第 31 卷 第 20 期 计 算 机 工 程 2005 年 10 月
Vol.31 № 20 Computer Engineering October 2005
·软件技术与数据库· 文章编号：1000—3428(2005)20—0107—03 文献标识码：A 中图分类号：TP18

一种基于粒子群算法的分类器设计

段晓东[1,2]，王存睿[1]，王楠楠[2,3]，刘向东[1]，石 丽[3]
(1. 东北大学信息科学与工程学院，沈阳 110004；2. 大连民族学院非线性信息技术研究所，大连 116600；
3. 沈阳理工大学信息与工程学院，沈阳 110168)

摘 要：将粒子群算法应用于数据分类，给出了适用于粒子群算法的分类规则编码，构造了新的分类规则适应度函数来更准确的提取规则集，并通过修改粒子位置更新方程使粒子群算法适于解决分类规则挖掘问题，进而实现了基于粒子群算法的分类器设计。该文进一步用 UCI 基准数据集对作者提出的粒子群分类器进行了测试，并将几种不同速度与位置更新策略的粒子群算法分类器与遗传算法分类器进行对比，实验结果表明，这种粒子群分类器是一种有效、可行的分类器设计方案。

关键词：数据挖掘；粒子群；分类器；分类规则

Design of Classifier Based on Particle Swarm Algorithm

DUAN Xiaodong[1,2], WANG Cunrui[1], WANG Nannan[2,3], LIU Xiangdong[1], SHI Li[3]
(1. Faculty of Information Sci. and Eng., Northeastern Univ., Shenyang 110004; 2. Research Institute of Nonlinear Information Technology, Dalian Nationalities Univ., Dalian 116600; 3. Faculty of Information and Engineering, Shenyang Univ. of Technology, Shenyang 110168)

【Abstract】 This paper proposes the use of particle swarm optimization(PSO)as a method for classification rule discovery and accomplishes a particle swarm classifier. To make it suitable for PSO algorithm, it designs a new classification rule coding and defines a function for evaluating classification rule. To test its performance, UCI repository used for machine learning is used for experimental testing. The result indicates that PSO classifier is an available and feasible classifier for data mining.

【Key words】 Data mining; Particle swarm; Classifier; Classification rule

1 粒子群分类器的分类规则编码与适应度

1.1 分类规则编码

基于遗传算法分类器的典型分类规则编码有 Holland 的密歇根方案和 De Jong 的匹茨堡方案。两种方法的根本区别在个体与群体组成上，密歇根方法将一条规则对应为一个个体，而规则集对应一个种群；匹兹堡方案视规则集为一个个体，多个规则集对应一个种群。

在密歇根方案中，采用二进制串来表示分类规则。假定特征属性类型为离散型，如果一个特征属性有 n 个可能的取值，则在二进制串中为其分配 n 位，每一位与特定的取值对应，取 0 表示析取式中没有该取值，取 1 表示析取式中有该取值。对于类别属性，则采用连续二进制进行表示。

例如，数据集 D 包含 2 个特征属性 x{x_kind_1, x_kind_2, x_kind_3, x_kind_4}和 y{y_kind_1, y_kind_2, y_kind_3}，4 个类别属性 class_a、class_b、 class_c 和 class_d，则规则

(IF <x = x_kind_1 or x_kind_4> AND <y = y_kind_2 or y_kind_3> THEN class =class_b)

可以表示为二进制串((1001 011)，(01))。反过来，二进制串((0110101)，(11))对应规则为

(IF <x = x_kind_2 or x_kind_3> AND <y = y_kind_1 or y_kind_3> THEN class = class_d)

上述编码方式适合遗传算法对编码进行的交叉和变异操作，但却不适于粒子群算法对粒子速度和位置的更新操作。因为在粒子群算法中，每个粒子的位置由不同的维度组成，粒子之间交换信息仅限于在相同维度间进行，并且粒子群算法中的维度为一定范围内的实数。为此，本文在遗传算法密歇根方案规则编码的基础上，提出了一种适合应用于粒子群算法的分类规则编码方案。

粒子群分类规则编码以每个粒子表示一条规则，规则集对应某个粒子群。每个粒子由不同的维度(定义为整数)组成，数据集 D 中每个特征属性对应粒子的不同维度值。数据集中的类别属性也对应粒子的一个维度值，但不参与粒子间的信息交换，属于粒子的恒定属性。对于一条分类规则，首先按照密歇根编码方案编码为按特征属性分段的二进制串，然后将每段二进制串转化为十进制数，作为粒子的不同维度。例如，上述数据集 D 中分类规则

(IF <x = x_kind1 or x_kind4> AND <y = y_kind2 or y_kind3> THEN class = class_b)

其遗传算法对应的规则编码为((1001 011)，(01))，将编码中的特征属性对应部分按不同特征属性分段成((1001),(011),(01))，再转换为十进制的粒子群分类规则编码(9, 3, 1)。反之，粒子(10, 4 , 3)对应的分类规则为

粒子$(10, 4, 3)_{整型表达}$ => $((1010),(100),(11))_{分段的二进制编码}$

(IF <x = x_kind1 or x_kind3> AND <y = y_kind1> THEN class = class_d)

通过上述方法，分类规则可以表达为一个粒子，任一粒

基金项目：高校优秀青年教师资助计划项目；国家自然科学基金资助项目(69974008)；辽宁省自然科学基金资助项目(20021069)
作者简介：段晓东(1963—)，男，博士、教授，主研方向：人工智能与非线性线性信息处理技术；王存睿、王楠楠，硕士生；刘向东，博士、教授；石 丽，教授
收稿日期：2004-08-22 **E-mail：**aplnan@sohu.com

子也与某个规则惟一对应。

粒子的每个维度经编码后为在某个固定范围的自然数，在粒子的“飞行”过程中须保证其编码的合法性，如超出特定范围，则取其对应的边界值。

采用这种编码规则，粒子每个维度表达了不同的含义，可以方便地进行粒子群算法的粒子位置更新操作，同时也满足了粒子间不同维度进行信息交换的独立性的要求。

1.2 分类规则的适应度

分类规则的评估一般采用灵敏度(sensitivity)、特效度(specificity)和精度(precision)等指标。假设数据集 D 有类别属性 class_a 和其它类别属性，记为 $\overline{\text{class\_a}}$，规则 R 是关于类别 class_a 的一条分类规则。称被按规则 R 正确分类为类别 class_a 的样本数为真正样本数，记为 t_pos，D 中类别为 class_a 的样本个数为正样本数，记做 pos。D 中类别为 $\overline{\text{class\_a}}$ 的样本个数为负样本数，记为 neg。D 中被按规则 R 正确拒绝分为 $\overline{\text{class\_a}}$ 的样本数为真负样本数，记为 t_neg。灵敏度、特效度和精度分别定义为

$$\text{sensitivity} = \text{t\_pos/pos} \tag{1}$$

$$\text{specificity} = \text{t\_neg/neg} \tag{2}$$

$$\text{precision} = \text{t\_pos/(t\_pos+f\_pos)} \tag{3}$$

其中，精度(*precision*)表示被分类标记为正样本的数据实际是正样本的百分比，常用于分类器中的规则匹配；灵敏度表示规则对正样本的分类情况，特效度描述的是对负样本的拒绝情况。规则 R 的适应度一般取为

$$f(\text{R}) = \text{sensitivity}\times\text{specificity} \tag{4}$$

由于分类规则挖掘最终要得到的是能将数据尽可能正确分类的规则集，而不是某个单一的最优规则。如果某一规则的灵敏度不是很高，但特效度很高，这表示该规则能表达部分正样本数据，并且能拒绝绝大部分的负样本数据，这样的规则更适合于将数据进行正确分类。因此本文在定义分类规则适应度时，加入 θ_1、θ_2 两个权值来调整适应度中灵敏度与特效度的权重，采用的分类规则适应度计算公式为

$$f(\text{R}) = (\theta 1\times \text{sensitivity})\times(\theta 2\times\text{specificity}) \tag{5}$$

其中，θ_1, θ_2 为[1,2]之间的实数，分别是 *sensitivity* 和 *specificity* 的权重系数。并且，在算法过程中，取 $\theta_2>\theta_1$。

2 粒子群分类器

通过以上分类规则的粒子群编码和分类规则适应度的定义，就可以利用粒子群优化算法进行分类规则的挖掘，进而形成数据分类器。但由于分类规则挖掘问题的解决不仅要获得代表一条最佳规则的粒子，而且更注重最佳协调的规则组合的获得，正是由于分类规则挖掘问题的特殊性，因此针对分类规则挖掘问题，本文修改了粒子群算法粒子位置更新方程，并应用序列覆盖算法逐步挖掘数据集中的分类规则，然后采用信任分配算法(Credit Assignment Algorithm, CAA)使得规则集对数据进行分类，最终实现基于粒子群算法的分类器设计。

2.1 应用于分类规则挖掘的基本粒子群算法

设数据集 D 中包含 r-1 个特征属性 $\text{x}_1, \text{x}_2, \cdots, \text{x}_{r-1}$ 和 q 个类别属性 $\text{class}_1, \text{class}_2, \cdots, \text{class}_q$，它们共同构成 r 维规则搜索空间。在这个 r 维目标搜索空间内，由 n 个粒子组成一个群落，其中第 i 个粒子为一个 r 维向量 $\text{x}_i = (\text{x}_{i1}, \text{x}_{i2}, \cdots, \text{x}_{i,r-1}, \text{x}_{ir})$，$i = 1,2,\cdots,n$。

$$\begin{cases}\text{x}_{ij}\in\{1,2,3,\cdots,2^{m_j}-1\}, j=1,2,\cdots,r-1 & \text{特征属性x}_j\text{有}m_j\text{种不同取值}\\ \text{x}_{ir}\in\{1,2,3,\cdots,q\}\end{cases}$$

然后，通过式(5)即可计算出其适应值 $f(\text{x}_i)$。

第 i 个粒子的“飞行”速度是一个 r-1 维的向量，记作 $\boldsymbol{v}_i = (v_{i1}, v_{i2}, \cdots, v_{i,r-1})$，$i = 1, 2, \cdots, n$，$v_{ij}\in R$，$j= 1,2, \cdots, r$-1 且 $v_{ij}\in[-V_{\max}, V_{\max}]$，$V_{\max}$ 是常数，通过经验指定。记第 i 个粒子迄今为止搜索到的最优位置为 $p_i = (p_{i1}, p_{i2},\cdots, p_{i,r-1}, p_{ir})$，$i = 1,2, \cdots,n$。整个粒子群迄今为止搜索到的最优位置为 $p_g = (p_{g1}, p_{g2}, \cdots, p_{g,r-1}, p_{gr})$。

基于基本粒子群算法(Basic PSO，BPSO)的分类规则挖掘算法具体过程如下：

Step 1 对数据集 D 进行预处理，包括对数据集 D 的数据进行数据清理，清除非规范数据、平滑噪声数据和填写缺失值等，如果数据集数据未进行离散化，要对数据进行概念分层离散化处理。

Step 2 初始化权重因子 θ_1、θ_2，学习因子 c_1、c_2，惯性因子 ω，以及最大迭代次数 L，速度最大值 $V_{\max}$ 和数据量阈值 M 等参数。

Step 3 确定本次搜索的要挖掘的分类规则类别 class_k，$k \in\{ 1, 2,\cdots, q\}$。如果数据集 D 中类别 class_k 的数据个数小于阈值 M，这时意味着该类别数据量过小，可终止该类别的规则挖掘，进行下一个类别规则的挖掘，$k = k$+1，并将数据集恢复成完整的数据集；如果 $k>q$，则转 Step 7。

Step 4 在对应的范围内随机生成粒子 i 的位置向量 $\text{x}_i = (\text{x}_{i1}, \text{x}_{i2}, \cdots, \text{x}_{i,r-1}, \text{x}_{ir})$，表示规则结论的规则类别属性 $x_{ir} = k$，在速度设定范围内随机生成每个粒子的速度向量 v_i，置 $p_i =\text{x}_i$，$i = 1,2, \cdots,n$；$p_g = (0, 0, \cdots, 0, k)$。

Step 5 计算 n 个粒子的适应度 $f(\text{x}_i)$，如果 $f(\text{x}_i) > f(p_i)$，则 $p_i = \text{x}_i$，$i = 1,2, \cdots,n$；$f(\text{x}_j) = \max(f(\text{x}_i)$，$i = 1, 2, \cdots, n)$，如果 $f(\text{x}_j) > f(p_i)$，则 $p_g = \text{x}_j$；如已经达到最大迭代次数 L，则转 Step 7。

Step 6 按式(8)更新 n 个粒子的速度向量 v_i 和位置 x_i 。

$$v_{ij} = \omega v_{ij} + c_1\alpha_1 (p_{ij} - \text{x}_{ij}) + c_2\alpha_2 (p_{gj} - \text{x}_{ij})\ i = 1, 2, \cdots,n,\ j= 1, 2, \cdots,r\text{-}1 \tag{6}$$

其中，$\omega\geq 0$，称为惯性因子；学习因子 c_1 和 c_2 是非负常数；α_1 和 α_2 是介于[0,1]之间的随机数；对于粒子位置向量 x_i，由于粒子速度向量为浮点数，可是 x_i 要求为正整数，因此在原有算法位置向量更新方式的基础上，对 x_i 进行取整操作，如式(7)所示。

$$\text{x}_{ij} = \text{int}(\text{x}_{ij} + \text{v}_{ij}) \qquad \text{i} = 1, 2, \cdots,\text{n},\ \text{j} = 1, 2, \cdots,\text{r-1} \tag{7}$$

完成粒子群中所有粒子的位置和速度更新操作后，进一步检查，如果 $\text{x}_i (i = 1, 2, \cdots, n)$超出设定的范围，则取边界值。转 Step5。

Step 7 将搜索到的粒子最优位置 $\boldsymbol{p}_g$ 置入规则集 R 中，然后采用序列覆盖算法在数据集 D 中移去规则 $\boldsymbol{p}_g$ 覆盖的数据，即特征属性和分类属性均与规则相匹配的数据，其它类别属性的数据保留，以使在挖掘某一类规则时，能够保持负样本数量不变，以保证挖掘规则的准确性。完成操作后转 Step3。

对于上述算法，有以下两点说明：(1) 粒子代表规则类别的 x_{ir}，$i = 1,2, \cdots,n$ 在算法流程中不参加粒子 i 的位置和速度更新操作；(2) 算法采取限制信息策略，即只允许相同类别属性的粒子进行信息交换。

2.2 其他粒子群算法分类规则挖掘方案

除基本粒子群算法(BPSO)外，粒子群算法有很多改进算法，如自适应粒子群算法[5](Adaptive PSO, APSO)和中值基本粒子群算法(Middle Basic PSO,MBPSO)[6]。BPSO 是利用群体最优信息和个体经验信息调整自身方向和位置；MBPSO 是个基于群最优和所有个体的经验的平均位置来改变自身的参数；BPSO 和 MBPSO 都从信息利用的角度出发，分别通过个体经验信息和个体平均经验信息来平衡算法搜索效率和精度之间的矛盾。APSO 通过线性递减惯性因子 ω 使算法收敛速度加快。可以很容易地将基于基本粒子群算法的粒子群分类器推广到基于其他改进的粒子群算法的粒子群分类器设计。

例如，若采用 MBPSO，则需在 Step4 中计算所有个体的

经验的平均位置 $\bar{p}_v$ = (p_{v1}, p_{v2}, …, $p_{v,r-1}$, p_{vr})，其中 $p_{vj} = p_{1j} + p_{2j} + ... + p_{nj} / n$，$j$ = 1, 2, …, r；并将 Step7 的速度更新公式(6)改为如下速度更新公式：

$$v_{ij} = \omega v_{ij} + c_1 r_1 (p_{vj} - x_{ij}) + c_2 r_2 (p_{gj} - x_{ij})$$

i = 1,2, …,n，j= 1,2, …,r-1 (8)

若采用 APSO，则需将 Step7 中粒子速度更新操作中的惯性因子 ω 按迭代次数线性递减。

2.3 粒子群分类器设计

完成了对训练数据集 D 的分类规则挖掘后，首先，要对规则集 R 进行约简。由于 R 可能存在冗余规则，因此本文根据文献[3]的约简算法对规则集进行约简。具体做法为：若在同一类别的规则集中，2 条规则仅有一个特征属性描述不同，此时这 2 条规则要么有包含关系，可以除去被包含的规则；要么可以进行合并处理，归纳为一个更广义的描述。

然后，对约简后的规则集 R，采用信任分配算法(Credit Assignment Algorithm, CAA)，即根据分类规则的权值来决定数据分属的类别。本文用式(3)定义的精度(precision)作为分类规则的权值，根据分类规则的权值来决定矛盾数据分属的类别，来完成分类器的最终设计。

粒子群分类器实现过程如图 1 所示。

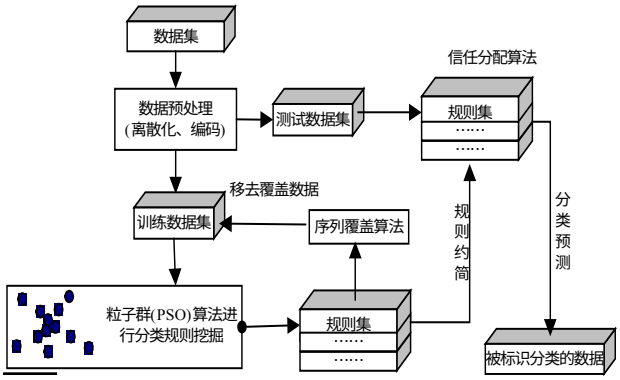

图 1 粒子群分类器实现过程

本文利用分类器对测试集进行分类，统计分类器分类的性能指标，来评价分类器的适用性。

3 实验结果及分析

本文采用来自于 UCI(美国加利福尼亚大学欧文分校)所整理的机器学习知识库[7]中的数据集作为测试数据集对粒子群分类器性能进行了性能测试。数据集 Breast cancer 数据集，来自于南斯拉夫卢布尔维那肿瘤医学研究中心，由 M. Zwitter 和 M. Soklic 搜集整理。

3.1 Wisconsin Breast Cancer 数据集

Wisconsin Breast Cancer 数据集是关于乳腺癌诊断数据。此数据集包括病人癌变部位的 9 个特征数据：簇厚度，细胞尺寸均匀性，细胞形状均匀性，边缘附着力，单层上皮细胞尺寸，裸核，染色质，正常核仁和有丝分裂，每个特征属性有 10 个量级；还有 2 个肿瘤性质分类属性——良性和恶性，共 699 条数据。

表 1 是利用基于基本粒子群算法的粒子群分类器挖掘出的规则集，及算法初始化的参数值、分类器对数据分类的准确率和对数据集中各类别分类的情况。

表 2 是基于基本粒子群算法的粒子群分类器和基于遗传算法的分类器对 Breast cancer 数据集进行分类准确率的均值和标准方差的对比情况。

表 1 粒子群分类器挖掘出的 Cancer 数据集中的规则集

	n =20，ω = 0.7，V_{max} = 5，c_1=c_2=2
良性	细胞尺寸均匀性(1 2 3 4 9)细胞形状均匀性(1 3 4 7 9 10)边缘附着力(1 2 3 4 6 7 10)正常核仁(1 2 3 7 8 9 10)=>良性 细胞尺寸均匀性(1 2 3 6 7 8 9)细胞形状均匀性(1 2 3 7 10)裸核(1 2 5 7 8 9 10)正常核仁(1 2 5 7 10)=>良性 细胞尺寸均匀性(1 2 3 7 9 10)单层上皮细胞尺寸(1 2 4 5 6 7 10)裸核(1 2 4 6 7)正常核仁(1 3 5 6 8 10)=>良性 簇厚度(1 2 3 4 5 6 9 10)裸核(1 3 4 5 8 9)=>良性
恶性	细胞尺寸均匀性(2 4 5 6 7 8 9 10)单层上皮细胞尺寸(3 4 5 6 7 8 10)=>恶性 细胞形状均匀性(3 4 5 6 7 8 10)裸核(1 2 3 4 7 8 10)=>恶性 细胞尺寸均匀性(3 4 5 8 10)裸核(1 2 5 8 9 10)=>恶性
准确率：95.13591%，肿瘤良性覆盖率：442/458，肿瘤恶性覆盖率：223/241	

表 2 Wisconsin-Breast-Cancer 数据集的两种分类器对比情况

	参数设置	准确率均值	准确率标准差
BPSO	n=20,ω=0.7,V_{max}=300,L=100	92.5464%	±0.01607
GA	n=60,变异率:0.05 交叉率:0.25	92.1380%	±0.02801

本文进一步将几种常用的粒子群算法，即基本粒子群算法、中值基本粒子群算法和自适应粒子群算法对应的粒子群分类器进行了性能对比，结果如表 3 所示。

表 3 Wisconsin-Breast-Cancer 数据集的分类情况

	参数设置	准确率均值	准确率标准差
BPSO	n=20,ω=0.7, V_{max} =300,L=100	92.5464%	±0.01607
MBPOS	n=20,ω=0.7, V_{max} =300,L=100	92.7911 %	±0.01953
APSO	n=20,ω=0.9~0.4,V_{max}=300,L=100	92.0472 %	±0.02038

3.2 实验结果分析

由实验结果表 1 和表 2 可以得出：(1)粒子群分类器是可行、有效的分类器设计方法，其效能与遗传算法分类器基本一致；(2) 在基于不同的粒子群算法的粒子群分类中，基于 MBPSO 的粒子群分类器预测的准确率较高，基于 MBPSO 的粒子群分类器的预测准确率较稳定。

4 结论

本文首先给出了适用于粒子群算法的分类规则编码方案，构造了新的可以更准确地提取规则集的分类规则适应度函数，给出了基于粒子群算法分类规则挖掘算法与粒子群分类器设计的完整方案，并通过基准数据集实例对粒子群分类器进行了测试。测试表明，本文提出粒子群分类器是一种有效、可行的分类器设计方案。

参考文献

1 Holland J H. Genetic Algorithm and Classifier System：Foundations and Future Directions[C]. In: Proceedings of the Second International Conference on Genetic Algorithms, Lawrence Erlbaum Associates, Publishers, 1987: 82-89

2 Kennedy J, Eberhart R. Particle Swarm Optimization[C]. In: Proc. of IEEE Int. Conf. on Neural Networks, Perth, Australia, 1995: 1942

3 形乃宁，孙志辉. 基于增量式遗传算法的分类规则挖掘[J]. 计算机应用研究，2001, 18(11): 13-15

4 Shi Yuhui, Eberhart R. ParameterSselection in Particle Swarm Optimization [A]. In: Proc. of the 7th Annual Conf. on Evolutionary Programming[C]. Washington DC, 1998: 591-600

5 Shi Yuhui, Eberhart R. A Modified Particle Swarm Optimizer [A]. In: Proc. of IEEE Int. Conf. on Evolutionary Computation[C]. Anchorage, 1998: 69-73

6 王存睿，段晓东，刘向东等. 改进的基本粒子群算法[J]. 计算机工程，2004, 30(21): 35