第 34 卷 第 11 期 计 算 机 工 程 2008 年 6 月
Vol.34 No.11 Computer Engineering June 2008
·软件技术与数据库· 文章编号：1000—3428(2008)11—0086—02 文献标识码：A 中图分类号：TP311

一种基于混合算法的分类器设计

邝艳敏，王自强，李 鹏
(河南工业大学信息科学与工程学院，郑州 450001)

摘 要：为了高效地从数据库中挖掘分类规则，提出一种将粒子群优化算法和遗传算法相结合的新算法。该算法的核心思想是对规则的前件进行固定长度编码，适应度函数的计算由分类规则的准确率、置信度、支持度和简洁度构成，从而实现基于两者混合算法的分类器设计。将该分类器与遗传算法分类器和粒子群算法分类器进行对比，实验结果表明，该分类器具有更高的分类准确率以及更快的收敛速度。

关键词：数据挖掘；粒子群；遗传算法；分类器；分类规则

Design of Classifier Based on Hybrid Algorithm

KUANG Yan-min, WANG Zi-qiang, LI Peng
(College of Information Science and Enginering, Henan University of Technology, Zhengzhou 450001)

【Abstract】 To efficiently mine the classification rule from database, a novel hybrid classification algorithm based on Particle Swarm Optimization(PSO) and Genetic Algorithm(GA) is proposed. The core idea of the proposed algorithm is as follows: a new rule code with fixed length is proposed, a novel fitness function combined with accuracy, confidence, support and simplicity is constructed, and a hybrid heuristic classifier is accomplished. Experimental results show that the proposed classification algorithm acheives higher classification accuracy and lower running time compared with other classification algorithms.

【Key words】 data mining; particle swarm; genetic algorithm; classifier; classification rule

分类规则挖掘是数据挖掘最重要的研究领域之一。分类规则挖掘就是研究一组已知其类别的数据对象(训练数据集)的属性与其类别间的关系，发现其规律(分类的规则)，用来对未知类别的数据对象作出类别判断[1]。分类规则的挖掘采用的方法有很多，目前，许多学者研究探讨了分类规则挖掘的具体算法和相关问题，研究显示没有一种算法在规定条件下的性能更为突出。本文提出将粒子群优化算法(Partical Swarm Optimization, PSO)[2]和遗传算法(Genetic Algorithm, GA)[3]相结合的方法来进行分类规则的挖掘，进而实现一个分类器。

1 分类器的分类规则编码与适应度函数

1.1 分类规则编码

本文采用Michigan方法来对规则进行固定长度编码，每个个体代表一条分类规则，只对分类规则的前件(IF部分)进行编码，编码结构如图 1 所示。在图 1 中，A_i表示第i个属性；W_i为权值域，其值是在属性上取当前值的个体数占所有个体数的百分比；O_i是与属性A_i相联系的运算符，对离散属性而言，取“=”和“≠”符号；对于连续属性，取“≤”和“＞”符号；V_i是属性在规则中的取值，若是离散属性，值等于实际值在取值域中的位置；若是连续属性，值等于实际值；G_i是属性的信息增益域，每个属性的信息增益值在混合算法执行前已经计算好并存储在个体的G_i中。

A_1	…	A_i	…	A_n
$W_1O_1V_1G_1$	…	$W_iO_iV_iG_i$	…	$W_nO_nV_nG_n$

图 1 个体的编码结构

1.2 分类规则的适应度函数

适应度函数应能评价个体(规则)的好坏。设有分类规则“IF A THEN C”，则数据集中共存在下面 4 类不同的规则，其个体数目如表 1 所示。

表 1 数据集中存在的四类规则及其数目

规则				个体数目
IF	A	THEN	C	Tp
IF	A	THEN	NOT C	Tn
IF	NOT A	THEN	C	Fp
IF	NOT A	THEN	NOT C	Fn

为了提高分类规则的挖掘效果，本文的适应度函数主要由算法的分类准确率、规则的置信度、规则的支持度以及规则的简洁度构成。

(1)准确率(*accuracy*)：规则的准确率越高，说明规则正确分类的样本越多。

$$accuracy = \frac{Tp + Tn}{pos + neg} \tag{1}$$

其中，*pos* 是正样本总数；*neg* 是负样本总数。

(2)置信度(*confidence*)：对于分类规则也称为精确度(*precision*)，规则的置信度表示规则在训练集上的正确程度，当置信度为 1 时，规则在训练集上恒真，此时只要条件为真，结论恒为真。其定义如下：

$$confidence = \frac{Tp}{Tp + Fp} \tag{2}$$

(3)支持度(*support*)：规则的支持度越大，说明规则在数据集空间所占的比例越大，规则的普遍意义越好。定义如下：

$$support = \frac{Tp + Fp}{pos + neg} \tag{3}$$

(4)简洁度(*simplicity*)：规则的简洁度越大，说明规则的

基金项目：河南省自然科学基金资助项目(0624010002)；郑州市科技攻关基金资助项目(2006-8-1)
作者简介：邝艳敏(1983－)，女，硕士研究生，主研方向：数据挖掘；王自强，副教授；李 鹏，硕士研究生
收稿日期：2007-06-30 **E-mail：**yanminkuang@yahoo.com.cn

结构越简单，规则越容易理解。其定义如下：

$$simplicity = \frac{attributes - v\_attributes}{attributes} \quad (4)$$

其中，*attributes* 是数据集上的属性总数；*v_attributes* 是规则中出现的属性个数。

本文综合考虑以上 4 种因素，给出适应度函数定义如下：

$$fitness = accuracy + confidence + support + simplicity$$

2 分类器

通过上述分类规则编码和分类规则适应度的定义，可以利用本文提出的粒子群优化算法和遗传算法相结合的混合算法进行分类规则的挖掘，进而形成数据分类器。

2.1 分类规则挖掘的算法的应用

核心规则挖掘算法描述如下：

```
begin
  确定 PSO 粒子种群规模 m 和 GA 种群规模 n
  初始化 PSO 粒子种群和 GA 种群
利用上文定义的适应度函数计算每个 PSO 粒子和每个 GA 个体的适应度
        选择适应度最优的 PSO 粒子和 GA 个体
  repeat
    repeat
      for each  PSO 粒子
          更新粒子的速度和位置
        end
        计算每个粒子的适应度
        选择适应度最优的 PSO 粒子
      until  算法达到 N 次迭代而最佳适应值不再提高，切换 PSO 粒子种群为 GA 种群
        repeat
          对 GA 种群执行选择、交叉、变异操作
          计算每个个体的适应度
          选择适应度最优的个体
        until 算法达到 N 次迭代而最佳适应值不再提高，切换 GA 种群为 PSO 粒子种群
  until  特定的迭代次数
  返回 PSO 粒子种群或 GA 种群即为优化生成的分类规则
  end
```

(1)核心挖掘算法从设置种群大小开始。此算法设计为允许 GA 和 PSO 并行混合工作和串行混合工作。如果 GA 和 PSO 的种群规模都设置为一个不为零的常数，将执行并行混合算法。该算法也可以通过设置种群规模来指定启发式算法的序列。如果一个算法的种群规模是零，则首先执行另一个启发式算法。

(2)初始化。种群通过启发式算法各算子的操作而改变。当每个算法进化超过特定代数，而最佳适应值不再提高的情况下，则切换算法，种群转换为下一个启发式算法的种群。此周期不断重复直到达到一个终止条件。本文测试的大部分例子，循环在第一周期已终止，因为已经达到收敛。

(3)启发式算法执行过程中，全局最优解在每次迭代中更新，并作为规则返回。

2.2 分类器设计

完成对训练数据集的分类规则挖掘后，为了使生成的规则更加简洁，易于用户理解以及降低过拟合的风险，可采取规则剪枝策略。本文根据文献[4]的剪枝算法对规则集进行剪枝。然后对剪枝后的规则集，采用信任分配算法(Credit Assignment Algorithm, CAA)[5]，即根据分类规则的权值来决定数据所属的类别。本文用式(2)定义的置信度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所属的类别，进而完成分类器的最终设计。

3 实验结果

为了验证本文提出的基于混合算法的分类器性能，采用了美国加州大学机器学习知识库中的 4 个数据集[6]，即Zoo, Breast cancer, Waveform和Wisconsin Breast Cancer作为测试数据。

将本文分类器与粒子群优化算法(PSO)和遗传算法(GA)对应的分类器进行了性能比较，结果如表 2~表 4 所示，其中，GA算法的参数设置为：交叉率pc=0.90，变异pm=0.01。PSO算法的参数设置为：惯性权重w =1，学习因子$c1$ = $c2$ = 2.05, V_{max}= ± 4。本文混合算法的参数设置与上述 2 种算法的参数一致。

表 2　4 种分类器预测准确率的比较　(%)

数据集	GA	PSO	PSO+GA	GA+PSO
Zoo	89	89	88	89
Breast cancer	75	75	75	76
Waveform	57	71	71	72
Wisconsin Breas	94	95	94	95

表 3　4 种分类器运行时间的比较　s

数据集	GA	PSO	PSO+GA	GA+PSO
Zoo	0.62	0.71	0.80	0.68
Breast cancer	1.78	2.21	4.21	2.18
Waveform	77.00	96.20	90.70	78.00
Wisconsin Breas	3.40	3.50	3.86	3.63

表 4　4 种分类器挖掘的规则集大小的比较

数据集	GA	PSO	PSO+GA	GA+PSO
Zoo	8/6	7/6	7/6	7/6
Breast cancer	4/5	4/6	4/6	4/6
Waveform	6/18	7/31	7/32	7/32
Wisconsin Breas	7/7	7/7	7/9	7/7

表 2 列出了 4 种分类器预测准确率的平均值，表 3 给出了 4 种分类器收敛于最优解的运行时间比较，表 4 列出了 4 种分类器挖掘的规则集大小，规则集大小表示为 2 个数字，第 1 个数字表示判定树的分枝个数或属性个数，第 2 个数字表示是树的叶节点的个数。从表的实验数据可以得出：(1)本文基于 GA+PSO 算法的分类器是一种可行、有效的分类器设计方法，在所有 4 种启发式算法中，GA+PSO 算法根据预测准确率和执行时间提供了最佳的结果。(2)在基于不同算法的分类器中，基于 GA 算法的分类器收敛速度较快但预测准确率不高，基于 PSO 算法的分类器预测准确率较高但收敛速度较慢，基于 PSO+GA 算法的分类器预测准确率接近于 PSO，但收敛速度太慢，基于 GA+PSO 算法的分类器收敛速度接近于 GA，预测准确率接近于 PSO。(3)4 种分类器发现的规则集大小相似。

4 结束语

本文给出了基于混合算法的分类规则编码方案，构造了新的分类规则适应度函数，提出了一种将 PSO 算法和 GA 算法相结合的分类规则挖掘算法和分类器设计的完整方案，并通过实验说明了本文提出的分类器有效、可行。

参考文献

[1] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2001.

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

（下转第 90 页）