蚁群算法在网页内容分类中的应用研究①

Application of Ant Colony Algorithm in Page Contents Classification

韩　杰　王自强　李春峰　谭明交　(河南工业大学 信息学院 河南 郑州 450052)

摘　要： 研究各种高效的分类算法是数据挖掘的重要问题之一[1]。蚁群算法作为一种新型的模拟进化算法，在求解复杂的组合优化问题中表现出了良好的性能[2]。文章介绍了蚁群算法在网页内容分类数据挖掘任务中的一种应用方案，阐述了算法的基本原理及特性，并使用少量类别的网页进行了分类实验，实验结果验证了该算法在应用中的有效性。

关键词： 蚁群算法 网页内容分类 分类规则 数据挖掘 文本分类

随着网络信息资源不断的快速增加，我们急需对这些资源进行合理地分类整理以便从海量数据中快速检索到期望的以及相关联信息。数据分类，包括对网页内容进行分类，是数据挖掘研究中的重要内容。常用的数据分类方法有决策树分类法、基于规则的分类法、神经网络、支持向量机和朴素贝叶斯分类法等。基于规则的分类有着其他分类方法不具备的特征，它能产生易于解释的描述性模型，性能却可与其他分类方法性媲美。另外，基于规则的分类允许一条记录触发多条规则，可以构造更加复杂的决策边界。

1　引言

蚁群算法(ant colony algorithm)是由意大利学者 M.Dorigo 等人提出的一种基于蚂蚁种群的新型模拟进化算法。该算法已成功应用于 TSP 问题、集成电路综合布线、指派问题、网络路由、数据聚类、组合优化等问题，并取得了较好的效果。蚁群算法在数据挖掘中的应用正逐步引起人们的关注。目前，蚁群算法在知识发现的过程中主要用于发掘聚类模型和分类模型，本文将讨论蚁群算法在网页分类规则提取中的研究及应用。

2　蚁群分类算法在网页内容分类中的应用

2.1 应用模型

蚁群分类规则提取算法在网页内容分类中的一个应用模型如图 1 所示。算法应用包含两个过程：

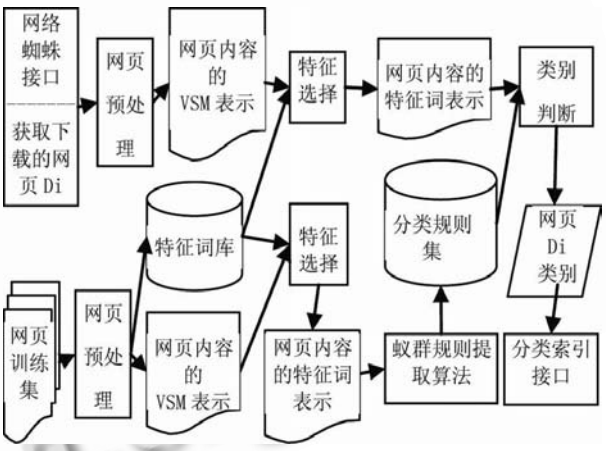

图 1　蚁群分类规则提取算法在网页内容分类中的一个应用模型

第一个是训练过程：首先对已知网页类别的网页训练集进行网页预处理，同时建立一个特征词库；接着进行特征选择，缩减特征词数量；最后使用蚁群规则提取算法可以得到一个分类规则集，用于对网页进行类别判断。

第二个过程是自动分类过程：首先获取待分类的网页 Di，对 Di 进行网页预处理，得到网页 Di 的向量空间模型(VSM)表示；然后根据特征词库进行特征选

① 基金项目:河南省自然科学基金项目(0624010002)
收稿时间:2008-09-20

择，减少 Di 的向量空间的维数，得到使用特征词和词频的网页 Di 的表示；最后就可以使用分类规则集进行类别判断，可以得到网页 Di 的类别，以供其它模块，如分类索引模块进行调用。

虽然蚁群算法处理大规模数据时速度比较慢，但是使用蚁群算法得到分类规则集过程只需要执行一次，而使用分类规则集进行类别判断的速度取决于规则集的数量。本应用模型用于处理分类类别数量比较少的应用，如主题搜索，是可行的。

2.2 网页预处理

网页预处理包括对网页进行 Html 解析，网页净化，中英文分词，去停用词，词频统计和对所有网页用向量空间模型表示。本应用主要采用文献[3]中所述的网页预处理方法，但是对文献[3]提供的处理作了如下修改：

中文分词采用正向最大匹配与逆向最大匹配相结合的方法，并且只需要分出名词和动词。Html 解析使用 DocView 模型完成解析网页信息和网页内容的提取，其中网页信息包括：网页标识、网页类型、内容类别、标题、关键词和摘要。如果网页信息与正文无关，如网页标题经常为“无标题文档”，则忽略此网页信息，否则，增加在网页内容中出现的词条与网页信息词条相等的词的权重。

对网页训练集处理时选择词条放入特征词库的方法：对网页的每一个类 i(0<i<=网页类别数)进行如下操作：(1).对属于 i 类的所有文章的所有词条 Iij(i 为类别，0<j<=S，S 为 i 类的所有词条数)，在第 i 类网页内累加每个文章中出现词条 Iij 的权重 Wij(i 为类别，0<j<=S)，这样得到一个词条和权重的数组 […<Iij,Wij>…]。(2).把所有词条按照权重大小从大到小排序。(3).根据 i 类的词条数量 S，按照一定的比率 θ 选择权重大的前 θ *S 个词条。(4).将选择的词条加入特征词库，如果特征词库中已存在该词，则不再加入。这种方法可以根据 θ 控制特征词库的规模。

特征选择[4]方法为：使用信息增益(Information Galn，IG)法选取特征词库中存在的词条。

3 蚁群算法的分类规则提取算法

分类规则[5,6]就是要求出 IF(规则前件)THEN<Class>其中规则前件是一组属性测试的合取：<term1>and<term2>and.....，每一个 term 是一个三元组合<attribute,operator,value>。Value 是属性 attribute 中的值，operator 为一个关系操作符，通常为相等关系。Class 称为该分类规则预测的类别。

蚁群构造规则的过程分为三个阶段:第一个阶段蚂蚁从一条空规则开始，重复选择属性节点到路径上，直到得到一条完整的路径，即生成一条分类规则，生成一个规则后，它所覆盖的训练样本从训练集中删除。然后进行剪枝，删除不相关的 term，以解决过渡拟合问题。最后进行信息素更新，对下一只蚂蚁施加影响。蚁群算法的分布性、灵活性和自组织性等特征，使得蚁群算法适合本质是分布、动态和需要内部容错问题求解。

3.1 分类规则挖掘中蚂蚁构造规则的方法

规则生成[4,6]是构造规则的核心操作。规则的生成使用从一般到特殊的策略进行规则增长，先建立一个初始规则 r：{}→y，其中左边是一个空集，右边包含目标类。m 只蚂蚁依次从类标号出发，重复选择节点，一次加入一个前件提高规则质量。在规则的增长过程中，属性被选择的概率由式(1)决定。

$$Pi_j(t)=\frac{\tau_{ij}(t)\eta_{ij}}{\sum_{i}^{a}\sum_{j}^{bi}\tau_{ij}(t)\eta_{ij}} \quad \forall i\in I \tag{1}$$

其中，$\tau_{ij}(t)$ 表示 termij 上的信息素值，η_{ij} 为 termij 的启发信息值，a 表示属性的数量，bi 表示属性 i 的取值个数，I 表示没有被加入的规则中的属性集合。

初始化时，所有路径的信息素都是相同的，其定义如下式：

$$\tau_{ij}(t=0)=\frac{1}{\sum_{i=1}^{a}b_i} \tag{2}$$

每条路径最初的信息素与所有属性值的数量呈反比，使得蚁群搜索出更简洁的规则。

其中，每个属性节点的启发函数值 ηij 为：

$$\eta_{ij}=\frac{C-E(d_{ij})}{\sum_{i=1}^{a}\sum_{j=1}^{b_i}(C-E(d_{ij}))} \tag{3}$$

$E(d_{ij})$ 是与属性 d_{ij} 相关的熵。算法计算每个属性的信息增益，根据启发函数公式确定属性节点被选择的概率。

根据选择概率选择属性节点加入到规则前件中，当节点的增加是规则不能覆盖足够多的样例时，蚂蚁搜索过程结束。在生成规则后，对规则的泛化误差进行估计，为了改善规则的泛化误差，对规则采取后剪枝的策略，对使规则拟合度有较大提高的属性节点进行移除，当规则不能在改进时终止剪枝步骤。

在建立一条规则时，用局部更新规则对信息素进行更新，增加蚂蚁通过路径的信息素，$\tau_{ij}(t+1)=\rho*\tau_{ij}(t)*Q$，其中 ρ 表示信息素的保留率，Q 表示规则的有效性。

每个蚂蚁生成一条规则，经多次优化，规则当中最有效的规则被保留。然后，它所覆盖的所有样例都要被删除。将最有效的规则加入到规则集中，重复上述迭代过程，直到规则集覆盖数据库中所有样例为止，最后形成一个对于初始训练集的分类规则集。

3.2 调整信息素增量

在蚂蚁建立规则的过程中，信息素的调整根据规则有效性及蚂蚁通过的路径调整。对于那些从未被搜索到的节点的信息量会有较大降低，降低算法全局搜索能力。通过减小 ρ 提高算法的全局搜索能力。而 ρ 的减小又会降低算法的收敛速度。因此对 ρ 自适应调整。具体方法为：将 ρ 乘以常数 C=0.9 得到新的 ρ，如果 ρ < ρ min，则 ρ = ρ min。自适应调整 ρ 既可以保留上次搜索得到的有效信息，在较好区域进行更精细的搜索，加快算法收敛，又可以保证大范围搜索的有效性。

3.3 剪枝

分类规则[4]剪枝的目的是以最少的属性集合和最少的分类规则数来描述对象的分类知识，达到对对象进行有效分类的目的。

设生成规则 ruleA={条件 i，类名，i=1，2，…，n}，则算法步骤如下：

①计算规则适应度 ruleA.Q

②从规则 ruleA 中移去条件 i，生成新的规则 ruleB，计算新规则的适应度 ruleB.Q

③比较 ruleA.Q 和 ruleB.Q，如果 ruleA.Q <ruleB.Q，ruleA=ruleB；如果 ruleA.Q>ruleB.Q，ruleA 不变

④i=i+1,重复上述(1)——(3)步骤，直到执行完对所有条件的操作为止，生成一条最简规则。

3.4 算法流程描述：

算法流程直接采用文献[6]中提供的过程：

```
Training Set=all training case;
WHILE(No of uncovered case in the Training Set>Max_Uncovered_Cases)
    i=0;
    REPEAT
        i=i+1;
        Arule= ConsructRule(i);
        PruneRule(Arule);
        UpdatePheromene(Arule);
    Until(i>No_of_ants)or(j>=No_ rules_ converg);
    Select the best rule among all constructed rules;
    Remove the cases correctly covered by the selected rule from the training set;
End While
```

4 算法仿真实验

为了研究改进的蚁群算法在网页分类中的有效性，以从腾讯网站上下载的 180 篇网页进行实验，人工分为 6 个类：经济，政治，文化，科技，军事，娱乐。每一类的训练样本数量都是 20，测试样本数量都是 10，采用 C4.5 算法改进后的蚁群算法分别对数据集进行仿真试验。试验中取 $\alpha=0.1$, $\beta=0.1$, $\rho_{min}=0.1$，Max_Uncovered_Cases=4，No_of_ants =2000。试验结果如表所示：

实验结果充分说明，改进的蚁群算法发现的分类

(下转第 147 页)

(上接第 154 页)

规则集在准确率上有了较大的提高，发现的规则数量更少，分类精度更高，规则也更短；在发现规则的速度上也有较大提高。

表 1　蚁群算法仿真分类结果

算法	规则集大小	准确率（%）	训练计算时间（s）
C4.5	17	75	163
改进蚁群算法	9	83.33	187

5　结论

本文将自适应的调整信息素增量的蚁群规则挖掘算法应用到网页分类中。通过实验验证算法能够发现更好的分类规则，能够获得较高的网页分类的准确率，表明该算法方案具有可行性。

参考文献

1 毛国君,段立娟,王实,石云.数据挖掘原理与算法.北京:清华大学出版社. 2006:64－153,211－251.

2 Dorigo M, Gambardella LM. Ant colony system:a cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computing, 1997,1(1):53－56.

3 李晓明,闫宏飞,王继民. 搜索引擎—原理、技术与系统.北京:科学出版社, 2004:271－290.

4 李鹏,王自强,邝艳敏.基于改进蚁群算法的分类规则挖掘.农业网络信息, 2007,10:13－15.

5 张惟皎,刘春煌,尹晓峰. 蚁群算法在数据挖掘中的应用研究.计算机工程与应用, 2004,(28):171－173.

6 Holden N,Freitas A.Web page classification with an ant colony algorithm, 2004:18－22.