概念的判别描述规则。其特点是：算法简明，便于平行计算。因而适用于较大规模的归纳学习问题。

参　考　文　献

[1] Michalski R.S. A theory and methodology of inductive learning. *Artificial Intelligence*, 1983, **20**:111—161.

[2] 王正欧,林燕,一种新的归纳学习算法—基于特征可分性的归纳学习算法。自动化学报，1993，**19**(3)：328—331。

[3] Stanfil C, Waltz D. Toward memory-based reasoning. *Communication of the CAM*, 1986, **29** (12): 1213—1228.

A NEW MEMORY-BASED INDUCTIVE LEARNING ALGORITHM

WANG ZHENGOU　　LIN YAN

(*Institute of Systems Engineering, Tianjin University, Tianjin* 300072)

Key words: Memory-based reasoning, inductive learning, dissimilarity measure, classifying tree.

1) 林燕。专家系统中基于记忆的归纳学习模型的研究。天津大学硕士学位论文,1991。