基于遗传算法的α-离群约简搜索算法*)

金义富[1,2] 朱庆生[1] 邹咸林[1]

(重庆大学计算机学院 重庆 400044)[1] (湛江师范学院计算机系 湛江 524048)[2]

摘 要 离群数据挖掘与分析在网络入侵控制、信用卡检测、通信欺诈分析等诸多领域具有十分重要的意义。结合粗糙集理论的属性约简技术,定义了α-离群约简等概念,提出了一种以属性离群贡献率和离群划分相似水平为基础的基于遗传算法的α-离群约简算法。这种方法通过维数更小的属性子空间去获得相同或相近的离群数据集,使对离群数据来源及出现原因的分析和理解更加集中于较小的目标域。通过对现实数据集的实验表明,该算法可有效地产生出约简并具有较好的规模适应性。

关键词 离群约简,遗传算法,粗糙集,离群相似水平

A Searching Algorithm for α-Outlying Reduction Based on Genetic Algorithm

JIN Yi-Fu[1,2] ZHU Qing-Sheng[1] ZOU Xian-Lin[1]

(College of Computer, Chongqing University, Chongqing 400044)[1]

(Department of Computer, Zhanjiang Normal College, Zhanjiang, Guangdong 524048)[2]

Abstract Mining and analyzing for outliers is of great importance in many applications, including network invasion control, credit card and telecom fraud detection, etc. A concept of α-outlying reduction is defined in the paper based on the approach of attribute reduction in the theory of rough set. Along with the discussion of outlying contribution rate of attributes and the level of outlying partition similarity, this paper proposes a searching algorithm for α-outlying reduction based on genetic algorithm. The approach can help us obtain similar outlier sets by means of searching in an attributes subspace with lesser dimension, which leads to that analyzing for origins and appearance reasons of outliers is focused better on narrow and specific object fields. Experimental results on real world data sets show that the proposed algorithm is scalable and efficient and it can result in optimal eduction.

Keywords Outlying reduction, Genetic algorithm, Rough set, Outlying similarity level

1 引言

离群数据(Outlier)是那些偏离大部分数据分布的数据点[1],从大量甚至海量数据中发现离群点是数据挖掘的重要组成部分。在网络入侵控制、信用卡及通信欺诈检测、气象预测以及很多科学研究领域,离群数据可能比常规数据更有价值,因此离群数据挖掘(Outlier Mining, DM)与分析具有重要意义。现已提出的离群挖掘算法包括基于聚类的、统计的、距离的、深度的以及基于密度的方法等多种类型[2,3]。这些方法挖掘出的离群点均以单个形式给出,较少对挖掘出的离群点进行进一步的分析,如数据为什么会离群,离群是如何产生的等等。本文结合粗糙集(Rough Set)中划分与约简技术[4],探讨在多维数据集中属性域子集对数据离群的贡献度;提出以离群划分相似水平为基础进行离群约简;认为,如果仅在某个属性域子集上即能产生出离群数据集的全部或其中大部分,那么这个域子集就是导致数据离群的全部或主要因素,也即是在离群意义下的一个属性约简。

遗传算法(GA, Genetic Algorithm)由 Michigan 大学的 Holland 在 20 世纪 60 年代末提出,是一类模拟 Darwin 自然进化论和 Mendel 遗传变异理论的仿生优化技术[5]。GA 通过对求解问题进行编码,根据个体的适应度值,按照优胜劣汰与适者生存原则,借助于选择、交叉与变异等遗传算子实现不断进化寻求最优解,它是一种全局导向随机搜索算法[6]。本文以属性离群贡献率为基础生成初始种群,利用遗传算法搜索满足给定离群相似水平的离群约简,即 α-离群约简。

本文第 2 节定义了离群划分及 α-离群约简等概念,第 3 节分析属性重要度并定义离群贡献率,第 4 节详细讨论了基于遗传算法的 α-离群约简算法,描述了算法框架,分析了算法复杂度,第 5 节给出实验结果,最后是本文的结论部分。

2 α-离群约简

离群挖掘与分析系统可表示为一个五元组(U,A,R,B,V,F),其中 U 为论域,代表一组有限对象集 $X=\{x_1,x_2,\cdots,x_n\}$,其中 $x_i=(x_{i1},x_{i2},\cdots,x_{id})$ 为第 i 个数据对象;属性集 $A=\{A_1,A_2,\cdots,A_d\}$ 是一有界域集合,R 为关系集,它使 x_{ij} 对应于域 A_j, $i=1,2,\cdots,n$, $j=1,2,\cdots,d$。B 为决策属性集,它仅有一个值域为{0,1}的属性,代表是否为离群数据对象;V 是对象集 X 的取值约束;F 为分类规则,此处表示一种离群挖掘算法。

我们知道,如果一组集合 $C_k\subseteq X$, $k=1,2,\cdots,K$, $K\leqslant n$, 满足:1)$C_k\neq\Phi$,2)$C_{k1}\cap C_{k2}=\Phi(k_1\neq k_2)$,3)$C_1\cup C_2\cup\cdots\cup C_K=X$, 则$\{C_1,C_2,\cdots,C_K\}$构成 X 的一个划分[4]。为探讨同一

*)重庆市自然科学基金资助项目(2005BB2224)。**金义富** 博士生,副教授,研究方向为智能系统、数据挖掘;**朱庆生** 博士,教授,博士生导师,研究方向为商务智能、图像处理;**邹咸林** 博士生,副教授。

种离群挖掘算法作用于不同属性域子集产生离群对象的能力，首先引入如下概念。

定义 1(离群划分) 记 X 的离群数据集为 XO。若 $XO\neq\Phi$，则称 $\hat{C}=\{XO, X-XO\}$ 为 X 的一个离群划分(Outlying Partition, OP)。

用 $\Omega(X)$ 表示这种离群划分的全体，设 $L^d=A_1\times A_2\times\cdots\times A_d$ 为一个 d 维数字空间，记 $\hat{C}_d=\{XO_d, X-XO_d\}\in\Omega(X)$ 为 L^d 中对论域 U 产生的划分，XO_d 为对象集 X 的离群数据集。

显然，我们也可以在属性子空间上对论域进行离群划分。设 $L^s(j_1,j_2,\cdots,j_s)=A_{j_1}\times A_{j_2}\times\cdots\times A_{j_s}$ 由 L^d 中 $j_1,j_2,\cdots,j_s$ 域构成的 s 维子空间，$\forall x_i\in X$ 为 L^d 中的一个数据对象，记 x_i 在 $L^s(j_1,j_2,\cdots,j_s)$ 中的投影为：

$$p_{j_1,j_2\cdots j_s}(x_i)=(x_{ij_1},x_{ij_2},\cdots,x_{ij_s}) \tag{1}$$

其中 $1\leqslant j_1<j_2<\cdots<j_s\leqslant d$，$s\in\{1,2,\cdots,d-1\}$。在子空间 $L^s(j_1,j_2,\cdots,j_s)$ 上对投影集 $\{p_{j_1,j_2\cdots j_s}(x_i), i=1,\cdots,n\}$ 执行相同的离群挖掘算法，设挖掘出的离群数据集为 XO_s，则 $\hat{C}_s=\{XO_s,X-XO_s\}$ 构成 X 的一个离群划分。

定义 2(等价离群划分) $\forall s、r$，$0<s、r\leqslant d$，设 XO_s、XO_r 分别为根据子空间 L^s、L^r 发现的离群数据集，记 $\hat{C}_s=\{XO_s,X-XO_s\}$，$\hat{C}_r=\{XO_r,X-XO_r\}$，显然 $\hat{C}_s、\hat{C}_r\in\Omega(X)$。若 $XO_s=XO_r$，则称 $\hat{C}_s$ 与 $\hat{C}_r$ 互为等价离群划分，记为 $\hat{C}_s=\hat{C}_r$。

定义 3(离群约简) 设在属性子空间 L^s、L^r 上对应的离群划分分别为 $\hat{C}_s$、$\hat{C}_r$，$0<s、r\leqslant d$。若 $L^s\subset L^r$ 且 $\hat{C}_s=\hat{C}_r$，则称 L^s 为 L^r 的一个离群约简(Outlying ReDuction, ORD)。

离群约简在保证等价离群划分的同时减少了描述离群数据集所需要的属性个数，更进一步明确了离群数据有关含义，但离群约简有时仍然可能具有较高的维数。因此，在实际使用中，除了这种等价关系以外，我们经常会考察更为普遍的近似情况。如果只根据少数几个属性域投影数据集产生的离群划分 $\hat{C}_s=\{XO_s,X-XO_s\}$ 与在 L^d 中的离群划分 $\hat{C}_d$ 十分接近，则有理由对这几个属性特别关注。

设 $w=card(XO_d\cap XO_s)$，card 表示集合的势。记

$$\begin{aligned} Sup &= w / card(XO_d\cup XO_s)\\ Con &= w / card(XO_s)\\ Inc &= w / card(XO_d)\end{aligned} \tag{2}$$

分别表示离群划分 $\hat{C}_s=\{XO_s,X-XO_s\}$ 相对于 $\hat{C}_d$ 的支持度、置信度和包含度，它们从不同角度表明离群数据集 XO_s 与 XO_d 的趋近程度。支持度越大，说明 XO_s 与 XO_d 总体更相似，置信度代表 XO_s 自身的正确程度，而包含度则表示 XO_s 正确反映 XO_d 的程度。

定义 4(α-离群约简) 设 $\hat{C}_d=\{XO_d,X-XO_d\}$ 为 L^d 中对论域 U 产生的离群划分，$\hat{C}_s=\{XO_s, X-XO_s\}\in\Omega(X)$ 为在子空间 L^s 上获得的离群划分，$0<s\leqslant d$，令

$$\alpha(L^s)=\begin{cases}(Sup+Con+Inc)/3, & XO_d\cap XO_s\neq\Phi\\ 0, & XO_d\cap XO_s=\Phi\end{cases} \tag{3}$$

称 L^s 为 L^d 的一个 α-离群约简。

可以看出，参数 α 代表了离群集 XO_s 与 XO_d 的相似程度，从而反映了离群划分 $\hat{C}_s$ 和 $\hat{C}_d$ 的相似程度，因此称 α 为离群相似水平。它的值越大，说明离群划分 $\hat{C}_s$ 和 $\hat{C}_d$ 越趋向于一致。显然，$0\leqslant\alpha\leqslant1$，而 $\alpha=1$ 当且仅当 $\hat{C}_s=\hat{C}_d$，此时若 $s<d$，则 L^s 为 L^d 的离群约简。

3 离群贡献率

属性约简是粗糙集的重要研究内容，离群约简是在离群划分基础上的属性约简，已经证明约简的复杂性是随着问题规模增大呈指数增长的，是一个典型的 NP 完全问题[5]。因此，计算离群约简也是 NP 难的。

现有约简算法主要以粗糙集(Rough Set)理论的可辨识矩阵(Discernibility Matrix)以及单个属性的必要度为基础[4]。可辨识矩阵 M_D 的定义如下：当对象 x_i 和 x_j 不同类，即 $B(x_i)\neq B(x_j)$ 时，$M_D(i,j)=\{A_k\mid A_k\in A$ 且 $A_k(x_i)\neq A_k(x_j)\}$；反之，当 $B(x_i)=B(x_j)$ 时，$M_D(i,j)=0$，$i,j=1,2,\cdots,n$。任给属性集 $P\subset A$，$\mathrm{Ind}(P)$ 是一个不可分辨关系，如果 $\exists x_i, x_j\in X$，$\forall p\in P$，满足 $p(x_i)=p(x_j)$，此时对象 x_i 和 x_j 对于属性集 P 不可分辨。若 $\exists a\in P$，满足 $\mathrm{Ind}(P)=\mathrm{Ind}(P-\{a\})$，则称属性 a 在 P 中是可缺的或冗余的，否则称 a 是不可缺的或必要的。当 P 中所有属性都是必要的，则称 P 是独立的。若 P 和 Q 为论域 U 上两个等价关系簇，$Q\subset P$ 是独立的，$\mathrm{Ind}(Q)=\mathrm{Ind}(P)$，则 Q 是 P 的一个约简。这里，等价关系即不可分辨关系。因此，一方面基于这种离散数据的约简在大规模数据集中其使用受到一定限制，同时，这里以单个属性的必要度为基础定义独立与约简也不完善。在离群数据分析中有时不能正确反映其离群贡献。

如数据对象集如表 1 所示，对象 x_1、x_2、x_3 组成一个簇，x_5、x_6、x_7 组成一个簇，而 x_4 是一个离群点。现从中去掉任何一个属性，其聚类效果和离群集均无任何变化，但很明显是由属性 A_3 和 A_5 的取值过大直接导致对象 x_4 的离群。因此，本文针对离群约简，以属性重要度及一维投影数据集子空间中形成的离群划分与在 L^d 中的离群划分 $\hat{C}_d$ 的相似水平来共同定义离群贡献率。

表 1 一个说明属性重要度的实例

X	A_1	A_2	A_3	A_4	A_5	A_6
x1	1	2	1	2	1	1
x2	1	1	2	1	2	2
x3	1	1	1	2	1	1
x4	2	1	888	1	999	1
x5	11	10	12	10	11	12
x6	10	11	11	11	12	11
x7	10	12	11	12	10	11

定义 5(离群贡献率) 设 $\alpha_1(A_j)$ 和 $\alpha_2(A-A_j)$ 分别为在一维属性子空间 $\{A_j\}$ 和 $d-1$ 维子空间 $\{A-A_j\}$ 上发现的离群数据集，按(2)、(3)式计算所得，称

$$\alpha(A_j)=\max(\alpha_1(A_j), 1-\alpha_2(A-A_j)),\ j=1,2,\cdots,d \tag{4}$$

为属性 A_j 的离群贡献率。

按表 1 所示的例子，计算得：$\alpha(A_3)=\alpha(A_5)=1$，其他为 0，可见与实际情况吻合。在约简算法中，随着问题规模增大会越来越复杂，而遗传算法由于它本身具有全局优化和隐含并行性等优点，因此适合求解这类问题。文[7,8]讨论了以属性重要度为基础先计算相对核并以此作为特定模式展开遗传进化，而文[9]提出了一种基于协同进化的属性约简方法，可以有效回避单个属性重要度方法的不足。本文提出的基于遗传算法的离群约简算法 GORA(Outlying Reduction Algorithm based on GA)根据离群贡献率形成遗传算法所需的初

始种群，获得了较好的效率。

4 遗传 α-离群约简算法

遗传 α-离群约简算法(GORA)在遗传算法主框架下在不同属性子空间投影数据对象集上通过离群挖掘算法进行分类，目的是探索不低于给定的离群相似水平 α_0 ($0<\alpha_0\leqslant 1$)且属性个数最少的子空间。

4.1 个体(染色体)编码

考虑离群约简问题的实际特点，一个很自然的编码方案就是用长度为 d 的二进制串来表示一个个体，其中每一位对应于条件属性集 A 中一个属性，为 1 时表示包含对应属性，为 0 时则表示不包含对应属性。如当 $d=10$ 时，L^d 的一个可能约简是 $L^s=\{A_1, A_3, A_8, A_9\}$，则相应的编码为 1010000110。

4.2 适应度函数

个体适应度值是评价个体性能的惟一确定性指标，是在进化过程中对个体进行优胜劣汰选择操作的主要依据，因此适应度函数的形式直接决定着群体的进化行为。

由 α-离群约简的定义可知，一个优良的个体 y 应该体现以下两个方面：

1) $\alpha(y)\geqslant\alpha_0$；

2)所含属性个数尽量少。

为此，构造如下适应度函数：

$$F(x)=\alpha(y)(1-S(y)/d) \tag{4}$$

其中 $S(y)$ 为染色体 y 中包含 1 的个数。

适应度函数 $F(y)$ 由两部分组成：第一个因子 $\alpha(y)$ 即 $\alpha(L^s)$ 由(3)式给出，此处 L^s 即是由 y 中取值为 1 的基因位对应的属性子集张成的投影子空间，所以 α 的值代表由 y 产生的离群划分 $\hat{C}_s$ 和 $\hat{C}_d$ 的相似程度，其越大说明 y 的适应度越强。第二个因子 $(1-S(y)/d)$ 表示个体 y 中为 0 的基因位数所占的比率，也即是不包含在个体 y 中的属性数量所占的比率，显然，y 中所含属性个数越少，其值越大。由此可见，函数 $F(y)$ 满足离群约简的求解要求。

4.3 遗传算子设计

遗传算子包括选择、交叉与变异运算。选择运算建立在对染色体适应度评价基础上，它将优良个体进行复制，从而体现遗传算法优胜劣汰的进化准则。交叉算子是产生新个体的最主要方法，而变异算子模拟生物进化变异，也是产生新物种不可忽视的因素。GORA 算法采用了繁殖池(Breeding Pool)比例选择策略与最优保持策略相结合的选择技术，通过单点交叉和限制性变异实现子代群体生成。

1)选择运算

设 m 为种群规模，其中每个个体 y_i 的适应度值为 $F(y_i)$，则

$$M_i=\text{round}(m\cdot F(y_i)/\sum_{i=1}^{m}F(y_i)) \tag{5}$$

为个体 y_i 的繁殖数目，其中 round(t)表示与 t 相距最小的整数，$i=1,2,\cdots,m$。每个个体分别复制成 M_i 个个体，组成一个繁殖池，从中随机抽取成对个体，进行杂交操作后取代当前个体，从而形成下一代个体。

为了保证进化过程中子代比父代有更强的适应能力，GORA 算法在子代形成过程中采用了最优保持策略。设 $F_{\max}$ 代表所有父代最优个体 $y_{\max}$ 的适应度，设当前种群中个体 y_i 的适应度最高，若 $F(y_i)>F_{\max}$，则把 y_i 作为新的 $y_{\max}$，然后用最优个体 $y_{\max}$ 替换当前种群中适应度最低的个体 y_j，该个体不参与交叉变异操作。可见，这种策略可以保证获得最优解。

需要说明的是，由于在计算每个个体 y_i 的适应度 $F(y_i)$ 时将会执行一次离群挖掘算法，因此种群规模 m 不宜选得过大。

2)交叉与变异运算

交叉与变异运算的结果是产生新个体。在 GORA 算法中采用单点交叉方法，即从繁殖中随机选择成对个体，随机设置一交叉点，以一定概率 P_c 在该点处把两个个体的部分染色体进行交换，从而生成两个新的子代个体。

变异算子根据个体适应度函数 $F(y)$ 的因子 $\alpha(y)$ 的值进行限制性变异。如果 $\alpha(y)\geqslant\alpha_0$，即该个体已达到离群相似水平要求。若要成为最优解，则应减少属性个数，因此这时只进行基因 1 到 0 的变异；否则，当 $\alpha(y)<\alpha_0$ 时，说明该个体尚未达到离群相似水平要求，若要成为最优解，则应增加属性个数，因此这时只进行基因 0 到 1 的变异。显然，这种限制性变异技术降低了变异的盲目性，使变异结果逐步往最优解的方向接近。因此，尽管变异的概率很小，也有利于提高算法性能。

4.4 GORA 算法描述与分析

根据以上分析，下面给出遗传 α-离群约简算法(GORA)的一般框架。

算法名称：GORA；输入：数据集 X，属性集 A，参数 α_0；输出：α-离群约简 α-ord；

Step1. 获取离群集。XO ←OM(X,L^d,ps)，OM(.)为一种离群挖掘算法，ps 代表 OM 算法中所涉及的一些参数，每次调用 OM 算法时这些参数保持不变，以保证各个体产生的离群划分的一致性；

Step2. 产生初始种群。首先计算每个属性的离群贡献率 $\alpha(A_j)$，$j=1,2,\cdots,d$。按如下方式产生 q 个个体：$t_k=(b_1 b_2\cdots b_d)$，$k=1,2,\cdots,q$，其中 q 为 $\alpha(A_j)>0$ 的属性个数。每一个个体 t_k 只有一个基因位为 1，其他全是 0，而 $b_j=1$ 当且仅当 $\alpha(A_j)>0$。取 $m(2\leqslant m\leqslant q)$ 为种群规模，随机产生 m 个正整数 $N_i(1\geqslant N_i\geqslant q)$，令

$$y_i=\bigcup_{l=1}^{N_i}t_{k_l},\ 1\leqslant k_l\leqslant q, i=1,2,\cdots,m \tag{6}$$

则 $G=(y_1,y_2,\cdots,y_m)$ 为初始种群；

Step3. 选择。对每个个体 y_i 调用相应的离群挖掘算法，获得离群集 XO_i←OM(X,y_i,ps)。按(4)、(5)式计算每个个体 y_i 的适应度及繁殖数目，形成繁殖池，随机成对选择个体进行复制。为了输出所有已发现的 α-离群约简，在计算个体适应度时，若离群相似水平 $\alpha(y_i)\geqslant\alpha_0$，则记录个体 y_i 的相应信息(y_i，$\alpha(y_i)$，d-$S(y_i)$)入数组 α-ord。设该步中个体 $y_{\max}$ 的适应度最大为 $F_{\max}$；

Step4. 生成新种群。从繁殖池中随机成对取出个体，根据交叉概率 P_c 进行交叉操作，同时根据计算所得的新个体 y_i 的离群相似水平 $\alpha(y_i)$ 按变异概率进行限制性变异，计算最终新个体的适应度。设其中 y_j 的适应度最高，若 $F(y_i)>F_{\max}$，则把 y_j 作为新的 $y_{\max}$，并用 $y_{\max}$ 替换本次适应度最低的个体。同样，计算适应度时若 $\alpha(y_i)\geqslant\alpha_0$，则 α-ord←(y_i,$\alpha(y_i)$,d-$S(y_i)$)；

Step5. 判断是否连续 w 代的最优个体适应度不再提高。如果是，则输出数组 α-ord 及最优 α-离群约简($y_{\max}$，$\alpha(y_{\max})$，

d-$S(y_{max})$)并终止算法，否则转 Step3 重复遗传算子。

对于给定的离群相似水平 α_0，GORA 算法获得了一系列不低于 α_0 的 α-离群约简，并以列表形式输出，从而可以清晰地显示各属性域子空间对离群数据集的贡献。设离群挖掘算法 $OM(.)$的时间复杂度为 $O(T(n,d))$，除 Step1 外，其他调用 $OM(.)$算法均是在其一子空间进行。每次时间小于 $O(T(n,d))$，其中 Step1 调用 d 次，Step3、4 共调用 $2m$ 次。设遗传算子重复次数为 M，其他计算时间为 $o(T(n,d))$，因此 GORA 算法总的时间复杂度不超过 $O((2mM+d+1)T(n,d))$。

5 实验结果

利用实际的数据集对 GORA 算法进行了实验。实验环境是 Pentuim2. 66G，内存 256M，操作系统为 Windows xp，编程语言采用 VC7. 0。测试数据集 X 来自广东某市移动通信业务数据库，由其一个子集、18 个属性、10 万条记录组成。离群挖掘算法采用文[1]提出的基于分区(Partition-Based)的方法，这种方法通过对大量正常数据对象的剪枝而更加专注于离群数据发现，具有较好的离群发现效率。

首先，执行 GORA 第一步，调用离群挖掘算法，获取离群数据集 XO，共发现离群点 57 个。然后计算每一个属性的离群贡献率 $\alpha(A_j)$，取种群规模 $m=30$，交叉概率 $P_c=0.7$，变异概率 $P_m=0.05$，算法在连续 3 代($w=3$)适应度无变化时结束。分别取不同的离群相似水平 α_0 进行实验，计算结果见表 2。

表 2 不同相似水平时的 α-离群约简

$\alpha 0$	α-离群约简
1	{call_duration, call_count, gn_charge, gj_charge, newfun_fee, gprs_duration }
0.95	{call_duration,call_count,gn_charge,newfun_fee, gprs_duration }
0.90	{call_duration,call_count,gn_charge,gprs_duration}
0.85	{call_duration,call_count,gn_charge,gj_charge} or {call_duration,call_count,gn_charge,newfun_fee}
0.80	{call_duration, call_count, gn_charge }

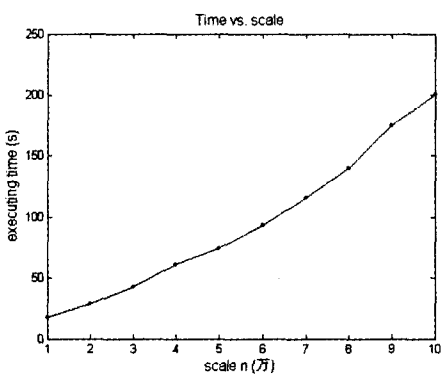

图 1 执行时间随问题规模的变化趋势

其中 call_duration 为呼叫时长、call_count 为呼叫次数、gn_charge 为国内漫游费、gj_charge 为国际漫游费、gprs_duration 为 GPRS 流量以及 newfun_fee 为新业务费。从表 2 看出，仅由属性子空间{call_duration, call_count, gn_charge}即可产生与 XO 具有相似水平达 0.8 的离群数据集。换句话说，这 3 个属性对 XO 离群贡献率达 80%，可见它们是致使数据离群的重要原因。而当 $\alpha_0=1$ 时，属性子空间 {call_duration, call_count, gn_charge, gj_charge, newfun_fee, gprs_duration }能产生与{XO, X－XO}的等价离群划分，它是 L^d 的一个离群约简。

取 $\alpha_0=0.95$，测试 GORA 算法对问题规模的适应性。图 1 显示了 $n=1\sim10$ 万条记录时算法执行时间与问题规模 n 的关系图，结果可看出其具有接近线性的增长率。

结论 离群数据分析对于诸多应用领域均具有重要意义。本文结合粗糙集理论的划分与属性约简技术，提出了离群约简与 α-离群约简等思想以及基于遗传算法的 α-离群约简搜索算法 GORA。通过定义属性贡献率等概念，构建了初始种群以及合理的适应度函数。同时，在选择和变异算子中分别引入了最优个体替代和限制性变异策略。实验表明，GORA 具有较好的搜索效率。本文的方法说明了如何通过较少的属性子空间去获得相同或相近的离群数据集，从而使对离群数据来源及出现原因的分析和理解更加集中于较小的目标域。

下一步我们将结合本文提出的 α-离群约简方法探索离群数据集内部分布规律及其在约简子空间中数据的离群特征，从而预测离群数据的离群趋势。

参 考 文 献

1 Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for Mining Outliers from Large Data Sets. In: Proc. of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2000. 427～438

2 Angiulli F, Pizzuti C. Outlier *mining* in large high dimensional data sets. IEEE Tran on Knowledge and Data Engineering, 2005, 17(2): 203～215

3 Tolvi J. Genetic algorithms for outlier detection and variable selection in linear regression models. Soft Computing, 2003

4 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法. 北京：科学出版社，2001

5 高隽. 智能信息处理方法导论. 北京：机械工业出版社，2004

6 Ware M, Wilson D, Ware A. A knowledge based genetic algorithm approach to automating cartographic generalization. Knowledge-Based Systems, 2003(16): 295～303

7 陶志，许宝栋，汪定伟，等. 基于遗传算法的粗糙集知识约简方法. 系统工程，2003,21(4):116～122

8 Dai J, Li Y. Heuristic genetic algorithm for minimal reduction decision system based on rough set theory. In: Proc. of the First International Conference on Machine Learning and Cybernetics. IEEE Press, 2002. 833～836

9 王立宏，吴耿锋. 基于并行协同进化的属性约简. 计算机学报，2003,26(5):630～635