第 27 卷第 5 期　　电 子 与 信 息 学 报　　Vol.27No.5
2005 年 5 月　　Journal of Electronics & Information Technology　　May 2005

基于 AdaBoost 的改进模糊分类规则集成学习

方　敏[*,**]　　王宝树[**]

[*](西安电子科技大学综合业务网国家重点实验室　西安　710071)

[**](西安电子科技大学计算机学院　西安　710071)

摘　要：基于集成学习提出了一种新的模糊分类规则的产生算法。将分类规则的前件、后件模糊化，在自适应提升(Adaptive Boosting, AdaBoost)算法的迭代中，调整训练实例的分布，利用遗传算法产生模糊分类规则。并在规则学习的适应度函数中引入训练实例的分布，使得模糊分类规则在产生阶段就考虑相互之间的协作，产生具有互补性的分类规则集。从而改善了模糊分类规则的整体识别能力，提高了分类识别精度。

关键字：模糊分类规则，AdaBoost 算法，分类器集成

中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2005)05-0835-03

Advance Ensemble Learning of Fuzzy Classification Rules Based on AdaBoost

Fang Min[*,**]　　Wang Bao-shu[**]

[*](*National Key Laboratory of Integrated Services Networks, Xi'an* 710071, *China*)

[**](*Institute of Computer Science, Xidian University, Xi'an* 710071,*China*)

Abstract　A new learning algorithm of fuzzy classification rules is presented based on ensemble learning algorithm. By tuning the distribution of training instances during each AdaBoost iterative training, the classification rules with fuzzy antecedent and consequent are produced with genetic algorithm. The distribution of training instances participate in computing of the fitness function and the collaboration of rules which are complementary is taken into account during rules producing, so that the classification error rate is reduced and performance of the classification based on the fuzzy rules is improved.

Key words　Fuzzy classification rule, Adaptive Boosting (AdaBoost) algorithm, Classifiers ensemble

1　引言

模糊规则作为一种描述不确定知识的有效方法，在模式识别等领域被证明十分有效。在模式识别中，通过选择合理的规则集可以完整地描述模式的特征属性，并且通过推理来完成模式分类。但是模糊系统缺乏从训练样本集中提取分类知识的能力，因此，过去几年里的研究致力于采用进化算法和神经网络提高模糊系统的分类能力[1]，它利用进化算法调整模糊规则的参数，形成基于模糊规则的分类系统。

通常的模糊分类规则的获取方法，往往首先对模式空间进行划分，对于每种划分训练出一套分类规则。假设特征空间的维数为 m，如果将每一维划分为 s 个模糊子集，就会有 s^m 种划分，划分数巨大，算法比较耗时。由于不同的划分可能有重叠的或部分重叠的模糊子集，它们构成的规则就可能存在包含关系。另外还有些划分出来的模糊子集可能覆盖不到任何样本，所构成的模糊分类器中，存在着大量的冗余规则。

相比之下，集成学习法在初始模式空间划分下，利用输入属性集在不同分布下的重复取样，产生不同分类器的集成，比试图研究单个最优分类器更加容易达到同样的分类精度[2]。这种方法的研究近几年来得到了机器学习领域研究人员的重视，AdaBoost(Adaptive Boosting)算法在这方面作用更为显著[3]。

本文首先将分类规则的前件、后件模糊化，扩大分类界面的变化范围；然后提出了在规则学习的适应度函数中引入学习样本的分布作为适应度函数中实例样本的权值因子，基于 AdaBoost.M2 迭代的模糊多输出分类规则学习算法。在迭代规则学习中，利用进化模糊规则产生算法，根据训练实例

2003-11-21收到，2004-09-02改回
国家部级基金(413150801)和综合业务网国家重点实验室开放基金(ISN6-7)资助课题

的分布调整分类规则的模糊后件，得到分类规则集，融合各分类规则的输出达到提高整体分类精度的目的。文中将这种集成进化学习算法用于解决 Iris 数据集的分类识别问题。实验说明，该方法能够很好地提取分类规则，并有效地提高分类精度。

下面首先描述基于模糊规则的分类系统，在此基础上，研究了基于 AdaBoost.M2 算法的改进模糊分类规则集成学习算法，并对其分类效果进行分析。

2 基于模糊规则的分类系统

设 $X=\Omega$ 是输入论域，$Y=\{c_1,c_2,\cdots,c_m\}$ 是输出论域。输入论域是指待分类的模式集合，即模式空间；输出论域为模式类别的集合，模式空间的模式数为 m，模式类分别标记为 $c_1,c_2,\cdots,c_m$。假设输入模式为欧氏空间的一个矢量，$x^i=(x_1^i,x_2^i,\cdots,x_n^i)\in\Omega^n$，模式 x^i 的特征用模糊集描述，分类则用一组规则来实现，规则集中包含 L 条模糊规则 R=$\{r_1,r_2,\cdots,r_L\}$。下面构造一个基于模糊规则的分类器。如果模式 ω 具有属性或特征 A_i，那么 ω 是：$U_i=\{\beta_{i1}/c_1,\beta_{i2}/c_2,\cdots,\beta_{im}/c_m\}$，其模糊规则具有以下形式：

$$\text{if } (x_1^i \text{ is } A_{k1}\wedge x_2^i \text{ is } A_{k2}\wedge\cdots\wedge x_n^i \text{ is } A_{kn}) \text{ then}$$
$$x^i \text{ is } (\beta_{i1}/c_1,\beta_{i2}/c_2,\cdots,\beta_{im}/c_m)$$

其中 A_k 是输入论域 X 上的模糊子集；U_i 描述了该模式 ω 属于 c_1 的可能性为 β_{i1}，属于 c_2 的可能性为 β_{i2}，属于 c_m 的可能性为 β_{im}，其中 $\beta_{ij}\ge 0$，且有 $\sum_{j=1}^{m}\beta_{ij}=1$；权值 $W_i>0$，表示规则 r_i 的强度或可信度[4]。

3 改进的模糊分类规则集成学习

AdaBoost 是 Freund 在 1995 年提出的自适应提升算法(Adaptive Boosting)。该方法构造多个基本分类器并集成到一个组合分类器。训练中，通过对原始训练集进行概率分布估计 D_t，使得基本分类器越来越集中到那些难分的训练样本上，并在每一轮，使用这个分布来训练分类器。AdaBoost 需要不稳定的分类器，不稳定的分类器是指对训练实例较敏感。下面给出模糊分类规则集成学习算法，它利用遗传算法和 AdaBoost 算法相结合产生基分类规则，主要思想是：将每个分类规则看作一个不完整的、弱的基分类器[5, 6]，即每分类规则都能够对其前件覆盖的实例进行分类，但不能对其它训练实例进行预测分类。利用迭代法重复激活进化的规则产生算法来产生新规则，在集成训练过程中，被已有分类规则前件充分覆盖的训练实例被赋以小的权值，未被覆盖的或被错分的训练实例被赋予大的权值，使进化操作集中在这些实例上。不断完善分类规则集。

3.1 模糊分类规则集成学习算法

本文目的采用重复激活遗传模糊规则产生算法迭代产生规则集，遗传模糊系统识别能够正确分类及最优匹配当前训练实例的模糊规则，利用 AdaBoost 算法在规则产生阶段促进模糊规则之间的协作。

由于分类规则的输出需要提供比类别标号更多的信息，可以将分类规则的输出看作类别的后验概率逼近，所以利用 AdaBoost.M2 训练基分类规则，它可以处理分类器对每类信任值的计算。在训练第 t 个分类器的结果时，输出假设为 h_t: $X\times Y\to[0,1]$，并定义分布 $D_t(x^i,y)$ 为错配类别标号集合上的分布。通过改变这个分布使得下一轮学习不仅集中于难分类的样本，并且提高正确分类和错误分类的差异。

算法 基于 AdaBoost.M2 的模糊分类器训练

输入训练样本：$A=\{(x^1,y^1),(x^2,y^2),\cdots,(x^n,y^n)\}$，标号 $y^k\in Y=\{1,2,\cdots,m\}$，$n$ 为训练样本个数；

初始化：$B=\{(x^k,y)\mid k\in\{1,2,\cdots,n\},y\ne y^k\}$，分布 $D_1(x^k,y)=1/[n(m-1)]$，$(x^k,y)\in B$，置 R 为空；For $t=1,\cdots,L$ //L 为迭代次数。

(1) 基于分布 D_t 训练模糊分类规则 r_t。基分类规则的学习采用遗传模糊分规则的训练法，通过最大化正确分类数和错误分类数差值进行训练。其中：

(a) 编码方式：若训练样本 $A=\{(x^1,y^1),(x^2,y^2),\cdots,(x^n,y^n)\}$，标号 $y^k\in Y=\{c_1,c_2,\cdots,c_m\}$，$n$ 为训练样本个数；每个个体是由 K_t 条规则组成，个体基因串的编码为：

$$(A_1,\beta_{11},\beta_{12},\cdots,\beta_{1m},W_1)\cdots(A_r,\beta_{r1},\beta_{r2},\cdots,\beta_{rm},W_r),\quad r=1,2,\cdots,K_t \tag{1}$$

(b) 适应度函数：遗传算法需要根据个体适应度函数的值进行搜索，在每一个学习周期，适应度较低的个体将被淘汰，适应度最高的个体被看作满意解。适应度定义为规则 r_i 覆盖的具有类标签 c_i 的训练实例数与具有类标签 c_i 的训练实例数的比值与该规则覆盖的错分样本所占比值的差，见式(2)，即正确分类的样本数越大、错误分类的样本数越小，则适应度函数越高。

$$\text{适应度}=\Big(\sum_{k:\ y=y^i}D_t(x^k,y)\mu_{r_i}(x^k)\Big)\Big/\sum_{k:\ y=y^i}D_t(x^k,y)-\Big(\sum_{k:\ y\ne y^i}D_t(x^k,y)\mu_{r_i}(x^k)\Big)\Big/\sum_{k}D_t(x^k,y)\mu_{r_i}(x^k) \tag{2}$$

其中 $D_t(x^k,y)$ 反映了训练实例在训练集中的权值，由 AdaBoost.M2 算法调整训练实例分布确定，因此，适应度函数中考虑了训练实例权值的影响。

(c) 遗传算法的算子：使用了遗传算法的 2 种基本算子：选择和变异。

(2) 计算模糊分类规则的差错率。模糊规则 r_t 的分类差错 $E(r_t)$ 由第 k 个训练实例 (x^k,y^k) 与规则前件的匹配度 μ^k

以及训练实例的权值 $D(x^k,y^k)$ 共同确定，见式(3):

$$\varepsilon_t=\frac{\sum_{(x^k,y)\in B} D_t(x^k,y)\mu_{r_t}(x^k)}{\sum_{(x^k,y)\in A} D_t(x^k,y)\mu_{r_t}(x^k)} \tag{3}$$

(3) 置 $\alpha_t=\varepsilon_t/(1-\varepsilon_t)$。

(4) 将 r_t 加入 R_t。

(5) 更新分布 $D_{t+1}(x^k,y)=\frac{D_t(x^k,y)}{Z_t}\cdot\alpha_t^{\frac{1}{2}(1+h_t(x^k,y^k)-h_t(x^k,y))}$，

其中 Z_t 为标准化常量因子。

输出：训练好的 L 个模糊分类规则。

3.2 模糊分类规则的集成

对于一个给定的多分类器集成问题，决策信息的来源主要根据：分类器当前的输出以及单一分类器的置信度。一般来说，可以用分类器的识别率作为置信度，但是实际上识别率并不能完全代表单一分类器的置信度。基于上述模糊分类规则的集成学习算法，设 $\alpha_t=E(r_t)/(1-E(r_t))$，形成每条规则 r_t 的权值因子 $\log(1/\alpha_t)$，它反映了分类的误差和输入模式的分布信息。这样具有较小分类差错率的规则将具有较大的规则权值，所有分类规则都参与决策，每条分类规则与输入的匹配度以及规则的权值共同决策分类的结果。由此根据式（4）可计算出基于模糊分类规则的分类器最终的输出。

$$f(x^k)=\arg\max_{y\in Y}\sum_{t=1}^{L}\log(1/\alpha_t)\sum_{r_t:y=y^k}\mu_{r_t}(x^k)U_t(x^k) \tag{4}$$

4 实验验证

下面利用 Iris 数据作为训练样本，该数据集为四维输入模式空间，有 3 类 150 个实例样本，每类 50 个。GA 算法中的参数设定为：群体大小=50，变异概率为 0.01，模糊分类规则学习的繁衍代数为 50。外部 AdaBoost.M2 算法每轮迭代产生一个模糊分类规则，一个分类规则视为一个分类器，共产生 L 个规则的集成分类器。当 L 变大时，集成后识别率增高，但测试中发现约 9 条规则后，集成分类精度的提高趋于停滞。原理上，AdaBoost.M2 算法可以迭代直到集成分类器对所有训练样本正确分类，但是为了防止训练样本过配，算法迭代 11 次停止。为了评估算法的准确率，实验中采用了 k-次交叉验证法。将数据集分为 k 个子集，用 $k-1$ 个子集作训练集，1 个子集作测试集，然后 k 次交叉验证；根据平均分类结果测试算法性能。表 1 给出了基于模糊分类规则集成学习算法关于 Iris 数据集的分类效果同基于 GA 的模糊分类规则[7]分类效果以及模糊联想分类效果的对比，表中模糊划分数的最大值是指输入模式空间初始模糊划分数。可见本文提出的基于模糊分类规则集成学习具有较好的分类性能。

表 1　基于集成学习的模糊规则数及平均识别率
(10 次实验的均值)

方法	模糊划分数的最大值					
	4		5		6	
	正确分类率(%)	规则数	正确分类率(%)	规则数	正确分类率(%)	规则数
文献[7]中方法	96.67	8.8	99.20	11.8	99.47	12.6
模糊联想分类	97.33	8.3	99.47	10.8	99.60	12.0
集成学习	97.57	8	99.64	9	99.75	11

5 小结

结合 AdaBoost.M2 算法和遗传算法，本文提出了一种模糊分类规则的迭代产生方法。首先它把规则的前件、后件扩展为模糊集，使得分类界面的变化范围更大，因而分类性能比通常的模糊分类器要好。 基于 AdaBoost.M2 集成训练算法，在训练的每轮的迭代中，根据当前训练样本的分布，基于遗传算法对规则的前件、后件进行优化，产生分类规则，由于在遗传算法的适应度函数中考虑了训练实例的分布，使得在规则产生阶段就考虑的规则之间的相互协作，改善了模糊分类规则的整体识别能力。

参 考 文 献

[1] Cordon O, del Jesus M J. Genetic learning of fuzzy classification systems cooperating with fuzzy reasoning methods [J]. *International Journal of Intelligent Systems*, 1998, 13(10/11): 1025 – 1053.

[2] Merz C J. Using correspondence analysis to combine classifiers[J]. *Machine Learning*, 1999 36(1/2): 33 – 58.

[3] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C]. Proc. of the Thirteenth International Conference on Machine Learning, Morgan Kaufmann, 1996: 148 – 156.

[4] Gonzalez A, Herrera F. Multi-stage genetic fuzzy systems based on the iterative rule learning approach[J]. *Mathware & Soft Computing*, 1997, 4: 233 – 249.

[5] Schapire R E. Theoretical views of boosting[C]. In Proc. 4th European Conference on Computational Learning Theory, 1999: 1 – 10.

[6] Freund Y. Boosting a weak learning algorithm by majority[J]. *Information and Computation*, 1995, 2(121): 256 – 285.

[7] Ishibuchi H, Nozaki K, Yamamoto N, *et al*.. Selecting fuzzy if-then rules for classification problems using genetic algorithms[J]. *IEEE Trans. on Fuzzy Systems*, 1995, 3(2): 260 – 270.

方　敏：　女，1965 年生，副教授，计算机软件专业，研究方向为模式识别、网络安全.

王宝树：　男，1942 年生，博士生导师，研究领域：多传感器信息融合.