第17卷 第9期 计 算 机 学 报 Vol. 17, No. 9
1994年9月 CHINESE J. COMPUTERS September, 1994

示例学习的广义扩张矩阵算法及其实现

赵美德 李星原 洪家荣 陈 彬
(哈尔滨工业大学计算机科学与工程系 哈尔滨 150001)

摘 要 本文对扩张矩阵理论加以扩充，提出关于公式的扩张矩阵的概念，并据此实现一个广义扩张矩阵算法叫做 AE_9. 本文还将 AE_9 和 AQ_{15} 应用于几个实际领域的学习问题，如睡眠状态的分类，手写数字识别等，结果都表明 AE_9 比 AQ_{15} 分类精度更高.

关键词 示例学习，知识获取，扩张矩阵.

A GENERALIZED EXTENSION MATRIX ALGORITHM OF LEARNING FROM EXAMPLES AND ITS IMPLEMENTATION

Zhao Meide, Li Xingyuan, Hong Jiarong and Chen Bin
(*Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin* 150001)

Abstract In order to optimize extension matrix algorithm, this paper presents a new version of the extension matrix algorithm called AE_9. To make a comparison with AQ_{15}, AE_9 is tested for several real-world problems, such as learning the decision rules of human sleep stages, and learning the classification rules of unconstrained handwritten numerals. The results show that AE_9 is always more accurate than AQ_{15}.

Keywords Learning form examples, knowledge acquisition, extension matrix.

1 引 言

示例学习是机器学习的一个核心领域，根据它的知识表示可分为两大类——决策树归纳与决策规则归纳[1]. 前者以 ID_3[2] 为代表，其特点是训练与分类速度都很快，适用于大规模的学习问题；后者以 AQ_{15}[3,4] 和 AE_1[5] 为代表，其特点是分类精度高，其知识表达力

本文1993年2月5日收到，修改文1993年8月22日收到. **赵美德**，副教授，获博士学位，主要从事神经网络、机器学习等方面的研究. **洪家荣**，教授，主要从事机器学习、专家系统等方面的研究. **李星原**，讲师，主要从事模式识别和神经网络方面的研究. **陈 彬**，硕士研究生，主要从事基于机器学习的模式识别方面的研究.

强，适合于专家系统的知识自动获取，因而在专家系统领域引起更大的关注. AE_1 以扩张矩阵理论为基础，比 AQ_{15} 训练速度快很多，但是其产生的决策规则有时不如 AQ_{15} 产生的决策规则优化（以其包含的公式的长短和数目度量[5]）. 为了提高扩张矩阵算法的优化程度，本文将扩张矩阵推广到适合于公式，并引入合并（refunion）运算[6]，实现了一个广义扩张矩阵算法 AE_9. 我们还将 AE_9 应用于一些实际领域的学习问题，如人类睡眠六个阶段状态的知识获取，自由手写数字识别规则的自动归纳等，结果表明 AE_9 比 AQ_{15} 产生的规则更优化.

2 基本概念和理论

本文引用文献［7］中的有关概念.

设 d_i，$i=1$，2，…，n 为第 i 个属性 x_i 取值的个数，$D_i=\{0,1,\cdots,d_i-1\}$ 为 x_i 的值域，$E=D_1\times D_2\times\cdots\times D_n$ 是 n 维有穷向量空间，E 中的元素 $e=(v_1,\cdots,v_n)$ 叫做例子，其中 $v_j\in D_j$. 设 PE 和 NE 是 E 的两个子集，为区别起见，分别叫正例集和反例集. 选择子是形为 $[x_j=A_j]$ 或 $[x_j\neq B_j]$ 的关系语句，其中 $A_j\subseteq D_j$，$B_j\subseteq D_j$，并且规定 $[x_j\neq B_j]=[x_j=D_j-B_j]$. 公式是一个选择子或几个选择子的合取式，记为 $L=\bigwedge_{j\in J}[x_j=A_j]$，其 $J\subseteq N$，$N=\{1,2,\cdots,n\}$. 注意：1）对在公式 L 中不出现的属性规定它取值为该属性的值域，即任何 $j\in N$，如果 $j\notin J$，则等价于在 L 上逻辑乘选择子 $[x_j=D_j]$；2）例子可看作公式的一种特殊形式而统一处理，即 $e=(v_1,\cdots,v_n)=\bigwedge_{j\in N}[x_j=\{v_j\}]$. 选择子 $S=[x_i=A_i]$ 覆盖一个公式 $L=\bigwedge_{j\in J}[x_j=A'_j]$ 当且仅当 $i\in J$ 并且 $A'_i\subseteq A_i$. 已知公式 $L=\bigwedge_{j\in J}[x_j=A_j]$ 及 $L'=\bigwedge_{j\in J'}[x_j=A'_j]$，$L$ 覆盖 L' 当且仅当 $J\supseteq J'$ 并且对任何 j，$j\in J$，$A'_j\subseteq A_j$. 一个公式叫做（同已知例子集 $PE\cup NE$）一致的，如果它不覆盖反例集 NE 中的任何反例，一个规则叫做（同已知例子集）一致的，如果它的每一公式都是一致的.

现在，把文献［7］中的对例子的扩张矩阵推广到对公式的扩张矩阵. 以下把 PE 和 NE 当成矩阵处理.

定义 1. 已知反例矩阵 NE 和一个公式 $L=\bigwedge_{j\in J}[x_j=A_j]$. 对 NE 的每一列 $j\in N$，如果 $j\notin J$，则用死元素“*”对 NE 中第 j 列的所有元素作代换；如果 $j\in J$，则用“*”对 NE 中第 j 列属于 A_j 的所有元素作代换. 这样得到的矩阵叫做公式 L 的扩张矩阵，记为 $EM(L)$. 在扩张矩阵中，分别来自不同行的非死元素组成的集合叫做 $EM(L)$ 的一条路.

易见，当定义 1 中的公式用正例 e^+ 代替时，就得到正例 e^+ 的扩张矩阵 $EM(e^+)$.

设有由 4 个反例组成的反例矩阵 NE 及一个公式 L，如图 1 所示，其中 $D_1=\{0,1,2,3,4\}$，$D_2=\{0,1,2\}$，$D_3=D_4=\{0,1,2,3\}$.

文献［8］中关于例子的扩张矩阵的一些性质对公式的扩张矩阵也成立. 例如下面定理成立.

定理 1. 设 $EM(L)$ 是一致公式 L 的扩张矩阵，则存在一个从 $EM(L)$ 中的路到覆盖公式 L 的一致公式的映射. 如果公式 L 覆盖公式 L'，则 $EM(L)$ 中的一条路必定

也是 $EM(L')$ 中的一条路.

	x_1	x_2	x_3	x_4
公式 L	{0,3}	{0,1}	{1,3}	
1	1	1	1	1
2	0	2	2	2
3	3	1	2	2
4	4	0	3	3

NE

(a) 公式 L 与反例矩阵 NE

	x_1	x_2	x_3	x_4
公式	{1,4}	{2}		
1	1	*	*	*
2	*	2	2	*
3	*	*	2	*
4	4	*	*	*

(b) L 的扩张矩阵 $EM(L)$ 箭头经过的路对应于公式 $[x_1\neq 1,4][x_3\neq 2]=[x_1=0,2,3][x_3=0,1,3]$

图 1

证明．　第一部分参看文献 [8] 中的定理 1 之证，其映射方法参看表 1 (b) 之例. 第二部分只需注意 $EM(L)$ 中的非死元素必定是 $EM(L')$ 中的非死元素.

定义 2.　设 $EM(L)$ 是一致公式 L 的扩张矩阵，如果在 $EM(L)$ 中的某一行中只有一个非死元素，则该元素叫必选元素.

由于 $EM(L)$ 的必选元素必在所有覆盖 L 的公式中出现，因此有着重要作用.

定义 3.　已知公式 $L=\bigwedge\limits_{j\in J}[x_j=A_j]$ 及公式 $F=\bigwedge\limits_{j\in J'}[x_j=B_j]$，则将 L 和 F 对应的选择子的取值合并得到一个新的公式，叫做 L 和 F 的合并，记为 $L\oplus F$. 即 $L\oplus F=\bigwedge\limits_{j\in N}[x_j=A_j\cup B_j]=\bigwedge\limits_{j\in J\cap J'}[x_j=A_j\cup B_j]$. 注意，对任何 j，$j\in N$ 但 $j\notin J\cap J'$，$[x_j=A_j\cup B_j]=[x_j=D_j]$ 省略不写.

合并运算取自文献 [6] 的 Refunion. 它具有性质.

定理 2.　公式 L 既覆盖公式 A 又覆盖公式 B，当且仅当它覆盖 $A\oplus B$.

证明．　设 $L=\bigwedge\limits_{j\in J}[x_j=L_j]$，$A=\bigwedge\limits_{j\in J'}[x_j=A_j]$，$B=\bigwedge\limits_{j\in J''}[x_j=B_j]$.

必要性：因 L 覆盖 A，则 $L_j\supseteq A_j$，对 $j\in J$；因 L 覆盖 B，则 $L_j\supseteq B_j$ 对 $j\in J$. 因此对任何 $j\in J$，$L_j\supseteq A_j\cup B_j$，即 L 覆盖 $A\oplus B$.

充分性：因 L 覆盖 $A\oplus B$，则对任何 $j\in J$，$L_j\supseteq A_j\cup B_j$，因而 $A_j\subseteq L_j$ 且 $B_j\subseteq L_j$，即 L 覆盖 A 且 L 覆盖 B.

3　广义扩张矩阵算法 AE_9

广义扩张矩阵算法 AE_9 的基本步骤是：

1. 按文献 [7] 中的方法从正例集 PE 中选择一个种子 e. $F\leftarrow e$. path$\leftarrow\Phi$. CPE$\leftarrow\Phi$.

2. 做 F 的扩张矩阵 EM（F）. 如果有必选元素则放入 path 中，同时删去 NE 中该必选元素出现的行（反例），如果 NE 空则终止；如非空则删去 PE 和 CPE 中出现该必选元素的对应行，重复执行 1 直至 EM（F）中不存在必选元素为止.

3. 如果 PE 非空，检查 PE 中的每一正例，看它与 F 的合并是否是一致公式，如不是则从 PE 中删去该正例；若是则保留一个覆盖正例数目最多的一个合并并取代 F，将 PE 中被 F 覆盖的正例放入 CPE 中. 重复步骤 2 和 3，直至 PE 变空.

4. 如果 PE 空而 CPE 非空，则检查 CPE 中的每一个正例，看它与 F 合并后是否为一致公式，若不是则从 CPE 中删去该正例；若是则生成合并公式，保留一个覆盖最多正例的合并并取代 F，从 CPE 中删去被新的 F 覆盖的正例，重复 2.4 直至 CPE 变空.

5. 此时 PE 和 CPE 均空，但 NE 非空，做 EM（F），将其中含有最多非死元素的列中的非死元素放入 path 的相应列中，并从 NE 中删去含这些非死元素的行，重复这一过程，直到 NE 变空. 这时将 path 转变为相应的公式.

AE_9 的思想是，第 2 步选择 EM（L）中的必选元素；第 3 步是寻找覆盖最多数目正例（包括种子）的合并 F，然后重复第 2 步使 EM（F）产生新的可能的必选元素；第 4 步的目的是由于第 3 步中最后选定的合并 F，可能不覆盖它的祖先 F 已覆盖的正例，这些正例在选这些祖先时已被存入 CPE，现在再检查一下在 CPE 中存放的被这些祖先覆盖的正例是否还被当前的 F 所覆盖，如果是则删除（不再存入 CPE 中），否则进行合并；第 5 步是对 NE 中含非必选元素的行进行处理，选择涉及较少数目的列，可能使生成的公式较短[7].

AE_9 的优点是每次试图寻找覆盖数目最多的一致公式，这样有可能使覆盖整个正例集的一致公式的总数较少. 但这并不能保证 AE_9 总能找到最优解. 事实上，对这个问题的最优解之一是两个公式，即 $[x_1=0, 1, 3][x_2=0]$ 覆盖 e_1, e_2, e_7，及 $[x_1=0, 2, 3, 4][x_2=1, 2][x_3=0, 1, 3]$ 覆盖 e_3, e_4, e_5, e_6.

4 结果比较

我们举两个实际问题，将 AE_9 同 AE_5[7] 和 AQ_{15}作一比较.

（1） 人睡眠状态的知识获取

这里所提供的数据来自美国 Illinois 大学[4]. 第一到第六阶段的例子数分别为 115、103、688、158、38 和 134，共 1236 个例子，每例有 11 个特征. 表 1 列出了 AE9 及 AE5 和 AQ15 得到的每阶段的规则数和规则总数. 从表 1 可以看出 AE9 得到的每一阶段的规则数是最少的，且六阶段的规则总数为 53，而 AQ15 为 63，AE5 为 85，其精度与 AE5 相比提高了约 38%，与 AQ15 相比提高了 约 16%.

（2） 自由手写数字识别

这里将 AE9 应用于自由手写体数字分类的知识获取，并与 AQ15 进行了比较[9]. 首先对扫入的图象进行分割、预处理、归一化和细化，然后提取包括外部轮廓、几何和笔划密度在内总共 29 个特征，用 6000 个不同人书写的数字作为训练例子. 表 2 列出了 AE9 和 AQ15 对 0—9 十个数字进行归纳学习的每个数字的规则数和规则总数. 从表 2 我们看

出 AE9 得到的每一数字的规则数都少于 AQ15，AE9 的总规则数为 96，而 AQ15 为 117，公式数减少了约 18%. 根据 AE9 和 AQ15 抽取的规则，我们对 4000 个测试例子进行测试，AE9 的识别率、拒识率和误识率分别为 98%、1.8%和 0.2%，AQ15 的识别率、拒识率和误识率分别为 96%、3%和 1%. 由此可见，与 AQ15 比较，AE9 具有更好的普化能力. 此外，每个数字的识别速度也提高了约 6%.

表 1　AE9 与 AE5 和 AQ15 运行结果比较

系　统	阶　段						
	1	2	3	4	5	6	总数
AE9	8	10	13	10	8	4	53
AE5	15	15	20	17	11	6	84
AQ15	10	11	15	13	9	5	63

表 2　AE9 与 AQ15 结果比较

系　统	数　字										
	0	1	2	3	4	5	6	7	8	9	总数
AE9	4	5	10	8	17	14	5	11	9	13	96
AQ15	5	6	13	11	23	16	5	13	9	16	117

5　结　论

本文对扩张矩阵算法加以推广，并结合合并运算，实现一个广义扩张矩阵算法 AE9，运行结果表明，AE9 比 AQ15 产生的决策规则更优化，有较大的实用与推广价值.

参 考 文 献

[1] Quinlan J R. Knowledge acquisition from structured data. *IEEE Expert*. 1991, 6 (6): 32—37.

[2] Quinlan J R. Induction of decision tree. *Machine Learning*, 1986, 1 (1): 81—106.

[3] Hong J R *et al*. AQ15: Incremental Learning of Attribute-Based Descriptions from Examples, the Method and User's Guide. Dept of Compat Sci, Vniv of Illinois: Rept ISG 86-5 UIUCDCS-F-86-949, 1986.

[4] Michalski *et al*. R S Multi-purpose incremental learning system AQ15 and its testing application to three medical domains. In: Proceedings of the Fifth AAAI, 1986, 1041—1045.

[5] Hong J R. AE1: Extension matrix approximate method for the general covering problem. Int Journal of Computer & Information Science, 1985, 14 (6): 421—437.

[6] Michalski R S, Stepp R E. Learning from observation: conceptual clustering. In: Michalski R S Machine Learning: An Artificial Intelligence Approach, Palo Alto, CA: Tioga Pub Co, 1983.

[7] 洪家荣. 示例学习及多功能学习系统 AE5. 计算机学报，1989，12 (2).

[8] 洪家荣. 示例学习的扩张矩阵理论. 计算机学报，1991，14 (6).

[9] 洪家荣等. 示例式学习系统 AQ15 同神经网络在手写体数字识别应用中的比较. 见：第三届全国机器学习研讨会论文集，1991.