改进蚁群算法 MMAS 在分类规则挖掘中的研究

陈宝钢[1],唐 飞[2],蔡 铁[2],陆芸婷[2],刘寿强[3]
(1. 河南农业大学 信息与管理科学学院,河南 郑州 450046;
2. 深圳信息职业技术学院,广东 深圳 518029;
3. 华南师范大学 物理与电信工程学院,广东 广州 510006)

摘 要:为深入研究和评估蚁群算法在分类规则挖掘应用中具有的特点和作用,针对目前基本蚁群算法在数据挖掘方面所存在的不足,引入了改进的蚁群算法模型最大最小蚂蚁系统(MMAS)。并根据分类算法比较原则,通过实验分析对分类规则挖掘算法进行比较。根据使用不同数据集实验结果的对比分析,从仿真的精确度、速度等方面展示和证实了基于改进的蚁群算法模型 MMAS 的数据分类规则挖掘工具 AntMiner+在分类规则挖掘中体现出的特点和优势。

关键词:数据挖掘;分类规则;蚁群算法;最大最小蚂蚁系统;AntMiner+
中图分类号:TP312 **文献标识码**:A **文章编号**:1673-629X(2014)06-0179-05
doi:10.3969/j.issn.1673-629X.2014.06.044

Research on Improved Ant Colony Algorithm MMAS in Classification Rule Mining

CHEN Bao-gang[1], TANG Fei[2], CAI Tie[2], LU Yun-ting[2], LIU Shou-qiang[3]
(1. College of Information and Management Science, Henan Agriculture University, Zhengzhou 450046, China;
2. Shenzhen Institute of Information Technology, Shenzhen 518029, China;
3. School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, China)

Abstract: In order to study and evaluate the features and functions of ant colony algorithm in classification rule mining applications in-depth, aiming at the deficiencies of the basic ant colony algorithm, introduce improved ant colony algorithms, the Max-Min Ant System (MMAS). And according to the comparison principle of the classification rule mining algorithm, make a comparison for classification rule mining algorithms through the experimental analysis. The results show that the AntMiner+ based on MMAS model has great advantages in the classification rule mining from simulation accuracy and speed.

Key words: data mining; classification rule; ant colony algorithm; Max-Min Ant System (MMAS); AntMiner+

0 引 言

蚁群算法是由意大利学者 M. Dorigo 等人通过模拟自然界中蚂蚁集体寻食的行为而提出的一种人工智能算法[1]。蚁群算法具有较强的鲁棒性,对基本蚁群优化算法模型略加修改,就可以应用到其他问题。目前,蚁群算法在数据挖掘方面的分类模型和聚类模型中的应用已经引起大量关注。

1 基于蚁群算法的分类规则挖掘算法

2002 年,R. S. Parepinelli 等首次结合蚁群算法,提出了基于蚁群算法的分类规则发现系统(Ant-Miner)[2]。蚁群算法用于构造规则的过程具体体现为分成三个阶段来构造一条路径。首先从一条空的路径开始重复选择路径节点增加到路径上,直到得到一条完整的路径,也即一条分类规则;然后进行规则的剪枝,以考虑分类规则对样例的过度拟合的问题;最后更新所有路径上的外激素浓度,对下一只蚂蚁构造规则施加影响[3-5]。生成一个规则后,它所覆盖的训练样本将从训练集中删除。

收稿日期:2013-09-30 **修回日期**:2014-01-02 **网络出版时间**:2014-02-24
基金项目:广东省自然科学基金项目(S2011010003890,S2013010012669,S2011010006115);深圳市科技计划项目(JC201105190829A);河南省科技攻关计划项目(11210221019)
作者简介:陈宝钢(1973-),男,博士,讲师,CCF 会员,主要从事 P2P 网络、无线传感器网络、人工智能等方面的研究。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0930.062.html

1.1 规则生成

规则生成模仿了蚂蚁的爬行行为,这是搜索的核心操作。通常可以设计一个与问题相关的启发式函数,来引导蚁群的搜索行为。定义每个属性节点Term_{ij}的启发式函数值η_{ij}为:

$$\eta_{ij}=\frac{|\mathrm{Term}_{ij}|}{|\mathrm{Trainingset}|}D_{\mathrm{N}}(P_{\mathrm{AI}},P_{\mathrm{C}}) \tag{1}$$

其中,|Trainingset|为训练集中的样例数;$|\mathrm{Term}_{ij}|$为训练集中属性Term_{ij}取值为j的样例数。当第一只蚂蚁开始构造路径时,所有路径节点的信息素浓度用下式表示:

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

其中,a表示数据库中的属性总数;b_i表示属性i的所有可能取值的数量。

在规则的增长过程中,属性被选择的概率公式如下:

$$P_{ij}=\frac{\eta_{ij}\tau_{ij}(t)}{\sum_{i}^{a}\sum_{j}^{b}\eta_{ij}\tau_{ij}(t)} \tag{3}$$

选出的属性节点将被加入到路径中去。其中,η_{ij}为启发信息,即为每个条件项的启发式参数值;$\tau_{ij}(t)$表示第i个属性的第j个取值在t时刻的信息痕迹量。

当所有的属性都包含在路径当中,蚂蚁重复选择属性节点的过程结束,接着它会选择一个类标号节点,形成一条完整的规则,并且该规则的有效性最大。规则的有效性按下面公式计算:

$$Q=\left(\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}\right)*\left(\frac{\mathrm{tn}}{\mathrm{fp}+\mathrm{tn}}\right) \tag{4}$$

其中,tp 为规则前件后件都符合的样例数;fp 为符合规则前件但不符合规则后件的样例数;fn 为符合规则后件但不符合规则前件的样例数;tn 为既不符合规则前件也不符合规则后件的样例数。

1.2 分类规则剪枝

在规则产生之后要对其进行剪枝的处理。剪枝的策略可以是:移去使规则的有效性得到最大提高的属性节点,直到任一属性节点的移去将会降低规则的有效性。当从规则中移去属性节点而使规则改变时,可能需要重新给它赋予一个类标号节点,以使规则的有效性仍然为最大。

1.3 信息素更新

当一只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要按照下面的公式更新:

$$\tau_{ij}(t+1)=\tau_{ij}(t)+\tau_{ij}(t)*Q,\forall\,\mathrm{term}_{ij}\in\text{the rule} \tag{5}$$

信息素的挥发是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的,如下式所列:

$$\tau_{ij}(t+1)=\frac{\tau_{ij}(t)}{\sum_{i}^{a}\sum_{j}^{b}\tau_{ij}},\forall\,i,j\notin\mathrm{Rule} \tag{6}$$

2 改进的蚁群算法模型 MMAS

最大最小蚂蚁系统(MMAS)是德国科学家 Thomas Stützle 等提出的[6]。它的基本思想是在信息素更新过程中,只允许最好的解增加信息素以实现对已有经验的利用,并利用一个限制信息素强度的简单机制,避免了蚂蚁过早地集中到同一条路径上。且通过加入局部搜索,MMAS 很容易进行拓展。MMAS 模型在蚁群系统基础上进行改进,主要不同在对信息素的控制和管理上[7-9]。如下:

(1)为了充分利用循环最优解和到目前为止已经找出的最优解,在每次循环之后,只有一只蚂蚁进行了信息素的更新。在 MMAS 模型中,经修改的轨迹更新规则如下:

$$\tau_{ij}(t+1)=(1-\rho)\tau_{ij}(t)+\triangle\tau_{ij}^{\mathrm{best}}(t)\quad\rho\in(0,1) \tag{7}$$

$$\triangle\tau_{ij}^{\mathrm{best}}=1/f(s^{\mathrm{best}}) \tag{8}$$

其中,$f(s^{\mathrm{best}})$表示迭代最优解或者全局最优解的值。

(2)为了避免算法的过早收敛,对算法中的信息素浓度引入了最值(最大值和最小值)的限制。即规定$\tau_{ij}(t)$在$[\tau_{\min},\tau_{\max}]$区间内,满足下面公式:

$$\tau_{\min}\leqslant\tau_{ij}\leqslant\tau_{\max}(\tau_{\min}>0) \tag{9}$$

当每次循环结束之后信息素更新时,如果$\tau_{\max}\leqslant\tau_{ij}$,则$\tau_{ij}=\tau_{\max}$;同样,$\tau_{\min}\geqslant\tau_{ij}$,则$\tau_{ij}=\tau_{\min}$。若 MMAS 收敛,信息素量最大的解元素所构造的解将与算法找出的最优解相一致。

(3)信息素轨迹初始化为$\tau_{\max}$,可在开始循环时扩大蚁群的搜索范围,增强对收敛性的控制。信息素挥发因子为ρ,第一次迭代后,信息素增强的路径上信息素最大值也是$\tau_{\max}$,而其他没有得到信息素更新的路径的信息素浓度为$\tau_{\max}*(1-\rho)$,两者的最大差距为$\triangle\tau=\rho*\tau_{\max}$。那么,第二次迭代后搜索路径上信息素的最大差距是:

$$\triangle\tau=(2\rho-\rho^2)*\tau_{\max} \tag{10}$$

(4)信息素轨迹的平滑化。在 MMAS 中为了提高性能和进一步解决算法的收敛性问题,MMAS 引入了一种信息素的平滑化机制,它的表达公式为

$$\tau_{ij}^{*}(t)=\tau_{ij}(t)+\delta*[\tau_{\max}(t)-\tau_{ij}(t)] \tag{11}$$

其中,$0<\delta<1$,$\tau_{ij}(t)$和$\tau_{ij}^{*}(t)$分别为平滑化之前和之后的信息素轨迹量。

通过对δ作下面的设置,可以得到不同的算法特

性:

1)当 $\delta=1$,信息素轨迹恢复成 τ_{max},即重新初始化;

2)当 $\delta=0$,关闭平滑化机制;

3)当 $0<\delta<1$,算法运行过程中所积累的信息部分保留。

3　改进蚁群算法模型在分类规则挖掘中的对比分析

3.1　分类规则挖掘算法比较原则

不同分类方法对同一个数据集分类结果可能不同。可以从几个方面对分类方法进行比较:

(1)分类的精度。对于预测型的分类任务,分类精度是指元组被正确分配到其所在的类别中的个数占元组总个数的比例。

(2)分类速度。

(3)模型描述的简洁性和可解释性。模型描述越简洁,则越容易理解,也越受欢迎。可解释性是指分类的结果要让人容易看懂,尽量以可视化的方式或规则展示出来。

(4)分类模型对各种数据具有的适应度。由于所分析的数据对象中可能会存在噪声数据、不完整数据、不一致的数据或者分布稀疏的数据,因此好的分类器需要对各种不同类型的数据集都要具有较强的适应能力。

(5)可伸缩性。可伸缩性是指分类算法对于海量数据应具有能够有效构建模型的能力。

3.2　分类规则挖掘算法的比较

WEKA 是一款非商业化的,基于 JAVA 环境下开源的机器学习和数据挖掘软件[10]。对应决策树分类算法、贝叶斯分类算法、关联规则算法、支持向量机算法,在 WEKA 里面用的是 J48,NaiveBayes,Jrip 和 SMO 分类算法。GUI Ant-Miner 模拟器是改进的 Ant-Miner,它是基于 Parpinelli,Lopes 和 Freitas 在 2002 提出的用来模拟蚁群算法的 Ant-Miner 工具改进得到的[11]。而 AntMiner+实现了 MMAS 算法,而且是第二种基于蚁群算法的数据分类规则挖掘算法[12]。AntMiner+在每次迭代中只有最好的蚂蚁可以进行信息素更新加强,可以更好地利用发现的最好的解决办法[12-14]。

(1)WEKA、GUI Ant-Miner 和 AntMiner+的比较。

当采用 iris、glass、wine、wisconsin-breast-cancer、auto-mpg、ttt、balance-scale、artificial 等八个数据集和设置蚂蚁数为 100 时,比较三种工具 WEKA、GUI Ant-Miner、AntMiner+对数据分类效率的影响,见表 1。

从表 1 和图 1、2 可见,当用 WEKA 中的分类算法时,所用的时间很多都不够 1 s,说明 WEKA 技术对分类规则挖掘的准确度和速度都比较高,但是对于伸缩性和分类模型的简洁度来讲,比不上 AntMiner+。而用 AntMiner+处理数据集所达到的精确度比较高,所用的时间也比较短,模型也简单易于理解,对各种数据集的

表 1　设置蚂蚁数为 100 时,三种分类规则挖掘工具比较

参数	数据集	算法					
		GUI Ant-Miner	AntMiner+	J48	NaiveBayes	Jrip	SMO
精确度/%	iris	93.33	95	96.08	94.12	92.16	96.08
	glass	66.98	58.79	65.75	49.32	58.90	72.60
	wine	89.35	91.52	86.89	98.36	86.89	98.36
	wisconsin-breast-cancer	94.42	89.97	94.64	94.96	94.96	95.80
	auto-mpg	72.37	72.93	79.26	68.15	73.33	77.78
	ttt	72.03	99.58	83.13	71.17	97.55	99.08
	balance-scale	68.95	74.98	69.06	66.51	69.34	68.40
	artificial	53.19	69.14	59.41	53.25	36.15	64.25
时间/s	iris	7	5.411 2	0.09	0.02	0.02	0.33
	glass	12	6.558 6	0.02	0.02	0.08	0.2
	wine	21	8.552 2	0.02	0.02	0.02	0.05
	wisconsin-breast-cancer	16	8.132 5	0.02	0.02	0.08	0.13
	auto-mpg	18	7.791	0.03	0.02	0.05	0.16
	ttt	78	8.541	0.03	0.02	0.13	0.77
	balance-scale	17	17.106 2	0.04	0.02	0.02	0.27
	artificial	25	3.431	0.86	0.03	3.97	136.11

适应度也有优势。

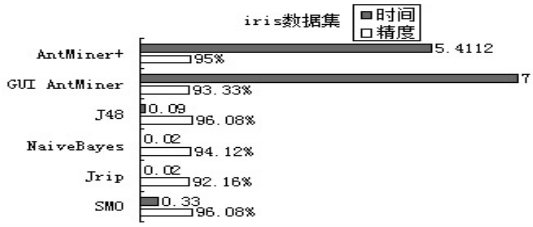

图 1　各种算法处理 iris 数据集对比

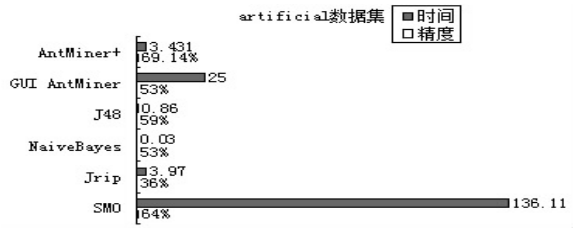

图 2　各种算法处理 artificial 数据集对比

(2) AntMiner+和 GUI Ant-Miner 比较。

使用 AntMiner+和 GUI Ant-Miner 两个工具来分析蚂蚁数为 100 和 1 000 时的运行情况。比较结果如表 2 所示。由表可见,当使用 GUI Ant-Miner 工具模拟时,运行时间变得非常长。无论蚂蚁数是100还是1 000,AntMiner+在分类精确度上都有六个数据集的实验结果超过 GUI Ant-Miner,而且在时间、简洁度以及可伸缩性上和 GUI Ant-Miner 相比都具有绝对的优势。

图 3、图 4 和图 5 是用 AntMiner+和 GUI Ant-Miner 两个工具处理 iris、artificial 和 glass 数据集得到的对比图。由图可见,AntMiner+工具所用的时间比较短,精确度也优于 GUI Ant-Miner 工具,而且 AntMiner+工具比 GUI Ant-Miner 工具更简洁、易于理解。在图 5 中,AntMiner+和 GUI Ant-Miner 在处理 glass 数据集时,GUI Ant-Miner 得到的运算精确度会比 AntMiner+高大概 8%,但是却是以牺牲运行时间以及规则的数目换来的,运行时间大概是 AntMiner+的 2 倍,分类规则数更是 AntMiner+的 1 倍多。可见,AntMiner+在处理数据分类时是比 GUI Ant-Miner 有优势的。

从所有分类结果看到,WEKA 工具在处理分类时精确度比较高,对于小的数据集分类速度也很快。但是,WEKA 分类模型比较复杂,不易于理解,而且某些算法在训练大数据集的时候运行时间变得非常长,如 SMO。而75% 数据集在AntMiner+中的分类精确度都

表 2　AntMiner+和 GUI Ant-Miner 处理数据时的对比分析

参数	数据集	算法			
		ant = 100		ant = 1 000	
		GUI Ant-Miner	Ant-Miner+	GUI Ant-Miner	AntMiner+
精确度/%	iris	93.33	95	93.33	95.12
	glass	66.98	58.79	69.52	60.44
	wine	89.35	91.52	90.33	91.70
	wisconsin-breast-cancer	94.42	89.97	94.62	89.87
	auto-mpg	72.37	72.93	72.37	72.93
	ttt	72.03	99.58	74.03	99.58
	balance-scale	68.95	74.98	68.95	74.98
	artificial	53.19	69.14	52.19	69.24
时间/s	iris	7	5.411 2	47	5.011 2
	glass	12	6.558 6	373	6.541 6
	wine	21	8.552 2	355	8.552 2
	wisconsin-breast-cancer	16	8.132 5	246	8.032 5
	auto-mpg	18	7.791	353	7.791
	ttt	78	8.541	122	8.541
	balance-scale	17	17.106 2	217	17.106 2
	artificial	25	3.431	505	4.431
规则	iris	4.38	1.666 7	4.38	1.666 7
	glass	9	6.25	9	6.25
	wine	5.9	2.75	5.9	2.75
	wisconsin-breast-cancer	7.6	1.750 4	7.6	1.750 4
	auto-mpg	8.6	1.5	8.6	1.5
	ttt	10.5	2.05	10.5	2.05
	balance-scale	8	4.749 6	8	4.749 6
	artificial	23.7	0.5	23.7	0.5

比 GUI Ant-Miner 高;训练时间会比 GUI Ant-Miner 少很多,分类规则也比它少很多。特别是当蚂蚁数很大和数据集实例很多时,AntMiner+在分类精度、分类速度、分类模型的简洁度、伸缩性等方面都表现出绝对的优势。可见,基于最大最小蚂蚁系统的 AntMiner+在处理分类规则挖掘时具有很强的能力和扩展性。

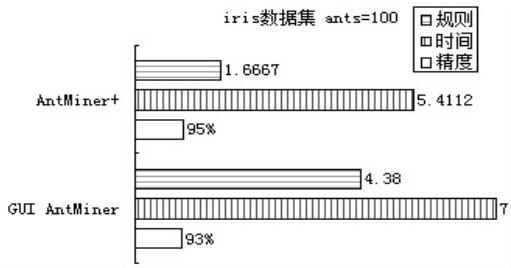

图 3　iris 数据集对比

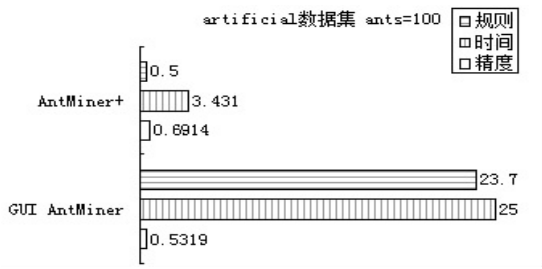

图 4　artificial 数据集对比

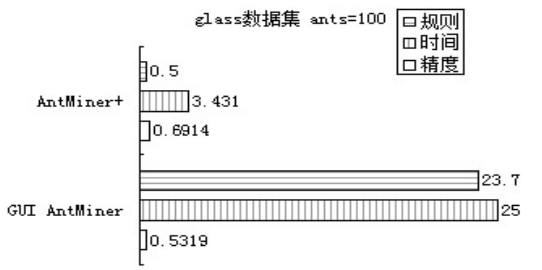

图 5　glass 数据集对比

4　结束语

蚁群算法在数据分类问题上的应用是一个比较新的研究课题。基于最大最小蚂蚁系统(MMAS)实现的 AntMiner+工具在处理分类规则挖掘中有很强的能力,相比原有算法在实验数据的算法优化性上有了显著的提高。但是 AntMiner+工具对环境的要求仍然有些苛刻,当针对不同的数据集时并不能都得出理想的效果。而且随着算法的不断改进和精度的持续提高,算法的时间复杂度和空间复杂度也都会随之增大。如何在基于蚁群算法的分类规则中提高规则的简单性和减少规则的数量是下一步的研究目标。

参考文献:

[1] Colorini A, Dorigo M, Maniezzo V, et al. Distributed optimization by ant colonies[C]//Proceedings of 1st European conf on artificial life. Paris: Elsevier Publishing, 1991: 134-142.

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

[3] 常晓磊,闫仁武. 一种基于蚁群算法的分类规则挖掘算法[J]. 计算机技术与发展,2007,17(7):114-116.

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

[5] 肖　菁,梁燕辉. 基于改进 Ant-miner 算法的分类规则挖掘[J]. 计算机工程,2012,38(17):162-165.

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

[7] 黄丽丰. 基于改进的蚁群算法在分类规则中的应用研究[D]. 重庆:重庆理工大学,2011.

[8] 赵朝卿. 蚁群算法的改进及其应用[D]. 重庆:重庆大学,2008.

[9] 杨剑峰. 蚁群算法及其应用研究[D]. 杭州:浙江大学,2007.

[10] Witten I H, Frank E. 数据挖掘实用机器学习技术[M]. 董琳,邱　泉,于晓峰,等,译. 北京:机械工业出版社,2007.

[11] GUI Ant-Miner[EB/OL]. 2006-02. http://read.pudn.com/downloads53/sourcecode/chinese/184586/GUIAntMiner 1.2.pdf.

[12] Bart M, David M, Manu D B, et al. To tune or not to tune: rule evaluation for metaheuristic-based sequential covering algorithms[R]. Ghent: University Ghent, 2012.

[13] Sumangala K, Nithya G. Comparative study on bio-inspired approach for soil classification[J]. International Journal of Computer Applications, 2012, 38(4): 32-37.

[14] Otero F E B, Freitas A A, Johnson C G. A new sequential covering strategy for inducing classification rules with ant colony algorithms[J]. IEEE Transactions on Evolutionary Computation, 2013, 17(1): 64-76.

(上接第 178 页)

voice and gestures[J]. Pioneering with Science and Technology, 2011, 4(4): 82-83.

[7] 朱　涛,金国栋,芦利斌. Kinect 应用概述及发展前景[J]. 现代计算机(专业版),2013(4):8-11.

[8] 余　涛,叶金永,邵菲杰,等. Kinect 核心技术之骨架追踪技术[J]. 数字技术与应用,2012(10):115-115.

[9] Merry B, Marais P, Gain E. Animation space: a truly linear framework for character animation[J]. ACM Transactions on Graphics, 2006, 25(4): 1400-1423.

[10] Pipho E. Focus on 3D models[M]. [s. l.]: Premier Press, 2003.

[11] 陆　平. 计算机三维动画制作教程[M]. 北京:人民邮电出版社,2005.

[12] 谢　飞. 人体骨骼与蒙皮制造高级应用技法[M]. 北京:北京兵器工业出版社,2009.

改进蚁群算法MMAS在分类规则挖掘中的研究

作者： 陈宝钢， 唐飞， 蔡铁， 陆芸婷， 刘寿强， CHEN Bao-gang， TANG Fei， CAI Tie， LU Yun-ting， LIU Shou-qiang

作者单位： 陈宝钢,CHEN Bao-gang(河南农业大学 信息与管理科学学院,河南 郑州,450046)， 唐飞,蔡铁,陆芸婷,TANG Fei,CAI Tie,LU Yun-ting(深圳信息职业技术学院,广东 深圳,518029)， 刘寿强,LIU Shou-qiang(华南师范大学 物理与电信工程学院,广东 广州,510006)

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(6)