基于双条件选择策略的 Ant-Miner 算法

李桂成,张惠萍
LI Gui-cheng,ZHANG Hui-ping

山西大学 计算机与信息技术学院,太原 030006
School of Computer & Information Technology,Shanxi University,Taiyuan 030006,China
E-mail:agui@sxu.edu.cn

LI Gui-cheng,ZHANG Hui-ping.Ant-Miner algorithm based on dual condition choose strategy.Computer Engineering and Applications,2009,45(11):147-149.

Abstract: This paper proposes a new condition choose strategy for Ant-Miner,is called dual condition choose strategy.Applied it to Ant-Miner algorithm,and compared it with original Ant-Miner on two standard data sets.The result shows that it is better than original algorithm on both predicted accuracy rate and run time.
Key words: ant colony optimization;Ant-Miner;dual condition choose strategy

摘 要:针对 Ant-Mine 算法提出一种新的条件选择策略-双条件选择策略。将该策略应用于 Ant-Miner 算法中,并与原 Ant-Miner 算法在两个公开的数据集上进行实验比较,结果表明应用了双条件选择策略的算法较原算法不仅具有更快的运行速度,而且获得了更高的预测精度。
关键词:蚁群优化;Ant-Miner;双条件选择策略
DOI:10.3778/j.issn.1002-8331.2009.11.045 **文章编号**:1002-8331(2009)11-0147-03 **文献标识码**:A **中图分类号**:TP18

分类是数据分析的一种重要手段,是数据挖掘的一个重要研究课题[1]。分类问题在人工智能、机器学习等领域已经得到了广泛地研究。目前,已经研究出的经典分类方法主要包括:决策树方法、神经网络方法、遗传算法和支持向量机方法等[2]。近年来,随着人工智能、机器学习和数据挖掘等领域中传统方法的不断发展及多种新方法和新技术的不断涌现,出现了一些新的分类方法,其中包括:基于粒度计算的分类方法、基于群的分类方法等[2]。其中基于群的分类方法可以看作是进化计算的一个新的分支。

受自然界蚂蚁觅食行为的启发,意大利学者 M Dorigo 等人于 20 世纪 90 年代初提出了一种新型的群智能优化算法—蚂蚁系统[3]。该算法有很强的发现好解的能力、较强的鲁棒性以及易于与其他算法相结合的优点,因此引起了许多研究者的注意,对该算法作了多种改进[4-5]并将其应用于更为广泛的领域,取得了一些令人鼓舞的成果。近几年,M Dorigo 等人将蚂蚁算法进一步发展成一种通用的优化技术—蚁群优化(简称 ACO),并将所有符合 ACO 框架的蚂蚁算法称为蚁群优化算法(ACO algorithm)[6]。

2002 年 Parpinelli 和他的同事们首次将 ACO 算法应用于分类规则的挖掘中,提出了一种新的分类算法—Ant-Miner 算法。文献[7]通过实验比较了 Ant-Miner 与著名的分类算法CN2,结果表明在预测精度方面 Ant-Miner 可与 CN2 相媲美,而 Ant-Miner 发现的规则比 CN2 发现的规则更为简洁。随后又有学者对其进行了不断地改进[8-10]。本文针对 Ant-Miner 算法提出了一种新的条件选择策略——双条件选择策略,将使用了该策略的 Ant-Miner 算法称为 Dual Ant-Miner 算法(简称 DA-M)。将 DA-M 与原算法在两个公开数据集中进行实验比较,结果表明 DA-M 算法较原算法不仅具有更短的运行时间,而且获得了更高的预测精度。

1 Ant-Miner 算法简介[7]

1.1 Ant-Miner 算法的具体步骤

Ant-Miner 算法是由 *Parpimelli* 和他的同事们首次提出的,它是第一次将 ACO 算法应用于分类问题的一个分类模型。该算法的目的就是从训练数据中发现一个可被用于预测测试数据的类型的分类模型。由于分类模型一般是由分类规则所组成的,故其目的也就是运用 ACO 技术发现以下形式的分类规则:

IF(条件 1 AND 条件 2 AND 条件 3 AND … AND 条件 m)THEN(类型值)

其中每个条件为一个三元组,其形式为〈属性=属性值〉。

Ant-Miner 算法的具体步骤如下[7]:

```
Begin
  Training Set=all training cases;
  While(NO.of uncovered cases in the Training Set>Max_Uncovered_Cases)
    i=0;
```

基金项目:教育部科学技术研究重点项目(the Key Scientific and Technical Research Project of Ministry of Education of China under Grant No.207018)。
作者简介:李桂成(1963-),男,副教授,主要研究领域为数据挖掘、人工智能;张惠萍(1979-),女,硕士,主要研究领域为数据挖掘、人工智能。
收稿日期:2008-02-27 **修回日期**:2008-05-05

Repeat

$i=i+1$;

Ant_i incrementally constructs a classification rule;

Prune the just_constructed rule;

Update the pheromone of the trail followed by Ant_i;

Until($i \geqslant No\_of\_Ants$)OR(Ant_i constructed the same rule as the previous $No\_Rules\_Converg$−1 Ants)

Select the best rule among all constructed rules;

Remove the cases correctly covered by the selected rule from the Training Set;

End while

End

1.2 信息素的初始化

在每次 while 循环开始时,每个条件的信息素被用公式(1)初始化为相同的值:

$$\tau_{ij}(0)=\frac{1}{\sum_{i=1}^{a} b_i} \tag{1}$$

其中 a 是属性个数,i 是属性标号,b_i 是第 i 个属性的属性值的个数。

1.3 问题依赖的启发式函数

在 Ant-Miner 中,启发式函数是基于信息熵理论的,由公式(2)给出:

$$\eta_{ij}=\frac{\mathrm{lb}(k)-InfoT_{ij}}{\sum_{i=1}^{a}\sum_{j=1}^{b_i}\mathrm{lb}(k)-InfoT_{ij}} \tag{2}$$

其中 $InfoT_{ij}$ 由公式(3)给出:

$$InfoT_{ij}=-\sum_{w=1}^{k}\left[\frac{freqT_{ij}^{w}}{|T_{ij}|}\right]*\mathrm{lb}\left[\frac{freqT_{ij}^{w}}{|T_{ij}|}\right] \tag{3}$$

其中 k 是样本的类型数,$|T_{ij}|$是划分 T_{ij} 中的样本数,$freq_{ij}^{w}$ 为划分 T_{ij} 中类型为 w 的样本数。由以上公式知 $InfoT_{ij}$ 的值越大,蚂蚁选择条件 $term_{ij}$ 加入到当前规则的可能性就越小。

1.4 条件选择所依赖的概率函数

每只蚂蚁都从一个空规则开始构造规则,对那些还没有被选入到规则中的属性来讲,其被选择的概率由公式(4)给出:

$$p_{ij}=\frac{\tau_{ij}(t)\cdot\eta_{ij}}{\sum_{i=1}^{a}(1-x_i)\sum_{j=1}^{b_i}\tau_{ij}(t)\eta_{ij}} \tag{4}$$

其中 x_i 取 0 或者是 1,当取 0 时表示属性 i 没有被当前规则所用到,反之亦然。

1.5 规则的剪枝

为了避免对样本的过度拟合,在规则产生之后要对其进行剪枝。规则剪枝的思路为:反复判断删除某一条件后能否增加该规则质量,若能则删除该条件,否则保留该条件,直到规则中只剩一个条件或删除任一条件都不能增加规则质量时剪枝结束。

规则的质量由公式(5)给出:

$$Q=\frac{TP}{TP+FN}*\frac{TN}{FP+TN} \tag{5}$$

其中:TP 表示训练集中满足规则条件,并且与规则的预测类型相同的样本数;TN 表示训练集中不满足规则条件,并且与规则的预测类型不相同的样本数;FP 表示训练集中满足规则条件,但与规则的预测类型不相同的样本数;FN 表示训练集中不满足规则条件,但与规则的预测类型相同的样本数;Q 的值越大,说明规则质量越高。

1.6 信息素的更新

在蚂蚁完成一条规则的构造后,在该规则中出现的条件的信息素按照公式(6)更新:

$$\tau_{ij}(t+1)=\tau_{ij}(t)+\tau_{ij}(t)*Q,\quad \forall term_{ij}\in the\ \ rule \tag{6}$$

在 Ant-Miner 算法中,信息素的挥发通过用该条件的信息素值除以所有条件的信息素值的总和来实现。

2 概述 Ant-Miner2[8]算法和 Ant-Miner3[9]算法

在 Ant-Miner 的基础上,文献[8-9]提出了对其的改进算法—Ant-Miner2 和 Ant-Miner3 算法。以下将简单介绍这两种改进算法对于原算法的改进之处。

2.1 Ant-Miner2

Ant-Miner2 对原算法的启发式函数进行了修改。原算法的启发式函数是基于信息熵理论的,而 Ant-Miner2 的启发式函数是基于密度的。其启发式函数由公式(7)给出:

$$\eta_{ij}=\frac{|majority\_class\,T_{ij}|}{|T_{ij}|} \tag{7}$$

其中$|majority_class\ \ T_{ij}|$表示 T_{ij} 中的类型值是主要类型的样本个数。

实验表明 Ant-miner2 在与原算法有相同预测精度的基础上,减少了运行时间。

2.2 Ant-Miner3

Ant-Miner3 算法主要在以下两个方面对原算法进行了改进。

2.2.1 信息素更新方法

在 Ant-Miner3 中,对信息素更新用公式(8)代替了公式(6):

$$\tau_{ij}(t)=(1-\rho)\cdot\tau_{ij}(t-1)+(1-\frac{1}{1+Q})\cdot\tau_{ij}(t-1) \tag{8}$$

其中 ρ 是信息素挥发率。ρ 的值越大挥发就越快。

2.2.2 条件选择策略

Ant-Miner3 为了加强算法对解的探索能力,使用了如下条件选择策略[9]:

If $q1\leqslant\varphi$

For 每一属性

If $q2\leqslant\sum_{j\in J_i}P_{ij}$

Then choose $term_{ij}$

End for

Else

Choose $term_{ij}$ with max p_{ij}

End if

其中 $q1$、$q2$ 是两个随机变量,φ 为参数。通过对两个随机变量的使用,使得算法在选择条件 $term_{ij}$ 时,不仅仅依靠启发式函数值和信息素,而且也依赖于随机产生的变量 $q1$、$q2$。这就增加了选择未被用到的条件的可能性,从而也提高了算法对解的探索能力。

3 双条件选择策略

在文献[9]中,通过改变其条件选择策略使得算法的分类预测精度得以提高。由此可知,使用原算法中的条件选择策略所得到的规则并不是最优的。因此,改进算法的条件选择策略对

于提高算法的预测精度是一个有效的方法。

分析 Ant-Miner3 中对条件选择策略的改进，算法中使用了两个随机变量，提高了算法对未使用过的条件的探索能力，增加了解的多样性，使得算法的预测精度总体上有所提高，但算法的运行结果并不稳定，有时也会得到不太理想的结果。

对 Ant-Miner 算法中的条件选择策略进行分析得知，对于一个条件是否被选择加入到当前规则中，只与该条件的信息素总量以及其启发式函数值有关。信息素总量反映了条件当前的被使用情况。公式(2)、(3)表明，启发式函数值反映了该条件的分布均匀度，即：启发式函数值越大，说明该条件的分布越集中。因此，当启发式函数值为 1 时，表示所有被该条件覆盖的样本都属于同一个类型。

规则生成阶段，在第一只蚂蚁构造规则时，各个条件的信息素总量都是相同的，此时哪个条件被选择加入到规则中，就取决于条件的启发式函数值了。如果多个条件的启发式函数值相同时，就会有多个条件的选择概率相同。例如：一个条件只覆盖了一个样本则该条件的启发式函数值是 1；而当另一个条件覆盖了 100 个样本，而且这 100 个样本都属于同一个类型时，该条件的启发式函数值也是 1。在这种情况下，如何来选择条件将影响到最终形成规则的质量。在原算法中并没有提到如何对多个具有相同选择概率的条件进行选择，而在实验中发现这种情况是存在的。

基于以上分析，本文提出了一种新的条件选择策略—双条件选择策略，其主要思想是：首先在条件概率矩阵中求出其最大值，然后在选择概率值等于该最大值的条件中选择出覆盖样本数最多的一个条件作为被选择的条件。该条件选择策略的伪码表示如下：

```
num=0
maxpr=max(p)
While
  If p_ij==maxpr
    If covernum(term_ij)>num
      num=covernum(term_ij)
      choose term_ij
    End if
  End if
End while
```

其中 $covernum(term_{ij})$ 表示训练集中被条件 $term_{ij}$ 所覆盖的样本数，p 表示当前的条件概率矩阵。使用该条件选择策略，可以选择在分布最集中的条件中覆盖的样本数最多的条件。从而使得所得到的规则具有更高的预测精度。

4　实验分析

为了验证本文所提出的双条件选择策略的性能，从 UCI[11] 公共数据库中选取了两个数据集作为实验数据集。表 1 给出了这两个数据集的主要特征。

表 1　实验数据集

数据集	样本个数	属性个数	类型数
Hepatitis	155	19	2
Tic-tac-toe	958	9	2

将本文所提出的双条件选择策略应用在 Ant-Miner 算法中，并和原 Ant-Miner 算法在上面两个数据集中进行实验比较。

实验中所使用的参数的取值如下所示：

(1)*No_of_ants*=1 000；

(2)*Max_uncovered_cases*=10；

(3)*Min_cases_per_rule*=5；

(4)*No_rules_converg*=5。

实验采用十次交叉验证法，每个数据集被分为十个部分，每种算法运行十次，每次用一个不同的部分作为测试集，而用其他部分作为训练集，将十次运行的预测精度的平均值作为该发现规则列表的预测精度。实验算法用 MATLAB 7.0 编写，表 2 为两种算法的实验结果。

表 2　实验结果

	Hepatitis		Tic-tac-toe	
	Ant-Miner	Dual Ant-Miner	Ant-Miner	Dual Ant-Miner
predicted accuracy rate/(%)	86.67±2.63	93.3±2.29	75.04±4.82	77.7±2.92
run time/s	0.47	0.45	5.75	3.91

由表 1 可知，DA-M 算法在两个数据集上都取得了很好的效果，在预测精度和运行时间上都比原算法有明显地改善。

5　总结

针对 Ant-Miner 算法提出了一种新的条件选择策略，并将其应用在原算法中。通过实验得知：DA-M 算法在预测精度以及运行时间上都比 Ant-Miner 算法有明显的改善。

参考文献：

[1] Han J W，Kamber M.Data mining concepts and techniques[M].San Francisco：Elsevier，2001：185-222.

[2] 张丽娟，李舟军.分类方法的新发展：研究综述[J].计算机科学，2006，33(10)：11-15.

[3] Dorigo M，Maniezzo V，Colorni A.Ant system：optimization by a colony of cooperating agents[J].IEEE Transaction on Systems，Man，and Cybernetics：Part B，1996，26(1)：29-41.

[4] Bullnheimer B，Hartl R F.Strauss C.A new rank-based version of the ant system：a computational study[J].Central European Journal for Operations Research and Economics，1999，7(1)：25-38.

[5] Stützle T，Hoos H.MAX-MIN ant system[J].Future Generation Computer Systems，2000，16(8)：889-914.

[6] Dorigo M，Caro D，Gambardella L M.Ant algorithms for discrete optimization[J].Artificial Life，1999，5(2)：137-172.

[7] Parpinelli R S，Lopes H S，Freitas A A.Data mining with an ant colony optimization algorithm[J].IEEE Transactions on Evolutionary Computing，2002，6(4)：321-332.

[8] Liu B，Abbass H.A，Mckay B.Density-based heuristic for rule discovery with ant-miner[C]//The 6th Australia-Japan Joint Workshop on Intelligent and Evolutionary System，Canberra，Australia，2002：180-184

[9] Liu Bo，Abbass H A，Mckay B.Classification rule discovery with ant colony optimization[C]//IEEE/WIC International Conference on Intelligent Agent Technology，2003：83-88.

[10] Wang Zi-qiang，Feng Bo-qin.Classification rule mining with an improved ant colony algorithm[C]//Gi W，Yu X H.Con on Artificial Intelligence，Australia，2004：357-367.

[11] 刘波，潘久辉.基于蚁群优化的分类算法的研究[J].计算机应用与软件，2007，24(4)：50-53.

[12] Hettich S，Bay S D.The UCI KDD archive [EB/OL].(2003).http://kdd.ics.uci.edu.