# 一种结合二元蚁群和粗糙集的连续属性离散化算法

曹 峰[1] 唐 超[2] 张 婧[3]

(山西大学计算机与信息技术学院 太原 030006)[1] (合肥学院计算机与科学技术系 合肥 230601)[2]

(太原学院数学系 太原 030006)[3]

**摘 要** 离散化是一个重要的数据预处理过程,在规则提取、知识发现、分类等研究领域都有广泛的应用。提出一种结合二元蚁群和粗糙集的连续属性离散化算法。该算法在多维连续属性候选断点集空间上构建二元蚁群网络,通过粗糙集近似分类精度建立蚁群算法适宜度评价函数,寻找全局最优离散化断点集。通过UCI数据集验证算法的有效性,实验结果表明,该算法具有较好的离散化性能。

**关键词** 离散化,二元蚁群算法,粗糙集

**中图法分类号** TP305 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.09.041

## Algorithm of Continuous Attribute Discretization Based on Binary Ant Colony and Rough Sets

CAO Feng[1] TANG Chao[2] ZHANG Jing[3]

(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)[1]

(Department of Computer and Science Technology, Hefei College, Hefei 230601, China)[2]

(Department of Math, Taiyuan College, Taiyuan 030006, China)[3]

**Abstract** Discretization is an important process of data preprocessing and has been widely applied in the research fields of rule extraction, knowledge discovery, and classification. A discretization algorithm of continuous attribute based on binary ant colony and rough sets was proposed in this paper. The algorithm constructs binary ant colony network on the cut points set generated by multidimensional continuous attributes. Meanwhile, it searches global optimal discretization cut points set by using fitness function constructed with the accuracy of approximation classification of rough sets. To validate the effectiveness of the proposed discretization algorithm, it is applied to seven UCI data sets. And the experimental results indicate that it has relative better performance.

**Keywords** Discretization, Binary ant colony algorithm, Rough sets

## 1 引言

连续属性离散化问题主要研究如何通过选取最优离散化断点集,将连续型属性划分为若干个离散化区间,并为每个区间赋予一个符号值,从而使连续型属性转变为离散型属性[1]。从现实世界中收集到的实体对象的描述属性大多为连续型属性,而数据挖掘、机器学习等研究领域的许多数据分析方法只能处理离散型属性或在离散型属性上表现出更好的数据分析性能,比如粗糙集、决策树、贝叶斯网等[2-3]。为了使这些数据分析方法能有效地应用于连续型属性,对属性进行离散化处理成为必要的数据预处理过程。

属性离散化已受到众多研究者的普遍关注,并取得了一系列的研究成果[2,4-5]。根据离散化过程是否考虑类别信息,离散化算法可以分为非监督和监督两大类[7]。非监督离散化算法在离散化过程中不考虑类别属性,如等间距划分(EW)和等频率划分(EF)等,这类算法易于实现,但往往难以取得令人满意的离散化效果。监督离散化算法在离散化过程中考虑了样本的类别分布信息,使同一离散化区间内的对象具有相同的类别值,而相邻离散化区间的类别值不同。由于监督离散化算法考虑了类别属性,往往比非监督离散化算法有更好的离散化效果,因此目前的离散化算法大多为监督离散化算法。

监督离散化算法主要包括基于信息熵的离散化算法[6,8]、基于卡方度量的离散化算法[9-10]、基于相关性度量的离散化算法[11]和基于粗糙集的离散算法[1,12]等。这些离散化算法的差异主要在于离散化过程中最优断点集的选取方式不同。比如,基于信息熵的离散化算法利用信息熵对断点划分后各离散化区间的类别信息分布状况进行评价,选取信息

到稿日期:2016-08-05 返修日期:2016-12-12 本文受国家自然科学基金项目(41401521,61403238,61502288),山西省青年科技研究基金(2015021101),智能信息处理山西省重点实验室开放课题基金项目(2004001,2016001),安徽高校自然科学研究项目(KJ2015A206),合肥学院人才科研基金项目(15RC07)资助。

曹 峰(1980-),男,博士,讲师,硕士生导师,CCF会员,主要研究方向为离散化、粗糙集、数据挖掘,E-mail:caof@sxu.edu.cn;唐 超(1977-),男,博士,讲师,主要研究方向为机器学习、模式识别、计算机视觉;张 婧(1982-),女,硕士,讲师,主要研究方向为离散化、粗糙集、数据挖掘。

熵最小的断点作为最优断点。基于卡方度量的离散化算法利用统计指数 $\chi^2$ 度量评价两个相邻离散化区间的相似性,并将相似性高的区间进行合并。该算法认为被合并的两个离散化区间之间的候选断点的重要性相对较低。基于相关性度量的离散化算法定义了基于属性与类别间交互依赖关系的指标 CAIM(Class Attribute Interdependence Maximum),在离散化过程中使得 CAIM 值最大的断点被选为较优断点。基于粗糙集的离散化算法以保持数据集不可区分关系不变为依据,在保持信息系统分类能力不变的基础上获取较优的离散化断点集。这些监督离散化算法在对具有多维连续属性的数据集进行离散化时,逐个选择单属性进行离散化,割裂了数据集中多维属性之间的关系。

本文提出了一种结合二元蚁群算法和粗糙集理论的连续属性离散化算法。该算法借助二元蚁群算法可以有效利用反馈信息进行快速全局寻优[13],在多维属性空间上寻找最优断点集,不割裂多维属性之间的关系。同时,该算法将粗糙集的不可区分关系不变作为约束条件,利用二元蚁群算法寻找保持数据集不可区分关系不变的全局最优断点集。通过 UCI 数据集测试本文所提算法的性能,发现其在多个数据集上都表现出较好的离散化效果。

## 2　粗糙集与离散化

### 2.1　粗糙集

粗糙集理论是一种可以有效分析不精确性、不完整性等不完备信息的数据分析和推理方法[14]。目前,粗糙集在人工智能以及其他领域已得到了广泛的应用。下面给出粗糙集的一些基本概念。

设 $S=\langle U,A\cup\{d\},V,f\rangle$ 为一个决策信息系统,也称决策表。其中,$U$ 为非空有限对象组成的集合,称为论域;$A$ 为非空有限属性组成的集合,称为条件属性;$d$ 为决策属性集合,$V=\bigcup_{a\in A\cup\{d\}}V_a$ 表示所有属性上的取值构成的集合;$f$ 是信息函数。

**定义 1**[15]　对任意对象集合 $X\subseteq U$ 和属性集合 $B\subseteq A$,$X$ 关于 $B$ 的下近似和上近似为:

$$B_-(X)=\{x\in U|[x]_B\subseteq X\} \tag{1}$$

$$B^-(X)=\{x\in U|[x]_B\cap X\neq\varnothing\} \tag{2}$$

其中,$[x]_B$ 表示包含对象 $x$ 的由属性集 $B$ 导出的等价类。$B_-(X)$ 指一定能包含于 $X$ 的对象的集合,$B^-(X)$ 指一定能和可能包含于 $X$ 的对象的集合。

**定义 2**[15]　对任意对象集合 $X\subseteq U$ 和属性集合 $B\subseteq A$,1)当且仅当 $B_-(X)=B^-(X)$ 时,集合 $X$ 是 $B$ 可定义集;2)当且仅当 $B_-(X)\neq B^-(X)$ 时,集合 $X$ 是 $B$ Rough 集。

**定义 3**[15]　集合簇 $F=\{X_1,X_2,\cdots,X_n\}(U=\bigcup_{i=1}^{n}X_i)$ 是论域 $U$ 上定义的知识,$B\subseteq A$ 是一个属性子集,$B$ 对 $F$ 的近似分类的精度 $d_B(F)$ 为:

$$d_B(F)=\frac{\sum_{i=1}^{n}|B_-(X_i)|}{\sum_{i=1}^{n}|B^-(X_i)|} \tag{3}$$

近似分类精度 $d_B(F)$ 定义了一种使用当前属性子集 $B$ 对对象进行分类时,近似分类能力的度量指标。当 $d_B(F)=1$ 时,$X_i$ 的边界域为空,此时 $X_i$ 为精确集,表明使用属性子集 $B$ 可以精确地对论域中的对象进行分类。$d_B(F)$ 越接近 1,使用属性子集 $B$ 对论域中的对象进行分类的精确度越高。

### 2.2　离散化

离散化问题主要研究如何有效地将连续型属性转变为离散型属性。如在信息系统 $S$ 中,对任一连续型条件属性 $a_i\in A$,设其取值范围为 $[V_{\min},V_{\max}]$。断点集 $C=\{c_1,c_2,\cdots,c_m|c_i\in[V_{\min},V_{\max}]\}$ 可以将属性 $a_i$ 的取值区间划分为 $m+1$ 个子区间,并用 $m+1$ 个标签代替对应区间内的所有原始值。$C$ 中每个元素 $c_i$ 都是一个断点,这些断点将连续型属性 $a_i$ 离散化为具有 $m+1$ 个属性值的离散型属性。离散化本质上可以归结为利用选取的最优断点集划分连续型属性取值空间,得到最优的离散化区间。不同离散化算法的主要差异在于如何确定最优离散化断点集。最优离散化问题,其实就是寻找最优断点集问题。由于离散化过程是从初始候选断点集中选取最优断点集,因此首先需要确定候选断点集。同样,在信息系统 $S$ 中,设属性 $a_i\in A$ 的属性值构成集合 $V_i=\{a_i^1,a_i^2,\cdots,a_i^{|V_i|}\}$,其中 $a_i^1<a_i^2<\cdots<a_i^{|V_i|}$,则属性 $a_i$ 上的所有候选断点构成的集合可定义为 $C_i=\{c_i^j|c_i^j=(a_i^j+a_i^{j+1})/2\}$,其中 $1\leqslant j\leqslant|V_i|-1$。对每一个属性 $a_i\in A$ 都进行上述操作,最终得到由所有属性上的候选断点构成的候选断点集。假设属性 $A$ 包含 $k$ 个属性,则候选断点集为 $C_{candidate}=\bigcup_{i=1}^{k}C_i$。

## 3　结合二元蚁群算法和粗糙集的离散化

### 3.1　二元蚁群算法

二元蚁群算法是在基本蚁群算法的基础上引入二进制编码而形成的一种求解方法[16]。在二元蚁群算法中,蚁群进化的过程就是寻找最优解的过程。蚂蚁根据路径上的信息素浓度决定前进的方向,前进的同时在路径上留下信息素以影响后代蚂蚁的前进方向。当连续几代蚁群的最优适宜度值相同或差别小于预定阈值时,认为蚁群达到了收敛状态,此时的路径就是最优路径。

二元蚁群算法中,蚂蚁在二维空间(见图 1)中搜索,得到以二进制串表示的最优解。

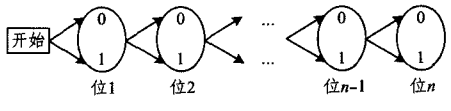


图 1　二元蚁群算法中蚂蚁的搜索空间

使用 $\Gamma$ 表示每一代蚂蚁的信息素浓度矩阵,$\Gamma_{ij}^{gen}$ 表示第 $gen$ 代蚂蚁在位 $i$ 和位 $i+1$ 之间的路径 $j\in\{0,1\}$ 上的信息素浓度。在二维搜索空间中,从每一位出发只有两条路径,即 $\Gamma$ 为一个 $2*n$ 的矩阵。初始时 $\Gamma$ 为一个常数,即所有路径上的初始信息素浓度都相等。

使用 $\Delta\Gamma$ 表示蚂蚁的信息素增量矩阵,$\Delta\Gamma_{ij}^{ant}$ 表示当代第 $ant$ 只蚂蚁在位 $i$ 和位 $i+1$ 之间的路径 $j\in\{0,1\}$ 上的信息素增量。$\Delta\Gamma$ 是一个初始值为 0 的 $2*n$ 的矩阵。在新一代蚁群开始搜索前,根据式(4)更新信息素浓度矩阵 $\Gamma$ 并将 $\Delta\Gamma$ 清零。

$$\Gamma_{ij}^{gen+1}=\rho\Gamma_{ij}^{gen}+\sum_{ant=1}^{numAnt}\Delta\Gamma_{ij}^{ant} \tag{4}$$

*gen* 为当前代数，*numAnt* 为蚂蚁总数，$\rho$ 为信息素持久度。

在搜索的过程中，蚂蚁根据式(5)和式(6)决定转移方向：

$$P_{i0}=\frac{[\Gamma_{i0}]^{\alpha}*[\eta_{i0}]^{\beta}}{[\Gamma_{i0}]^{\alpha}*[\eta_{i0}]^{\beta}+[\Gamma_{i1}]^{\alpha}*[\eta_{i1}]^{\beta}} \tag{5}$$

$$P_{i1}=1-P_{i0} \tag{6}$$

其中，$P_{i0}$ 和 $P_{i1}$ 分别为蚂蚁根据信息素浓度选择位 $i$ 到位 $i+1$ 的路径 0 和 1 的概率；$\Gamma_{i0}$ 和 $\Gamma_{i1}$ 分别表示路径 0 和 1 上的信息素浓度；$\eta_{i0}$ 和 $\eta_{i1}$ 分别表示路径 0 和 1 的可见度；$\alpha$ 和 $\beta$ 为权重调节因子，$\alpha,\beta\geqslant 0$。

### 3.2 适宜度函数

在二元蚁群算法中，适宜度函数是评价蚁群搜索结果优劣程度的重要指标。而在最优监督离散化问题中，要求在保证决策表相容性的前提下，选取尽可能少的断点。基于此，本文综合考虑近似分类精度和断点数两个因素，设计适宜度评价函数，用于在离散化过程中寻找尽可能保持决策表相容性且断点数较少的最优断点集。假设通过某次搜索得到的属性集 $B$ 的断点集为 $C$，由决策属性得到的对象集 $U$ 的划分为 $F$。根据式(1)－式(3)求出近似分类精度 $d_B(F)$，蚂蚁搜索得到的断点集 $C$ 的适宜度函数为：

$$fitness(C)=Q\cdot\frac{d_B(F)}{|C|^N} \tag{7}$$

其中，$Q$ 和 $N$ 为大于 0 的常数，$|C|$ 为最优断点集 $C$ 的断点数。

$Q$ 值的大小反映了前代蚁群经验对后代进化的影响程度，$Q$ 值越大，后代蚁群越容易受到前代蚁群的影响，蚁群收敛速度越快；$N$ 为权重调节因子；$d_B(F)$ 反映了断点集 $C$ 对决策表相容性的影响程度，$d_B(F)$ 越接近 1，越能保证原决策表的相容性。

一只蚂蚁完成搜索后，会在经过的路径上留下新增的信息素，可以使用适宜度值衡量该蚂蚁产生的信息素的增量。如果蚂蚁通过路径 $j\in\{0,1\}$ 从位 $i$ 到达位 $i+1$，则信息素增量 $\Delta\Gamma_{ij}=fitness(C)$；反之不留下信息素，即信息素增量 $\Delta\Gamma_{ij}=0$。

### 3.3 算法设计

利用二元蚁群算法寻找最优断点集时，首先对初始候选断点集的选择状态进行二进制编码，0 表示舍弃该断点，1 表示保留该断点。对于由 $k$ 个连续条件属性构成的信息系统 $S$，蚁群的搜索路径为长度 $L$ 的 0-1 串，记为 *bin*，其中 $L=\sum_{i=1}^{k}|C_i|$，$C_i$ 为第 $i$ 个属性的断点集。

为了改善蚁群的搜索能力，在决定蚂蚁的转移方向时加入一定的随机性，蚂蚁在搜索过程中根据式(8)决定路径 *bin* 上位 $i$ 的取值：

$$bin[i+1]=\begin{cases}0, & p_{i0}\cdot r_1<p_{i1}\cdot r_2\\ 1, & \text{else}\end{cases} \tag{8}$$

其中，$i$ 为蚂蚁当前所在的位($0\leqslant i\leqslant L-1$)，$p_{i0}$ 和 $p_{i1}$ 是由式(5)和式(6)求出的转移概率，$r_1$ 和 $r_2$ 是[0,1]内的两个随机数。$i=0$ 时蚂蚁位于开始位。下面给出算法的实现步骤。

输入：决策信息系统 S；

输出：最后一代蚁群中适宜度值最大的蚂蚁的路径对应的断点集为全局最优断点集 $C_{opt}$

1. 由 S 的多维连续条件属性计算候选断点集 $C_{cadidate}$；
2. 初始化蚂蚁数 AntNum、进化代数 MaxGen、蚁群最优适宜度值数组 $Fitness^{gen}$[MaxGen]、蚁群信息素浓度矩阵 Γ、蚂蚁的信息素浓度增量矩阵 ΔΓ、蚂蚁路径的可见度 η、权重调节因子 α 和 β、适宜度函数中的参数 Q 和 N；
3. for m=1,2,…,MaxGen
4. for n=1,2,…,AntNum
5. for i=1,2,…,|$C_{cadidate}$|
6. 根据式(5)和式(6)确定蚂蚁 n 的路径 bin；
7. end for
8. 将蚂蚁 n 的路径 bin 映射为断点集 C；
9. 根据式(7)计算蚂蚁 n 的适宜度值 fitness(C)；
10. if fitness(C)>$Fitness^{gen}$[m] then
11. $Fitness^{gen}$[m]=fitness(C)；
12. end if
13. 计算蚂蚁 n 在路径 ij 的信息素浓度增量：
14. $\Delta\Gamma_{ij}^{n}=\begin{cases}\text{fitness(C)}, & \text{蚂蚁经过路径 ij}\\ 0, & \text{否则}\end{cases}$
15. end for
16. 根据式(4)更新蚁群信息素浓度矩阵 Γ；
17. 将每一只蚂蚁的信息素浓度增量矩阵 ΔΓ 清零；
18. if 种群连续若干代最优适宜度值 $Fitness^{gen}$[m]相等或差别小于阈值时 then
19. break；
20. end if
21. end for

## 4 实验

### 4.1 实验数据

实验采用 7 个常用的 UCI 数据集，如表 1 所列。数据集中既包括连续属性个数较多的数据，也包括连续属性个数较少的数据。

表 1 实验数据集

| 数据集 | 样本数 | 连续属性数 | 决策类别数 |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| Machine | 209 | 7 | 8 |
| Breast | 699 | 9 | 2 |
| Glass | 214 | 9 | 7 |
| Heart | 270 | 6 | 2 |
| Wine | 178 | 13 | 3 |
| Sonar | 208 | 60 | 2 |

### 4.2 实验结果与分析

为了验证本文提出的离散化算法的有效性，比较了本文提出的离散化算法与等频率算法、等间距算法、基于熵的 MDL 算法[17]、CAIM 算法[11]、ModifiedChi2 算法[9] 以及 SMDNS[12]算法对分类器的分类精度的影响。评价离散化效果所使用的分类器为决策树分类器 C4.5[18] 和贝叶斯网分类器 BayesNet[19]。为了减少随机性导致的实验误差，本文计算分类精度时采用十倍交叉验证的方法，计算平台是 Weka 数

据挖掘软件。在等频率和等间距算法中，离散化区间数设为5。本文提出算法的初始基本参数设置为：蚁群信息素浓度矩阵 $\Gamma$ 的取值为 0.5，蚂蚁的信息素浓度增量矩阵 $\Delta\Gamma$ 的取值为0，蚁群最优适宜度数组 $Fitness^{gen}[MaxGen]$ 的取值为 0，蚂蚁数 $AntNum=50$，进化代数 $MaxGen=100$，信息素持久度 $\rho=0.5$，蚂蚁二值路径的可见度 $\eta_{i0}=2$，$\eta_{i1}=1$，权重调节因子 $\alpha=\beta=1$，适宜度函数中参数 $Q=1$，$n=2$。当连续 5 代蚁群搜索到的路径相同时，认为蚁群达到了收敛。不同离散化算法下 C4.5 分类器和 BayesNet 分类器的分类精度对比结果如表 2 和表 3 所列。

表 2　不同离散化算法下 C4.5 分类器的分类精度对比结果/%

| 数据集 | 未离散化 | 等宽 | 等频 | MDL | CAIM | Modified Chi2 | SMDNS | 本文算法 |
|---|---|---|---|---|---|---|---|---|
| Iris | 96.0 | 93.3 | 94.7 | 94.0 | 94.0 | 95.3 | 93.3 | **97.3** |
| Machine | 88.5 | 80.9 | 80.9 | 85.6 | 84.2 | 86.1 | 81.8 | **91.9** |
| Breast | 94.6 | 94.8 | 95.9 | 95.0 | 96.2 | 95.8 | 95.2 | **97.5** |
| Glass | 66.8 | 54.7 | 63.1 | 73.8 | **79.0** | 78.5 | 68.7 | 73.8 |
| Heart | 76.7 | 75.2 | 80.4 | 81.9 | 80.4 | 77.0 | 80.4 | **84.1** |
| Wine | 93.8 | 91.6 | 83.7 | 93.8 | 97.8 | **98.2** | 92.1 | 97.8 |
| Sonar | 71.2 | 71.6 | 70.7 | **79.8** | 79.3 | 72.6 | 73.1 | 78.4 |

表 3　不同离散化算法下 BayesNet 分类器的分类精度对比结果/%

| 数据集 | 未离散化 | 等宽 | 等频 | MDL | CAIM | Modified Chi2 | SMDNS | 本文算法 |
|---|---|---|---|---|---|---|---|---|
| Iris | 92.7 | 93.3 | 93.3 | 94.0 | 94.0 | 96.7 | 92.7 | **97.3** |
| Machine | 82.8 | 85.2 | 77.5 | 89.0 | 90.0 | **91.4** | 83.7 | **91.4** |
| Breast | 97.1 | 97.3 | 97.0 | 97.0 | 97.2 | 97.4 | 96.6 | **97.7** |
| Glass | 70.6 | 56.1 | 65.0 | 74.8 | **79.4** | 76.2 | 66.8 | 75.2 |
| Heart | 81.1 | 84.8 | 84.8 | 83.3 | 83.7 | **85.2** | 84.1 | **85.2** |
| Wine | **98.9** | 97.8 | 97.8 | **98.9** | **98.9** | 97.8 | 97.8 | **98.9** |
| Sonar | 80.3 | 71.6 | 76.9 | **85.6** | 84.6 | 77.4 | 75.5 | 79.8 |

通过对实验结果进行分析，可以得出如下结论：

(1)首先比较本文所提离散化算法与不进行离散化时 C4.5 与 BayesNet 两种分类器的分类精度。可以看出，使用本文提出的算法进行数据离散化时，除了 BayesNet 分类器在 sonar 数据集上的分类精度略低之外，两种分类器在离散化后的数据集上的精度都是最高的。由此可以看出使用本文所提离散化算法进行数据离散化可以有效提高 C4.5 和 BayesNet 分类器的分类精度。

(2)其次与等频率和等距离这两种非监督离散化算法相比，使用本文所提离散化算法时，C4.5 和 BayesNet 两种分类器在所有的数据集下的分类精度都是最高的。这与非监督的离散化算法不考虑样本的类别分布信息，往往导致离散化效率低有关。因此，本文所提的离散化算法要优于等频率和等间距离散化算法。

(3)最后与 MDL，CAIM 以及 ModifiedChi2 这 3 种目前使用较多的监督离散化算法相比，使用本文所提算法时，C4.5 分类器在7 个数据集中的 4 个数据集上分类精度最高。BayesNet 分类器在 7 个数据集中的 5 个数据集上分类精度最高。而在分类精度不是最高的数据集上，C4.5 分类器和 BayesNet 分类器的分类精度也相对较高。整体上比较，本文所提离散化算法在大部分数据集上的离散化效果要优于所比较的几种监督离散化算法。

**结束语**　本文提出了一种结合二元蚁群和粗糙集的连续属性离散化方法。该算法利用粗糙集的近似分类精度构建蚁群算法的适宜度函数，进而通过蚁群算法在多维属性空间的候选断点集上寻找最优断点集。利用 UCI 数据集测试算法的有效性，可以看出，与几种具有代表性的离散化算法相比，本文提出的算法具有较好的离散化性能，可以辅助 C4.5 和 BayeNet 等分类器获得更高的分类精度。

本文所提离散化算法所采用的粗糙集理论利用精确的包含关系来定义上近似集和下近似集。实际的数据往往含有噪音，如果仍使用精确的包含关系将难以实现对噪声数据的有效容错。Ziarko[20] 提出了一种粗糙集的扩展模型——变精度粗糙集模型。该模型放松了对经典粗糙集理论近似集的定义，通过设置阈值参数，实现了对噪声数据的容错。在以后的研究中，我们将利用变精度粗糙集来代替粗糙集，设计更适用于复杂数据的适宜度评价指标，进而取得更优的离散化结果。

另外，二元蚁群算法存在自身的局限性。当数据集过大时，算法易陷入局部最优，而且在搜索的初期，由于信息素匮乏导致蚁群算法初期的搜索速度较慢。为了克服二元蚁群算法的局限性，我们将采用结合蚁群算法和遗传算法的二进制蚁群进化算法[21]代替蚁群算法实现多维属性断点寻优，提高目前算法在较大数据集上的全局搜索能力。

## 参 考 文 献


[1] HOU L J, WANG G Y, NIE N, et al. Discretization in rough set theory[J]. Computer Science, 2000, 27(12): 89-94. (in Chinese)
侯利娟，王国胤，聂能，等. 粗糙集理论中的离散化问题[J]. 计算机科学，2000，27(12)：89-94.

[2] CAO F. Research on spatial data discretization[D]. Beijing: University of Chinese Academy of Sciences, 2013. (in Chinese)
曹峰. 空间数据离散化研究[D]. 北京：中国科学院大学，2013.

[3] SANG Y. Research on discretization methods for continuous data[D]. Dalian: Dalian University of Technology, 2012. (in Chinese)
桑雨. 连续数据离散化方法研究[D]. 大连：大连理工大学，2012.

[4] DOUGHERTY J, KOHAVI R, SAHAMI M. Supervised and unsupervised discretization of continuous features[C]// Proceedings of the Twelfth International Conference on Machine Learning. Morgan San Francisco: Morgan Kaufmann Publishers, 1995: 194-202.

[5] KHANMOHAMMADI S, CHOU C A. A Gaussian mixture model based discretization algorithm for associative classification of medical data[J]. Expert Systems With Applications, 2016, 58(c): 119-129.

[6] SHI Z C, XIA Y X, ZHOU J Z. Discretization algorithm based on granular computing and its application[J]. Computer Sciences, 2013, 40(6A): 133-135. (in Chinese)
史志才，夏永祥，周金祖. 基于粒计算的离散化算法及其应用[J]. 计算机科学，2013，40(6A)：133-135.

[7] LIU H, HUSSAIN F, TAN C L, et al. Discretization: a enabling technique[J]. Data Mining and Knowledge Discovery, 2002, 6: 393-423.

[8] XIE H, CHENG H Z, NIU D X. Algorithm of continuous attri-

bute discretization for rough set theory based on information entropy [J]. Chinese Journal of Computers, 2005, 28(9): 1570-1574. (in Chinese)
谢宏,程浩忠,牛东晓. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报,2005,28(9):1570-1574.

[9] TAY E H, SHEN L. A modified chi2 algorithm for discretization [J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 666-670.

[10] SU C T, HSU J H. An extended Chi2 algorithm for discretization of real value attributes[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(3): 437-441.

[11] KURGAN L A, CIOS K J. CAIM discretization algorithm[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(2): 145-153.

[12] JIANG F, SUI Y F. A novel approach for discretization of continuous attributes in rough set theory [J]. Knowledge-Based Systems, 2015, 73: 324-334.

[13] QIAN Q, CHENG M Y, XIONG W Q, et al. Reviews of binary ant colony optimization[J]. Application Research of Computers, 2012, 29(4): 1211-1215. (in Chinese)
钱乾,程美英,熊伟清,等. 二元蚁群算法研究综述[J]. 计算机应用研究,2012,29(4):1211-1215.

[14] PAWLAK Z. Rough sets[J]. Communications of the Acm, 1995, 38(11): 88-95.

[15] 王国胤. Rough 集理论与知识获取[M]. 西安交通大学出版社, 2001.

[16] XIONG W Q, WANG L Y, YAN C Y. Binary ant colony evolutionary algorithm[J]. International Journal of Information Technology, 2006, 12(3): 10-20.

[17] FAYYAD U M, IRANI K B. On the handling of continuous-valued attributes in decision tree generation[J]. Machine Learning, 1992, 8(1): 87-102.

[18] QUINLAN J R. C4. 5: programs for machine learning[M]. San Francisco: Morgan Kaufmann, 1993.

[19] KONONENKO I. Naive Bayesian classier and continuous attributes[J]. Informatica, 2010: 317-326.

[20] ZIARKO W. Variable precision rough set model[J]. Journal of Computer & System Sciences, 1993, 46(1): 39-59.

[21] XIONG W Q, WEI P. Binary ant colony evolutionary algorithm [J]. Acta Automatica Sinica, 2007, 33(3): 259-264. (in Chinese)
熊伟清,魏平. 二进制蚁群进化算法[J]. 自动化学报,2007,33(3):259-264.

(上接第 221 页)

以显著减少算法的时间花费,并且不会产生很大的内存使用量。

表 1 时间花费和内存使用量

| 查询 ID | All fuzzy GDMCTs /ms/MB | All fuzzy GDMCTs/ms/MB (未使用索引) |
|---|---|---|
| $DQ_1$ | 1985/89 | 16500/23.7 |
| $DQ_3$ | 1190/73 | 14000/20 |
| $DQ_5$ | 635/60 | 13200/18 |

**结束语** 本文针对模糊 XML 文档上的关键字近似查询方法进行了相关的研究。文中首先引入最小连接树(MCT)、距离最小连接树(DMCT)和成组距离最小连接树(GDMCT)的概念,给出最小连接树可能性值的计算方法。通过对传统的区间编码方式进行扩展以区分模糊 XML 文档中的模糊节点和一般节点,并记录节点在文档中所在路径上的模糊信息。提出关键字近似查询算法 All fuzzy GDMCTs 来求解在给定子树大小阈值 $K$ 和可能性阈值 $U$ 的条件下的 GDMCTs 结果。最后通过实验表明该算法能够获得较高质量的查询结果。

## 参 考 文 献

[1] XU Y, PAPAKONSTANTINOU Y. Efficient Keyword Search for Smallest LCAs in XML Databases[C]// Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2005: 527-538.

[2] GUO L, SHAO F, BOTEV C, et al. XRANK: Ranked Keyword Search over XML Documents[C] // Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2003: 16-27.

[3] LI G L, FENG J H, WANG J Y, et al. Effective Keyword Search for Valuable LCAs over XML Documents[C]// Proceedings of the sixteenth ACM Conference on Information and Knowledge Management. New York: ACM Press, 2007: 31-40.

[4] HRISTIDIS V, KOUDAS N, PAPAKONSTANTINOU Y, et al. Keyword Proximity Search in XML Trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(4): 525-539.

[5] BHALOTIA G, HULGERI A, NAKHE C, et al. Keyword Searching and Browsing in Databases using BANKS[C]// Proceedings of 18th International Conference on Data Engineering. New York: IEEE Press, 2002: 431-440.

[6] LI J, LIU C, ZHOU R, et al. Top-k Keyword Search over Probabilistic XML Data [C]// Proceedings of 2011 IEEE 27th International Conference on Data Engineering. New York: IEEE Press, 2011: 673-684.

[7] ZHOU R, LIU C, LI J, et al. ELCA Evaluation for Keyword Search on Probabilistic XML Data[J]. World Wide Web, 2013, 16(2): 171-193.

[8] MA Z M, LIU J, YAN L. Matching Twigs in Fuzzy XML.[J]. Information Sciences, 2011, 181(1): 184-200.

[9] LIU J, MA Z M, MA R Z. Efficient Processing of Twig Query with Compound Predicates in Fuzzy XML[J]. Fuzzy Sets and Systems, 2013, 229: 33-53.

[10] MA Z M, YAN L. Fuzzy XML Data Modeling with the UML and Relational Data Models[J]. Data & Knowledge Engineering, 2007, 63(3): 972-996.

[11] DBLP[EB/OL]. http://dblp.uni-trier.de/xml.

[12] LI T, MA Z M. Keyword Querying of Fuzzy XML[J]. Journal of Northeastern University: Natural Science, 2016, 37(7): 937-941. (in Chinese)
李婷,马宗民. 模糊 XML 关键字查询方法[J]. 东北大学学报:自然科学版,2016,37(7):937-941.