第 34 卷 第 8 期 计 算 机 工 程 2008 年 4 月
Vol.34 No.8 Computer Engineering April 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)08—0218—03 文献标识码：A 中图分类号：TP18

基于 Pareto 强度值演化的 FPRs 知识表示参数优化

安素芳，柴变芳，傅 玥，才秀凤
(石家庄经济学院信息工程学院，石家庄 050031)

摘 要：提出一种新的参数优化模型和求解算法，引入模糊熵来指导模糊产生式规则(FPRs)的参数优化，给出基于极大模糊熵定理的参数优化模型，提出求解该模型的 Pareto 强度值的演化算法。实验结果表明，该方法能够有效优化参数，一定程度上避免过度拟合，提高了 FPRs 的知识表示能力。

关键词：模糊产生式规则；知识表示参数；极大模糊熵定理；Pareto 强度值演化

Knowledge Representation Parameters Optimization of FPRs Based on Pareto Strength Evolution

AN Su-fang, CHAI Bian-fang, FU Yue, CAI Xiu-feng
(College of Information Engineering, Shijiazhuang University of Economics, Shijiazhuang 050031)

【Abstract】 This paper introduces fuzzy entropy into the procedure of exploring parameters of Fuzzy Production Rules(FPRs). A parameter optimization model based on maximum fuzzy entropy principle is proposed and a Pareto strength evolutionary algorithm is introduced to solve this model. Experimental results show that the trained parameters gained from above strategy are highly accurate, therefore this method can decrease the phenomenon of over-fitting and improve the knowledge representation capability of FPRs.

【Key words】 Fuzzy Production Rules(FPRs); knowledge representation parameter; maximum fuzzy entropy principle; Pareto strength evolution

现实世界中的许多知识难以用精确的方法来描述。模糊产生式规则(Fuzzy Production Rules, FPRs)是对这种模糊知识的常用表示方法。标准的 FPRs 是一种 IF-THEN 结构的模糊条件句。近年来，人们针对 FPRs 本身的特点提出了在 FPRs 中引入局权(local weight)、全权(global weight)、阈值、确定性因子等知识表示参数来增强其知识表示能力。因此，寻求合理高效的算法来优化这些知识表示参数，成为该研究领域的一个重要课题。大多数算法的共同思想是通过提高训练精度来指导知识表示参数的优化，这在一定程度上优化了知识表示参数，但同时会引起过度拟合(over-fitting)。

1 FPRs 系统

FPRs 是使用最多的模糊、不确定知识表示形式。为了提高 FPRs 的知识表示能力，文献[1]引入了确定性因子和阈值，用来表示模糊推理方法的合理性。文献[2]引入了局权来表示规则前件的不同因子对于该规则结论的相对重要程度。文献[3]引入全权表示为了达到最终推理结果，每个规则所做的相对重要程度。本文讨论的 FPRs 仅涉及局权、全权 2 个知识表示参数。

当 FPRs 用于表示模糊分类问题的模糊知识时，一条 FPR 可表示为

R ：IF (V_1 is A_1) [Lw_1]AND…AND (V_n is A_n)[Lw_n] THEN (U is B)[Gw]

其中，$V_1, V_2, \cdots, V_n$ 和 U 是变量；n 表示 FPR 前件的因子的个数；$A_1, A_2, \cdots, A_n$ 和 B 是模糊集表示的变量值；Lw_1，$Lw_2, \cdots, Lw_n$ 表示规则前件的局权；Gw 表示规则 R 的全权，局权和全权的定义域为(0,1)。

设模糊规则集 $S=\{R_i, i=1,2,\cdots,m\}$，其中，规则集中的任一元素$R_i$的结论为 U is c。则基于FPRs的系统模型可表示为

规则 R_i ：IF ($V_1^{(i)}$ is $A_1^{(i)}$) [$Lw_1^{(i)}$]AND…AND ($V_n^{(i)}$ is $A_n^{(i)}$)[$Lw_n^{(i)}$] THEN (U is c)[Gw_i]

示例：(V_1 is A_1'),(V_2 is A_2'),…,(V_n is A_n')

结论：(U is c), d

其中，$Lw_j^{(i)}(j=1,2,\cdots,n)$ 表示规则 R_i 的局权；Gw_i 表示规则 R_i 的全权；c 表示为分类结果；d 表示推理结论的确定性因子，确定性因子的定义域为(0,1)。根据已知示例，本文采用基于相似度的推理算法[4]匹配FPRs集，得到结论：(U is c),d。

(1)计算示例中的每个属性值 $A'_j(j=1,2,\cdots,n)$ 与规则 R_i 的前件因子 $A_j^{(i)}(j=1,2,\cdots,n)$ 的隶属度作为相似度，表示为

$$SM_j^{(i)} = A_j^{(i)}(A'_j) \tag{1}$$

(2)计算示例与规则 R_i 前件总的相似度为

$$SM_i = \mathrm{Min}_{1\le j\le n}(Lw_j^{(i)} \times SM_j^{(i)}) \tag{2}$$

(3)推理结果为 c，且推理结果的确定性因子为 d ：

$$d = \frac{Gw_i}{G} \times SM_i \tag{3}$$

其中，$G=\sum_{i=1}^{m} Gw_i$。

设 K 类分类问题将模糊规则分为 K 个子集 S_k $(k=1,2,\cdots,K)$，其中，S_k 是由推理结果为 k 类的 FPRs 组成。

作者简介：安素芳(1980－)，女，助教、硕士，主研方向：机器学习，数据挖掘；柴变芳、傅 玥，助教、硕士；才秀凤，助教、硕士研究生

收稿日期：2007-05-17 **E-mail：**ansf@mail.hbu.cn

一待分类的示例匹配这 K 组 FPRs，采用上述推理算法，得到结论 $(k,d_k),k=1,2,\cdots,K$。当要求示例仅属于某一类时，其分类结果为 $(\max,d_{\max})$，其中，$d_{\max}=\text{Maximum}_{1\le k\le K}\{d_k\}$，max 为 $d_{\max}$ 的类标签。

2 基于极大模糊熵定理的参数优化模型

在信息论中，熵[5]用于表征观察空间的平均信息量；在集合论中，模糊熵[6]用于度量模糊集合的模糊程度；在模糊推理中，引入模糊熵[7]度量模糊推理结果的模糊程度。模糊熵的定义见文献[6-7]。

设存在 K 个 FPRs 集 $S_k(k=1,2,\cdots,K)$，$F=\{f_i,i=1,2,\cdots,q\}$ 表示待分类训练集；当训练集中的第 $i-th$ 个示例 f_i 匹配 S_k 中的模糊规则，得到分类结果 $(k,d_k^{(i)})$，则示例 f_i 的推理结果的模糊熵可以表示为

$$e(f_i)=1-\frac{D_r(A,A^c)}{\sqrt[p]{K}} \tag{4}$$

其中，D_r 定义如下：

$$Dr(A,A^C)=\sqrt[p]{\sum_{k=1}^{K}\left|d_k^{(i)}-(1-d_k^{(i)})\right|^p},p\in\{1,2,\cdots\} \tag{5}$$

分类训练集 F 的模糊熵可以表示为

$$e(F)=\frac{1}{q}\sum_{i=1}^{q}e(f_i) \tag{6}$$

给定一数据集，将其随机地分为训练集和测试集。根据模糊规则抽取算法，可以由训练集抽取出 FPRs。在 FPRs 中加入知识表示参数，此时用测试集中的示例匹配该 FPRs，依据基于相似度的推理算法，得到示例的推理结果。目前常用的知识表示参数的优化准则为提高训练精度，但该方法容易引起过度拟合。本文采用模糊熵衡量模糊推理结论的模糊程度，采用极大模糊熵定理指导知识表示参数的优化。

极大模糊熵定理[7-8]：在运用模糊概念进行推理的过程中，用模糊集A表示推理结果。对于给定的一个示例，当知识表示参数取不同值时，用模糊集A表示推理结果也不同。本文偏向在满足已知条件约束下，使得该模糊集的熵最大的那个模糊集作为推理结果：

$$\text{Maximum}\ \ e(F) \tag{7}$$

$$\text{s.t.}\ \ d_{\max}^{(i)}=\text{Maximum}_{1\le k\le K}\{d_k^{(i)}\} \tag{8}$$

3 基于Pareto强度值的演化算法[9]

基于极大模糊熵定理的 FPRs 的参数优化模型，本质上是在式(8)的条件约束下，搜索一组参数，使得式(7)取得最大值的优化问题，本文采用基于 Pareto 强度值的演化算法，以降低搜索的复杂度、提高搜索的速度。其基本思想是将约束优化问题转化为 2 个目标优化问题，其中一个为原问题的目标函数，另一个为违反约束条件的程度函数。同时定义个体 Pareto 强度指标，根据强度指标，对上述 2 个目标函数组成的向量进行排序，使用演化算法求解原问题。

设 K 类分类问题 $\{x_1,x_2,\cdots,x_n\}$ 为知识表示参数向量，当训练集中的示例 f_i 匹配 FPRs，得到分类结果 $(k,d_k^{(i)}),(k=1,2,\cdots,K)$，用 $g(x_1,x_2,\cdots,x_n)$ 表示训练集的模糊熵，则 $g(x_1,x_2,\cdots,x_n)=e(F)$。

当参数向量中的 $\boldsymbol{x}_j=1(j=1,2,\cdots,n)$ 时，FPRs就相当于不加知识表示参数。训练示例 f_i 匹配这些FPRs时，得到分类结果 $(k,c_k^{(i)}),(k=1,2,\cdots,K)$。设训练示例 f_i 的正确分类为 j 类，那么训练集中能够被这些FPRs正确分类的示例组成的集合，记作 $B=\{i\mid 1\le i\le q,c_j^{(i)}=\text{Maximum}_{1\le k\le K}\{c_k^{(i)}\}\}$。

很难保证在训练集上所有示例满足约式(8)，因此选取训练集中的部分示例满足条件，从而能够保证训练精度不减。此时的规划问题可转化为

$$\text{Maximum}\ \ g(x_1,x_2,\cdots,x_n) \tag{9}$$

$$\text{s.t.}\ \ d_{\max}^{(i)}=\text{Maximum}_{1\le k\le K}\{d_k^{(i)}\},i\in B \tag{10}$$

设 $\{x_1,x_2,\cdots,x_n\}\in F'\subset S'$，其中，$S'$ 表示目标函数的搜索空间；F' 表示可行区域。定义 $s_1(\boldsymbol{x})=g\{x_1,x_2,\cdots,x_n\}$，$s_2(\boldsymbol{x})=\left|d_{\max}^{(i)}-\text{Maximum}_{1\le k\le K}\{d_k^{(i)}\}\right|$，那么优化问题可以转化为 $\text{Maximum}\ \ y=(s_1(\boldsymbol{x}),s_2(\boldsymbol{x}))$。

定义 1 $a\in S'$，$b\in S'$，称 a Pareto b (记为 $a\prec b$)，当且仅当 $\forall i\in\{1,2\}:s_i(a)\le s_i(b)$ 且 $\exists j\in\{1,2\}:s_j(a)<s_j(b)$。

定义 2 设 x_i 为群体 p_t 中的一个个体，用 $S(x_i)$ 表示群体中 Pareto 劣于 x_i 的个体个数，称为 x_i 的强度值，即

$$S(x_i)=\#\{x_j\in p_t,\text{且}\ x_i\prec x_j\} \tag{11}$$

其中，# 表示集合的基数。

基于 Pareto 强度值演化算法如下：

(1)初始化，种群规模为 N，置 t=0；

(2)从第t代群体p_t中随机选择 μ 个父体。

(3)对 μ 个父体使用杂交算子产生 λ 个后代。

(4)从 λ 后代中，选择 2 个后代，其中一个为 λ 后代中 Pareto 强度值最大的个体(计算 Pareto 值可根据式(11))；若强度值最大的个体不唯一，则比较它们的违反约束函数，即计算 $s_2(x)$，$s_2(x)$ 值小的个体优先。另一个父体为剩下的 λ -1 个后代中约束条件最小的个体。

(5)重复(2)~(4)，直到选取N个后代为止，由这几个后代将群体p_t中的个体整体替换掉；

(6)若满足停机条件则停止，否则 t=t+1，转(2)。

4 实验结果与分析

选取机器学习标准数据库UCI[10]数据库中的 5 个数据库进行实验。分别为Pima diabetes data, Glass identification, Mango leaf data, Ecolic data和Wine data。

实验步骤如下：

(1)模糊化已选数据库；

(2)模糊化后的数据库随机分为 2 部分，其中 70%的示例作为训练集，30%的示例作为测试集；

(3)将数据库训练集均通过模糊决策树[3]算法，产生一组 FPRs；

(4)设参数均为 1.0，采用基于相似度的推理算法，获取集合 B，记录训练精度和测试精度；

(5)采用基于 Pareto 强度值的演化算法，优化参数。设 N=20, μ=16, λ=4，系统终止条件为循环次数 1 000 次；

(6)根据(5)中求解的参数，使用基于相似度的推理算法，求得优化参数后的训练精度和测试精度；

(7)重复(5)~(6)10 次，求得优化参数后的训练精度和测试精度的平均值。

实验分析如下：

表 1 列举了部分 UCI 数据库的实验结果。由表可见基于极大模糊熵定理的参数优化模型，保证了训练精度不降低，明显提高 FPRs 的测试精度，从而增强了 FPRs 的知识表示能力。以 Ecoli data 数据库为例，未加参数时训练精度为 0.756 2，测试精度为 0.686 3；而通过本文采用的参数优化模型，训练精度提高到 0.801 0，测试精度提高为 0.712 5。采用 Glass identification 和 Pima diabetes data 数据库的训练精度和测试精度都提高了，测试精度分别提高了 4.31%和 2.71%。

表 1　优化参数前后训练精度、测试精度的比较

数据库	分类类别	未加参数时的训练精度	未加参数时的测试精度	加参数后的训练精度	加参数后的测试精度
Pima diabetes data	2	0.721 8	0.762 6	0.754 5	0.789 7
Glass identification	6	0.756 8	0.731 5	0.785 9	0.774 6
Mango leaf data	2	0.725 6	0.720 0	0.761 6	0.731 0
Ecoli data	5	0.756 2	0.686 3	0.801 0	0.712 5
Wine data	3	0.760 9	0.735 6	0.791 2	0.784 5

常用的参数优化方法是依据训练精度的提高来优化知识表示参数。该方法的缺点是会引起过度拟合。以 Pima diabetes data 为例，图 1 显示了训练精度和测试精度关系。可以看出，训练精度高于 0.723 时，极大化训练精度将导致测试精度降低，这印证了依据常用准则来优化，会引起过度拟合。

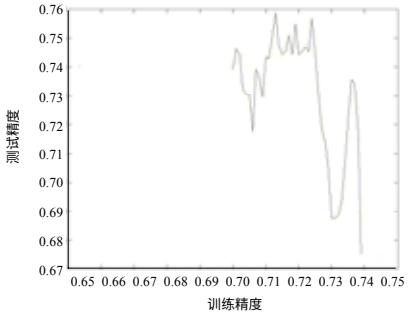


图 1　测试精精度与训练精度的关系

5　结束语

本文的方法在一定程度上克服了传统方法通过提高训练精度来获得知识表示参数值而引起的过度拟合的不足。如何降低知识表示参数求解算法的时间复杂度、如何更有效地提高 FPRs 的知识表示能力是进一步研究的方向。

参考文献

[1] Yeung D S, Tsang E C C. Improved Fuzzy Knowledge Representation and Rule Evaluation Using Petri Nets and Degree of Subsethood, Internet[J]. Journal of Intelligent Systems, 1994, 12(9): 1083-1100.

[2] Tseng H C, Teo D W. Medical Expert System with Elastic Fuzzy Logic[C]//Proc. of the 3rd IEEE Internet Conference on Fuzzy System. Osaka, Japan: [s. n.], 1994.

[3] Yeung D S, Tsang E C C. Weighted Fuzzy Production Rules[J]. Fuzzy Sets and Systems, 1997, 88(3): 299-313.

[4] 王永庆. 人工智能定理与方法[M]. 西安: 西安交通大学出版社, 2003.

[5] 喻传赞. 熵和信息与交叉学科[M]. 昆明: 云南大学出版社, 1994.

[6] Liu Xuecheng. Entropy, Distance Measure and Similarity Measure of Fuzzy Sets and Their Relations[J]. Fuzzy Sets and Systems, 1992, 52(3): 305-318.

[7] 郭方芳, 陈图云, 夏尊铨. 基于极大模糊熵定理的模糊推理三 I 算法[J]. 模糊系统与数学, 2003, 17(4): 55-59.

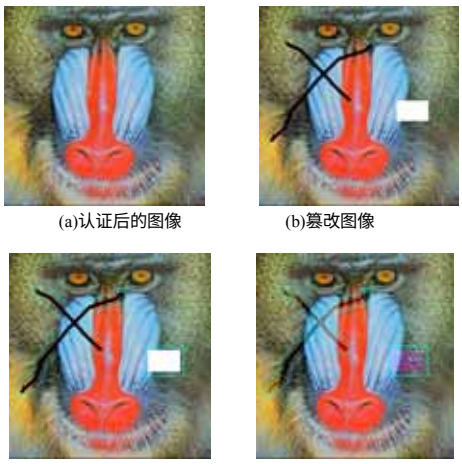
[8] Zhao Jian, Wang Xiaolong. Chinese POS Tagging Based on Maximum Entropy Model[C]//Proc. of IEEE International Conf. on Machine Learning and Cybernetics. Beijing, China: [s. n.], 2002.

[9] 周育人, 李元香, 王　勇, 等. Pareto 强度值演化算法求解约束优化问题[J]. 软件学报, 2003, 14(7): 1243-1249.

[10] UCI Repository of Machine Learning Databases and Domain Theories[Z]. (2006-08-12). ftp://ftp.ics.uci.edu/pub/machine-learning-databases.

(上接第 217 页)

改部位的进行恢复后，由于仅恢复了高两位的信息，因此恢复后的部分颜色会有一些变化，但是还是可以看到被篡改部位原始图像的大致情况。



(a)认证后的图像　(b)篡改图像

(c)篡改检测与定位　(d) 自恢复图像

图 5　实验结果

4　结束语

本文提出一种基于卷积码的彩色图像认证算法。该算法可以扩展到 JPEG 图像的 DCT 或 DWT 域，如果要在频域中的低频系数中嵌入认证信息，可以将原始彩色载体图像分解为 R,G,B 3 个彩色分量，并分别实施 DCT 或小波变换，然后对低频系数进行等长 8 位量化，分解为 8 个位平面。实验结果表明了该算法的有效性。

参考文献

[1] 吴金海, 林福宗. 基于数字水印的图像认证技术[J]. 计算机学报, 2004, 27(9): 1153.

[2] 任　娟, 王蕴红, 谭铁牛. 基于感兴趣区域的图像认证与自恢复算法[J]. 自动化学报, 2004, 30(6): 833-843.

[3] Wong P W, Memon N. Secret and Public Key Image Watermarking Schemes for Image Authentication and Ownership Verification[J]. IEEE Transactions on Image Processing, 2001, 10(10): 1593-1601.

[4] Viterbi A. Error Bounds for Convolutional and an Asymptotically Optimum Decoding Algorithm[J]. IEEE Transactions on Information Theory, 1967, 13(12): 260-269.