第19卷 第3期
1993 年 5 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 19, No. 3
May, 1993

短 文

一种新的归纳学习算法——基于特征可分性的归纳学习算法

王正欧　林　燕
（天津大学系统工程研究所，天津　300072）

摘　要

本文提出了一种新的基于特征可分性的归纳学习算法（SBI）. 与现有各种归纳学习算法相比，该方法直接从特征对不同类型的可分性出发，建立可分性判据，然后形成决策树，可对多种概念进行判别. SBI 算法具有直观且计算简便等优点. 本文以实例表明了 SBI 算法的有效性.

关键词： 归纳学习，特征，可分性判据，决策树.

一、前　言

在机器学习领域中较为普遍的一种学习方式是归纳学习，其中 Michalski 的研究[1]是有代表性的. Michalski 的方法实质是从训练事例出发，研究其分类规则，然后泛化这些规则形成归纳规则. 这种方法当描述元(特征)数量较大时，由于算法的繁杂而变为不实用. 本文提出了一种新的基于特征可分性的归纳学习算法，使计算大大简化，同时又不失其有效性.

二、基于特征可分性的归纳学习算法（SBI）

1. SBI 算法的提出

归纳学习的逻辑基础是完全性条件和一致性条件，即某一概念类型的描述覆盖该类型的所有训练事例，并拒绝所有其它概念类型的训练事例. 这是构造分类决策树的基本原则. 如果决策树中的特征对于不同类型，其特征值差别很大，那么就可以通过最少的层次使树中的各个分支达到唯一的结点，从而使树中所包含的归纳规则具有最少的描述元. 问题的关键在于如何选择那些具有最强分类能力的特征. 这里直接从特征对不同类型的可分性出发，建立可分性判据，提出基于特征可分性的归纳学习算法 SBI(Separability-Based Inductive learning algorithm).

本文于 1991 年 11 月 22 日收到.

2. 有关的概念和符号

1）数据记录的表示方法．这里采用基于记忆的推理[2]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）符号表示．D为所有训练事例的数据集合；TC 为目标记录类型；R为数据集合D中任一记录；$v.F_c$ 为数据子集合c关于特征F的特征值． 当 $v.F_c$ 是数值集合时，$v.F_c=\{m,M\}$，其中 $m=\min\{v.F_R:\forall R\in c\}$，$M=\max\{v.F_R:\forall R\in c\}$；当 $v.F_c$ 为非数值集合时取 $v.F_c=\{vF_R:\forall R\in c\}$；$v.F_{TC}$ 为目标记录类型 TC 关于特征F的值，以同 $v.F_c$ 相同的方式获取；$v.G_c$ 为子集合c的目标域值；$v.G_{TC}$ 为目标记录类型 TC 的目标域值．

3. 可分性判据的确定

设数据集合D中包含m个数据子集合c，每个子集合中具有相同的目标域值 $v.G_c$，则定义特征F关于目标记录类型集合 TC 与其它数据子集合的可分性判据如下：

$$J_{TC}(F,G)\triangleq\frac{1}{m-1}\sum_{c\neq TC}\left[1-\frac{|D[(v.F_R\in v.F_{TC})\wedge(v.G_R=v.G_c)]|}{|D[v.G_R=v.G_c]|}\right],\tag{1}$$

若记

$$J_{TC,c}(F,G)=1-\frac{|D[(v.F_R\in v.F_{TC})\wedge(v.G_R=v.G_c)]|}{|D[v.G_R=v.G_c]|},\tag{2}$$

则有

$$J_{TC}(F,G)=\frac{1}{m-1}\sum_{c\neq TC}J_{TC,c}(F,G).\tag{3}$$

在式(1)和(2)中 $|D|$ 表示集合D的容量，即D中包含的记录个数；$D[v.G_R=v.G_c]$ 为数据子集合 c，其元素具有相同的目标域值 $v.G_c$。

在式(1)中，$D[(v.F_R\in v.F_{TC})\wedge(v.G_R=v.G_c)]$ 是目标记录类型 TC 与其它数据子集合的交集，满足该集合特征值要求的元素可以对应于不同的目标域值 $v.G_{TC}$ 和 $v.G_c$，即该集合的特征值范围对于不同类型c和 TC 不存在任何差别，是不可分的．因此称该集合为不可分集合，其元素为不可分元素．

$$\frac{|D[(v.F_R\in v.F_{TC})\wedge(v.G_R=v.G_c)]|}{|D[v.G_R=v.G_c]|}$$

表示了不可分元素在数据子集合c中的比率，而 $J_{TC,c}(F,G)$ 则表示了可分性元素在c中所占的比率． 特征可分性判据 $J_{TC}(F,G)$ 正是在综合考虑目标记录类型 TC 对于所有$m-1$个其它数据子集合的可分性元素所占比率的基础上建立的．

$J_{TC}(F,G)$ 是在[0,1]范围内的值,其值愈近于 1，则特征 F 区分不同类型的能力愈大;其值愈近于零,则 F 的区分能力愈小.

4. SBI 算法的计算步骤

第一步,对训练事例,根据各类型的全部特征值应用闭区间规则[1]进行预处理;

第二步,计算各个特征分别关于各个目标类型记录(每个类型都可当作 TC) 与其它数据子集合之间的可分性判据 $J_{TC}(F,G)$;

第三步,对每个特征 F 计算 $J(F)=\sum_{TC} J_{TC}(F,G)$，并选择其最大者所对应的特征作为决策树的一层;

第四步,通过比较各个目标类型关于上一步所得特征的特征值,确定决策树的各个分支. 如果各个分支均能唯一地达到代表目标域中某个值的结点,则转入第六步,否则转第五步;

第五步,从原数据集合中限定某分支所对应的数据子集合,作为新的数据集合，重复第二至第四步;

第六步,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.

三、运 算 实 例

这里给出两种人(白种人和混血儿)分类的简单例子来说明 SBI 算法的有效性. 表 1 给出了 8 个训练事例的数据库. 每种人由 3 个特征来刻划. 特征的含义及其值域见表 2. 表 2 中 A,B,C,D,E,F,G 分别表示高、矮,金色,黑色,红色,褐色,蓝色. 它们分别是各特征相应的特征值.

表 1 种族分类的训练事例集合

事例 / 特征	1	2	3	4	5	6	7	8
height	*B*	*A*	*A*	*B*	*A*	*A*	*A*	*B*
hair	*C*	*C*	*E*	*D*	*D*	*C*	*D*	*E*
eyes	*G*	*F*	*G*	*G*	*G*	*G*	*F*	*F*
class	1	2	1	2	2	1	2	2

表 2 特征的含义及值域

特征	含义	值域
height	身高	$\{A,B\}$
hair	发色	$\{C,D,E\}$
eyes	眼睛颜色	$\{F,G\}$

用 SBI 算法按步骤逐步进行可产生决策树,并得以下两条产生式规则:

$$(v.\text{hair}_R = \text{红色}) \vee (v.\text{hair}_R = \text{金色}) \wedge (v.\text{eyes}_R = \text{蓝色}) \Longrightarrow (v.\text{class}_R = 1)$$

$$(v.\text{hair}_R = \text{黑色}) \vee (v.\text{hair}_R = \text{金色}) \wedge (v.\text{eyes}_R = \text{褐色}) \Longrightarrow (v.\text{class}_R = 2)$$

可以验证上述两条规则覆盖了全部正事例，排除了所有的反事例．因而可以当作两种族的分类规则，这对预言该领域的新事物是有用的．关于本例的详细运算过程可见作者硕士学位论文[1)]．

四、结　　论

SBI 算法从特征对不同类型的可分性出发，建立可分性判据，然后形成决策树，与一般归纳学习算法相比，具有概念直观、计算量减少，同时又不失其有效性的优点，因而该算法可适用于训练事例集合较大，特征数量多的较大规模归纳学习场合．

参　考　文　献

[1] Michalski, R. S., A Theory and Methodology of Inductive Learning, *Machine Learning*, 83—134, (1984).

[2] Craig Stanfil and David Waltz, Toward Memory-Based Reasoning, *Communication of the CAM*, **12**(1986), 1213—1228.

A NEW INDUCTIVE LEARNING ALGORITHM——SEPARABILITY-BASED INDUCTIVE LEARNING ALGORITHM

WANG ZHENGOU　LIN YAN

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

ABSTRACT

A new separability-based inductive learning algorithm is proposed in this paper. The algorithm is different from existing inductive learning algorithms. Starting directly from the separability of features for different classes, building a separability criterion, then forming a decision tree, the algorithm can classify multiclass concepts. The algorithm is intuitive, simple, and convenient for computation. Its effectineness is illustrated by an example in this paper.

Key words : Inductive learning; feature; separability criterion; decision 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)．