第4卷第4期
1993年8月

软件学报
JOURNAL OF SOFTWARE

Vol. 4, No. 4
Aug. 1993

一个基于信息论的示例学习方法

钟 鸣

(解放军防化研究院计算中心，北京 102205)

陈文伟

(国防科学技术大学，长沙 410073)

张凯慈

(解放军防化研究院计算中心，北京 102205)

AN INFORMATION—BASED METHOD IBLE FOR LEARNING FROM EXAMPLES

Zhong Ming

(*Computing Centre of Chemical Defence Institute, Beijing* 102205)

Chen Wenwei

(National University of Defence Technology, Changsha 410073)

Zhang Kaici

(*Computing Centre of Chemical Defence Institute, Beijing* 102205)

Abstract This paper presents a new method IBLE for learning from examples with the concepts of capacity, maximal plausible decode criterion of information theory. The method doesn't depend on the prior probability of class. In the method, the attributes are strongly associated, the knowledge representation is intelligible. We use IBLE in the interpretation of mass spectra, good result is obtained and the average predictive accuracy for eight classes of compounds is 93.96%. This result is superior to experts.

摘要 本文利用信息论中信道容量、最大似然译码准则等概念，提出一个新的示例学习方法IBLE. 此方法不依赖类别先验概率，特征间为强相关，具有直观的知识表示. 将它用于质谱解析，结果很好，八类化合物平均正确预测率为93.96%，高于专家水平.

§1. 引 言

机器学习是人工智能研究的一个重要课题. 机器学习方法有多种[1]，其中示例学习(Learning from examples)是一种较重要的方法. 由于它可用于知识获取，为解决专家系

本文1991年5月收到. 作者**钟鸣**，高级工程师，1992年硕士毕业于国防科学技术大学，主要研究领域为人工智能，化学计量学. **陈文伟**，副教授，主要研究领域为智能决策支持技术和机器学习. **张凯慈**，副研究员，主要研究领域为数据库，管理系统.

统的知识获取瓶颈问题提供重要的手段，得到广泛的关注和研究. 出现了一些有名的算法和系统，如 Quinlan 的 ID_3，Michalski 的 AQ_{11}，洪家荣的 AE_5 等.

ID_3 是当前国际上著名的一个算法，关于它的性能探讨，原理应用、算法改进方面的文章不断出现. 然而关于 ID_3 的各种改进，我们认为，一般都不能解决下述几个问题，即类别先验概率问题[2]、特征弱相关问题[3]和知识表示的不直观问题[4]. 针对这些问题，本文提出了一个新方法 IBLE(Information—Based Learning from Examples). 并用于实际问题，取得很好结果.

本文重点在于提出算法，关于示例学习和信息论的概念请参看文[4—6].

§2. 几个问题及解决办法

2.1 类别先验概率问题

设问题空间为 S，被分成 u_1、u_2 两类，u_1 称为正例类，u_2 称为反例类. 每个例子由 n 元组 $(A_1, A_2, \cdots, A_n)$ 表示，其中 A_k 值域为 $V=\{v_1, v_2, \cdots, v_q\}$. 示例学习归纳出的规则要判定一个未知实体是正例还是反例，这种判定是通过考察实体的各特征取值情况作出的.

ID_3 首先选择重要特征，然后以所选特征为根进行分支构造判定树[4]. 特征重要性用互信息度量. 互信息是 $P(u_i)$ 与 $P(v_j/u_i)$ 的函数[6]. $P(u_i)$ 表示正、反例在问题空间中的比例，$P(v_j/u_i)$ 代表正例和反例在 A_k 处的取值情况. 实际问题中，由于问题空间太大，不可能得到所有的正例和反例，只能利用已知的例子去近似求出 $P(u_i)$ 和 $P(v_j/u_i)$. 具体做法是选取一定数目的例子组成训练集 X，$X=PE \cup NE$，PE 为正例集，NE 为反例集.

令 $P(u_1)=|PE|/(|PE|+|NE|)$，$P(u_2)=|NE|/(|PE|/|NE|)$

$P(v_j/u_1)=|PE_j|/|PE|$，$P(v_j/u_2)=|NE_j|/|NE|$

其中 PE_j、NE_j 分别是属于 PE、NE 且在 A_k 处取值 v_j 的例子集合.

如何构造训练集，使从训练集得到的 $P(u_i)$ 和 $P(v_j/u_i)$ 与实际问题空间中的相符合或相接近？实际问题中，可以让训练集中正反例的数目尽量多，尽可能的使 $P(v_j/u_i)$ 与实际问题空间的相接近，但如何确定 $P(u_1)$ 和 $P(u_2)$ 呢？也就是如何确定训练中应含多少正例，多少反例呢？正确的做法是使训练集中正、反例的比例等于或接近问题空间中正、反例的比例. 但这一点在很多问题中难于做到，因此也难于正确地算出互信息. 这必然影响预测效果. 这个问题就是所谓类别先验概率问题，有人注意到这个问题，然而没有提出好的解决办法[2]. 我们认为用信道容量来度量特征的重要性就可以解决这个问题[6].

2.2 特征弱相关问题

ID_3 所建判定树的每个非叶结点皆是单个特征. 因此 ID_3 对未知物的分类可以视为一系列判定的组合，每次判定在一个特征上进行，特征间的联系很弱，这就是所谓特征弱相关问题. 此问题也必然影响预测效果. 有人注意到了这个问题，提出了解决办法，参见文[3]. 但他们的解决办法效率低，并且是非符号的，仍然存在知识表示不直观的问题.

我们在 IBLE 方法中构造一棵判定规则树，树中非叶结点是一条由多个特征组成的判定规则，这样，特征间的相关性得到较好体现，特征之间为强相关.

从几何角度来说，若每次判定在一个特征上进行，等于在问题空间中每次利用的判定面都必须与坐标轴之一垂直，产生问题是每次划分都不能使不同类别的例子间距离极大化. 而每次用多个特征组成规则，等于每次利用可以具有任意方向的判定面，能极大化不

同类别例子间的距离.

2.3 知识表示的不直观问题

ID_3以单个特征构造判定树,对于复杂的问题,树中将包含数千个非叶结点,专家看不到熟悉的东西,而实际问题往往都是复杂的.这就是所谓知识表示的不直观问题.Quinlan自己也认为这或许是积木中缺少了最基本的一块[4].知识表示不直观将会影响算法的实际应用.我们利用了信息论的几个概念,再结合模糊判断的思想,将多个特征组成判定规则,使得知识的表示具有较好的直观性.

§3. 示例学习方法 IBLE

人得到一复杂的事物后,若要判定该事物是否符合某个概念,即判定该事物属于u_1类还是u_2类,首先会从分析该事物的特征入手,经过分析会得出三种可能结论:①该事物属于u_1类,②该事物属于u_2类,③不能作出判定,需进一步分析再做结论.在进一步分析时又会出现上述三种情形.对具体的事物,这个过程一直进行到得出具体结论为止.

IBLE就是依据这种思想构造判定规则树的.判定规则树如图1所示,主干上的规则称为主规则,其它规则称为分规则.对一未知实体,首先从树根开始,用规则1对它进行判定,若判别不出则往下一层,一直到第K层的规则k给出类别,假设此处未知实体判成u_1,为了准确起见,还需对它用规则k_1确证一次,若仍判为u_1给出结论,多数确证都维持原结论,只有少数情况会改判为u_2.在判为u1时不用确证就可直接给出结论的情况下,规则k_1为空.在判为u_2时也有类似的工作,不再叙述.

3.1 判定规则树的结构

图1的判定规则树中,每个非叶结点的结构为

规则	左指针	中指针	右指针

左、中、右指针分别指向该结点的左、中、右后继结点.规则形式为:

特　征:$A_1,A_2,\cdots,A_m$

特征值:$v_1,v_2,\cdots,v_m$

逻辑符:$\#_1,\#_2,\cdots,\#_m$

权　值:$w_1,w_2,\cdots,w_m$

阈　值:n_1,n_2

其中逻辑符$\#_i=\{=,\neq\},i=1,2,\cdots,m$.

对任一例子,判定规则如下作出判决:

(1) sum:=0;

(2) if ($A_1\#_1 v_1$) then sum:=sum+w_1;

(3) if ($A_2\#_2 v_2$) then sum:=sum+w_2;

……

(m+1) if ($A_m\#_m v_m$) then sum:=sum+w_m;

(m+2) if (sum$\geqslant n_1$) then 例子为正例;

(m+3) if (sum$<n_2$) then 例子为反例;

(m+4) if ($n_2\leqslant$sum$<n_1$) then 含混,继续判别;

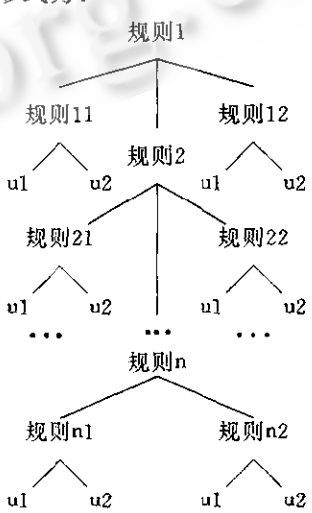

图1 IBLE判定规则树

由此,不难知道规则中各成份的意义及用途.对分规则来说必有 $n_1=n_2$.

判定规则树中,叶结点是一种示意,在存放真正的判定规则树时,并不需要有具体的结点,因此,分规则结点中几个指针皆空.

3.2 IBLE 算法

IBLE 算法如下:

1. 置判定规则树 T 为空;分配一新结点 p;T:=p;
2. 对当前训练集 PE∪NE,利用建规则算法构造主规则;
3. 用主规则测试 PE、NE,得子集 PEY、PEN、PEM、NEY、NEN、NEM;
4. 将主规则放入结点 P;
5. 若(|NEY|≠0) 则 PE:=PEY;NE:=NEY;分配一新结点 w_1;P 左指针指向 w_1;
(1)对当前训练集 PE∪NE,用建规则算法构造左分规则;
(2)将左分规则放入结点 w_1;
6. 若(|PEN|≠0)则 PE:=PEN;NE:=NEN;分配一新结点 w_2;P 右指针指向 w_2;
(1)对当前训练集 PE∪NE,用建规则算法构造右分规则;
(2)将右分规则放入结点 w_2;
7. 若(|PEM|≠0)∧(|NEM|≠0)则 PE:=PEM;NE:=NEM;分配一新结点 w_3;P 中指针指向 w_3;P:=w_3;转 2;
8. 结束.

说明:PEY、PEN、PEM 为 PE 中被当前规则判为是、非、不能判的例子集合,同样 NEY、NEN、NEM 为 NE 中被当前规则判为是、非、不能判的例子集合.

建规则算法如下:

1. 对各特征 A_k
(1)将 A_k 化为分特征 $A_{k1},A_{k2},\cdots,A_{kj}$;
(2)计算 A_k 的各分特征的信道容量,选择最大信道容量的分特征代表 A_k;
(3)A_k 的权值 w_k 等于(代表 A_k 的分特征的信道容量)*1000 取整;
(4)用最大似然译码准则求出代表 A_k 的分特征的取值,换成相应的逻辑符;
2. 取前 m 个信道容量较大的特征组成规则;
3. 对 PE、NE 依据选出的特征求权和,进行统计得出阈值 n_1,n_2;
4. 返回调用处.

说明:建规则算法中将各特征化为分特征的意思是,对于取值数目大于二的特征,将特征的每个值视为一个二值分特征.例如,鸟类描述中有羽毛颜色特征,取值为{白,黑,花},可视为羽毛白、羽毛黑、羽毛花三个分特征,各分特征取值{是,否},用{1,0}表示.这样做的目的是要避免偏向选择多值特征的问题[4].由于特征的各个值对特定的分类问题作用不同,有的作用大,有的作用小.因此选择作用最大的分特征代表特征.这样做同时可以避免将特征化为二值分特征时产生冗余信息增加问题的复杂度.

规则中特征 A_k,特征值 v_k,逻辑符 $\#_k$ 一起构成了类似于 AE_s[5] 中的选择子,[A_k $\#_k v_k$].建规则算法中步骤 1 的(4)步实际就是构造一个选择子.假如求出分特征羽毛黑应取值 1,则逻辑符为"=",构造的选择子为[羽毛颜色=黑].假如求出分特征羽毛黑应取值 0,则逻辑符为"≠",构造的选择子为[羽毛颜色≠黑].IBLE 算法不仅构造了选择

子，而且对选择子在分类中的重要程度进行了量化，赋予了权值.

对复杂程度较低的问题，一条规则就可解决问题，较复杂问题需多条判定规则组成一棵判定规则树才行.

§4. 质谱分类实验

质谱仪是一种化学分析仪器，它以高速电子轰击样品，使分子产生分裂碎片且重新排列，测量这些碎片的质量形成质谱图[7]. 谱图中横坐标为质/荷比 m/e，纵坐标为相对丰度，图中出现的竖线条称为谱峰或简称峰. 质谱解析就是要根据谱图中哪些 m/e 处有峰，哪些 m/e 处无峰以及峰的大小来推测未知物的特性，得到未知物的分子结构式，由于质谱数据量大又伴有噪声，并且质谱测定理论尚不完备，这种分析是很难的，每张谱的分析类似于解一道数学难题[7].

本实验中，以 m/e 为整数的位置为特征，m/e 顺序从 1 到 500，即共有 500 个特征，每个特征取 6 个值且不要求互斥[4]. 实验在 VAX—11/785 上进行，用 FORTRAN 编程，IBLE 的学习和测试程序共 1500 条左右. 对八种类型的化合物进行学习识别，其中前三种类型分别为 WLN 码中含 R、T60TJ 和 QR 的化合物；后五种为芬兰外交部为日内瓦国际裁军会议准备的技术报告中给出的五类有机磷战剂[8]. 从 3 万多张质谱中随机抽取 1 万多张构成八类化合物的训练集和测试集. 学习后，对八类化合物，IBLE 的平均预测率为 93.96%. 为了进一步的研究，还用相同的八类化合物对 ID_3 进行了试验，ID_3 的平均预测率为 81.76%，比 IBLE 的低得多. 而且 IBLE 获得的知识与专家知识在表示和内容上有较高的一致性，专家容易看懂.

§5. 结束语

本文提出的示例学习方法 IBLE 实现简单、学习正确性较高. 所得知识在表示和内容上与专家知识有较高的一致性. 特别适合于处理大规模的学习问题，将其形成系统后可用作专家系统的知识获取工具.

参考文献

1 徐立本、姜云飞，机器学习及应用，吉林大学社会科学丛刊，第 69 期，1988.

2 Giulia Pagallo, David Haussler. Boolean Feature Discovery in Empirical Learning, Machine Learning, 5(1990).

3 Peter de B. Harrington and Kent J. Voorhees, Multivariate Rule Building Expert System, Anal. Chem., Vol. 62, No. 7, 1990.

4 Quinlan, J. R., Induction of Decision Trees, Machine Learning, 1, 1986.

5 洪家荣，示例学习及多功能学习系统 AE_5，计算机学报，2(1989).

6 钟鸣，刘晓霞，用于示例学习的信息理论，计算机研究进展 '92，清华出版社，1992.

7 F. W. Mclafferty, Interpretation of Mass Spectra, Univerity Science Books Mill Valley, California, 1980.

8 Indentification of Potential Organophosphorus Warfare Agents, HELSINKI, 1979.