第 26 卷第 8 期　　电 子 与 信 息 学 报　　Vol.26No.8
2004 年 8 月　　Journal of Electronics & Information Technology　　Aug. 2004

一种基于 GA 的混合属性特征大数据集聚类算法 [1]

李　洁　　高新波　　焦李成

(西安电子科技大学电子工程学院　西安　710071)

摘　要：　在数据挖掘中，经常会遇到和分析大量具有数值和类属特征的数据．然而，现有的大多数算法只能单独处理数值特征数据或类属特征数据，而不能分析具有混合属性的数据．为此，该文提出了一种基于 GA 的模糊聚类新算法，通过改进聚类目标函数将数值特征与类属特征相结合，从而实现具有混合属性特征数据的聚类分析；通过引入 GA 算法能够快速得到全局最优解，而且不依赖于原型初始化．实验结果表明，基于 GA 的新聚类算法对于处理具有混合特征的大数据集聚类问题是相当有效的．

关键词：　聚类分析，数值特征，类属特征，遗传算法

中图分类号：　TP391　　**文献标识码**：　A　　**文章编号**：　1009-5896(2004)08-1203-07

A GA-Based Clustering Algorithm for Large Data Sets with Mixed Numerical and Categorical Values

Li Jie　　Gao Xin-bo　　Jiao Li-cheng

(*School of Electronic Engineering, Xidian Univ., Xi'an* 710071, *China*)

Abstract　In the field of data mining, it is often encountered to perform cluster analysis on large data sets with mixed numerical and categorical values. However, most existing clustering algorithms are only efficient for the numerical data rather than the mixed data set. For this purpose, this paper presents a novel clustering algorithm for these mixed data sets by modifying the common cost function, trace of the within cluster dispersion matrix. The Genetic Algorithm (GA) is used to optimize the new cost function to obtain valid clustering result. Experimental result illustrates that the GA-based new clustering algorithm is feasible for the large data sets with mixed numerical and categorical values.

Key words　Cluster analysis, Numerical data, Categorical data, Genetic Algorithm(GA)

1 引言

在数据挖掘中把样本集划分成各种不同的类是一种基本操作 [1]，并在许多工作中获得了广泛的应用，比如分类 (无监督)[2] 、聚合和划分 [3] 或解剖 [2] 等．聚类 [4,5] 就是一种流行的近似划分方法．它把一组样本划分成若干类，使得根据某一特定的标准，在同一类内的样本之间彼此接近，而不同类的样本间差异较大．

然而，数据挖掘与其它传统聚类分析不同 [4,6]，它常常需要处理大量高维数据集 (具有几十或者几百个特征的数千甚至几百万个记录)．这就使得许多现有的聚类算法不能用在数据挖掘中．同时，在数据挖掘中经常遇到的数据通常既包含数值也包含类属值．传统的将类属值转化为数值的方法不是总能得到有效的结果，这是因为类属域是无序的．大多数现有的算法或者能分析这两种数据类型，但不能处理大数据集，或者能有效处理大数据集，但仅限于数值型数据．只有很少几种算法能很好地处理这些问题，例如 k-原型算法等 [7,8]．

为了处理具有混合特征的大数据集聚类问题，我们定义了一种新的目标函数，通过修正传统聚类算法的目标函数——类散布矩阵的迹，来达到将不同属性特征相结合的目的．与 k-均值

[1] 2003-03-27 收到，2003-07-08 改回
国家自然科学基金 (60202004, 60372045) 、863 计划 (2002AA135080) 资助课题

算法类似，这种新算法也对原型初始敏感，容易陷入局部极值点．为了解决这个问题，我们在聚类过程中引入遗传算法 (GA)．由于 GA 是基于全局搜索策略的，并且能够并行处理 [9]，因此，基于 GA 的聚类算法具有很高的效率，因而适合大数据集聚类分析．

本文的安排如下：第 2 节给出聚类新算法的目标函数的定义，第 3 节把遗传算法引入混合属性聚类算法中，第 4 节为实验结果，并将本文提出的新方法与 k-原型算法进行了性能比较．最后总结本文提出的算法，并指出进一步的研究方向．

2 目标函数的定义

令 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$ 表示一组具有 n 个样本的数据集，其中 $x_i=[x_{i1},x_{i2},\cdots,x_{im}]^{\mathrm{T}}$ 表示第 i 个样本的 m 个特征值．令 k 是一个正整数，那么对 $\boldsymbol{X}$ 进行聚类的目的就是要找到一个划分，将 $\boldsymbol{X}$ 中的目标分为 k 类．

对于给定的 n 个样本，样本集可能的划分数目是非常巨大的 [4]，为了找到最好的一个划分，而去逐个研究每一个划分是不切实际的．通常的解决方法是选择一个聚类准则 [4,6] 来指导搜索划分．下面，我们定义一个目标函数作为聚类准则．

2.1 数值数据聚类的目标函数

目前大家广泛使用的目标函数是类散布矩阵的迹 [6]，如式 (1) 所示：

$$C(\boldsymbol{W},\boldsymbol{P})=\sum_{i=1}^{k}\sum_{j=1}^{n}w_{ij}(d(x_j,p_i))^2,\quad w_{ij}\in\{0,1\}\tag{1}$$

式中 $p_i=[p_{i1},p_{i2},\cdots,p_{im}]^{\mathrm{T}}$ 表示第 i 类的原型，w_{ij} 是目标 x_j 属于第 i 类的隶属度 [6]．$\boldsymbol{W}$ 是 $k\times n$ 阶的划分矩阵，且满足概率约束 $\sum_{i=1}^{k}w_{ij}=1,\forall j$．$d(\cdot)$ 是定义为欧几里德距离的相异性测度．对于具有实特征的数据集，即 $\boldsymbol{X}\subset R^m$，则有

$$d^2(x_j,p_i)=(x_j-p_i)^{\mathrm{T}}\cdot(x_j-p_i)\tag{2}$$

因为 w_{ij} 是样本 x_j 属于第 i 类的隶属度，当 $w_{ij}\in\{0,1\}$ 时，我们称 $\boldsymbol{W}$ 是硬 k-划分．在硬划分中，$w_{ij}=1$ 表示样本 x_j 属于第 i 类．

2.2 混合数据聚类中的目标函数

当样本具有数值和类属混合特征时，假设每个样本用 $x_i=[x_{i1}^r,\cdots,x_{it}^r,x_{i,t+1}^c,\cdots,x_{im}^c]^{\mathrm{T}}$ 表示，混合类型目标 x_i 和 x_j 之间的相异性测度可由式 (3) 计算：

$$d^2(x_i,x_j)=\sum_{l=1}^{t}|x_{il}^r-x_{jl}^r|^2+\lambda\cdot\sum_{l=t+1}^{m}\delta(x_{il}^c,x_{jl}^c)\tag{3}$$

式中第 1 项是数值特征上的欧几里德距离平方，第 2 项是类属特征上的简单的相异匹配测度．$\delta(\cdot)$ 定义为

$$\delta(a,b)=\begin{cases}0, & a=b\\ 1, & a\neq b\end{cases}\tag{4}$$

权值 λ 用来调节两种特征在目标函数中的比例，以避免偏向任何一种特征．关于 λ 取值对分类结果的影响我们将另文讨论．

对于混合类型的目标，我们可以通过修正式 (1) 中的相异性测度如式 (3) 而得到新的目标函数．此外，我们还将硬 k-划分扩展为模糊划分，这样对于模糊聚类，目标函数进一步修正为

$$C(\boldsymbol{W},\boldsymbol{P})=\sum_{i=1}^{k}\left[\sum_{j=1}^{n}w_{ij}^2\sum_{l=1}^{t}|x_{jl}^r-p_{il}^r|^2+\lambda\sum_{j=1}^{n}w_{ij}^2\sum_{l=t+1}^{m}\delta(x_{jl}^c,p_{il}^c)\right],\qquad w_{ij}\in[0,1]\tag{5}$$

令

$$C_i^r = \sum_{j=1}^{n} w_{ij}^2 \sum_{l=1}^{t} |x_{jl}^r - p_{il}^r|^2 \tag{6}$$

$$C_i^c = \lambda \sum_{j=1}^{n} w_{ij}^2 \sum_{l=t+1}^{m} \delta(x_{jl}^c - p_{il}^c) \tag{7}$$

我们可将式 (5) 重写为

$$C(\boldsymbol{W}, \boldsymbol{P}) = \sum_{i=1}^{k} (C_i^r + C_i^c) \tag{8}$$

对具有数值和类属特征的数据集进行模糊聚类时，式 (8) 就是其目标函数．因为 C_i^r 和 C_i^c 都是非负的，所以可以通过分别最小化 C_i^r 和 C_i^c，来达到极小化 $C(\boldsymbol{W}, \boldsymbol{P})$ 的目的．值得注意的是，我们给 w_{ij} 加上幂指数 2，从而保证了硬划分向模糊划分的扩展是非平凡的．

3 基于 GA 的混合类型数据聚类算法

为了在具有数值和类属特征的大数据集中获得最优的模糊聚类，我们采用遗传算法来最小化目标函数．因为遗传算法是以随机的方式进行全局搜索，因此它以较大的概率搜索到全局最优解．而且，遗传算法能够并行处理，所以基于 GA 的聚类算法特别适合处理大数据集．

3.1 遗传算法

遗传算法是建立在生物进化基础之上的算法，是一种基于自然选择和群体遗传机理的搜索算法．它模拟了自然选择和自然遗传过程中的繁殖、交配和突变现象．它将每个可能的解看作是群体 (所有可能解) 中的一个个体，并将每个个体编码成字符串的形式，根据预定的目标函数对每个个体进行评价，给出一个适应度值．在遗传算法开始时，总是随机地产生一些个体，根据这些个体的适应度值，利用遗传算子对这些个体进行操作，得到一些新个体，这些新个体由于继承了上一代的一些优良性状，因此明显优于上一代，这样逐步朝着最优解的方向进化．

3.2 基于 GA 的聚类算法

为了利用 GA 求解混合属性数据集的模糊聚类问题，首先需要解决以下 3 个问题 [10]：(1) 如何将聚类问题的解编码到基因串中；(2) 如何构造适应度函数来度量每条基因串对聚类问题的适应程度；(3) 如何选择各个遗传算子，确定各操作参数的取值，以确保快速收敛到最优解．

3.2.1 编码方案　我们由式 (1) 和式 (5) 定义的聚类目标函数可知，聚类的目的就是要获得数据集 $\boldsymbol{X}$ 的一个模糊划分矩阵 $\boldsymbol{W}$ 和聚类的原型矩阵 $\boldsymbol{P}$．而 $\boldsymbol{W}$ 和 $\boldsymbol{P}$ 是相关的，即已知其一则可求得另一个的解，因此，我们就有两种编码方案，对 $\boldsymbol{W}$ 矩阵编码或对 $\boldsymbol{P}$ 矩阵编码．因为我们的算法是处理大数据集的，$\boldsymbol{W}$ 矩阵必然具有极大的搜索空间，从而降低搜索效率，所以，我们选择第 2 种编码方案．

我们把原型中的 k 组特征连接起来，根据各自的取值范围，将其量化值 (用二进制串表示) 编码成基因串．

$$g = \left\{ \underbrace{\zeta_1, \zeta_2, \cdots, \zeta_m}_{\text{Encode}(p_1)}, \cdots, \zeta_i, \cdots, \underbrace{\zeta_{(k-1)m+1}, \zeta_{(k-1)m+2}, \cdots, \xi_{km}}_{\text{Encode}(p_k)} \right\} \tag{9}$$

式中参数集依据每个原型 p_i 取值．需要注意的是，由于我们处理的是混合特征，所以在基因串中，除了具有数值参数以外，还应具有类属参数．又因为，类属特征是无序的，因此，它们可以直接编码到基因串中，而不需要量化．

3.2.2 适应度函数构造 由聚类目标函数定义可知，目标函数越小，则聚类效果越好，而此时适应度应该越大．因此我们借助目标函数来构造适应度函数如下式：

$$f(g)=\frac{1}{1+C(W,P)}=\frac{1}{1+\sum_{i=1}^{k}\sum_{j=1}^{n}w_{ij}^{2}(d(x_j,p_i))^2} \tag{10}$$

3.2.3 遗传算子选择 在基于 GA 的聚类算法中，我们选用所有的算子 (选择、复制、交叉和变异)，只不过赋予其不同的使用概率，使其按概率操作．

在每一代的 N 个个体中，我们先将其按适应度值由高到低排序，并把其序号赋给每个个体．则选择概率定义为

$$P_s(g_i)=2(N-i+1)/[N(N+1)] \tag{11}$$

交叉率和变异率依据下式自适应选取：

$$P_c(g_i,g_j)=\begin{cases}\alpha_1(f_{\max}-f')/(f_{\max}-\overline{f}), & f'\geq\overline{f}\\ \alpha_2, & \text{其他}\end{cases} \tag{12}$$

$$P_m(g_i)=\begin{cases}\alpha_3(f_{\max}-f(g_i))/(f_{\max}-\overline{f}), & f(g_i)\geq\overline{f}\\ \alpha_4, & \text{其他}\end{cases} \tag{13}$$

式中 $f_{\max}=\max_{l=1}^{N}\{f(g_l)\}$，$\overline{f}=\frac{1}{N}\sum_{l=1}^{N}f(g_l)$，$f'=\max\{f(g_i),f(g_j)\}$，并且 $\alpha_i\in[0,1]$.

除了上述算子以外，我们在该聚类算法中又定义了一个新的算子——一步迭代算子．对于每一个个体，将基因串解码到聚类原型 $\boldsymbol{P}$ 之后，再进行一步迭代算子，一步迭代算子包含以下两个步骤：

$$w_{ij}=\left[\sum_{l=1}^{k}(d(x_j,p_l))^{-2}\right]^{-1}\Big/(d(x_j,p_i))^2,\qquad \forall i,j \tag{14}$$

$$p_{il}=\begin{cases}p_{il}^{r}=\sum_{j=1}^{n}w_{ij}^{2}x_{jl}^{r}\Big/\sum_{j=1}^{n}w_{ij}^{2}, & l=1,2,\cdots,t,\\ p_{il}^{c}=c_l^{\max}, & l=t+1,\cdots,m,\end{cases}\qquad \forall i \tag{15}$$

式中 $c_l^{\max}$ 表示属于第 i 类的样本中在第 l 维特征上占优势的类属特征值．在根据式 (15) 获得了新的聚类原型后，再将其编码到基因串中，并重新进行上述遗传算子的操作，直到聚类原型收敛到最优解．

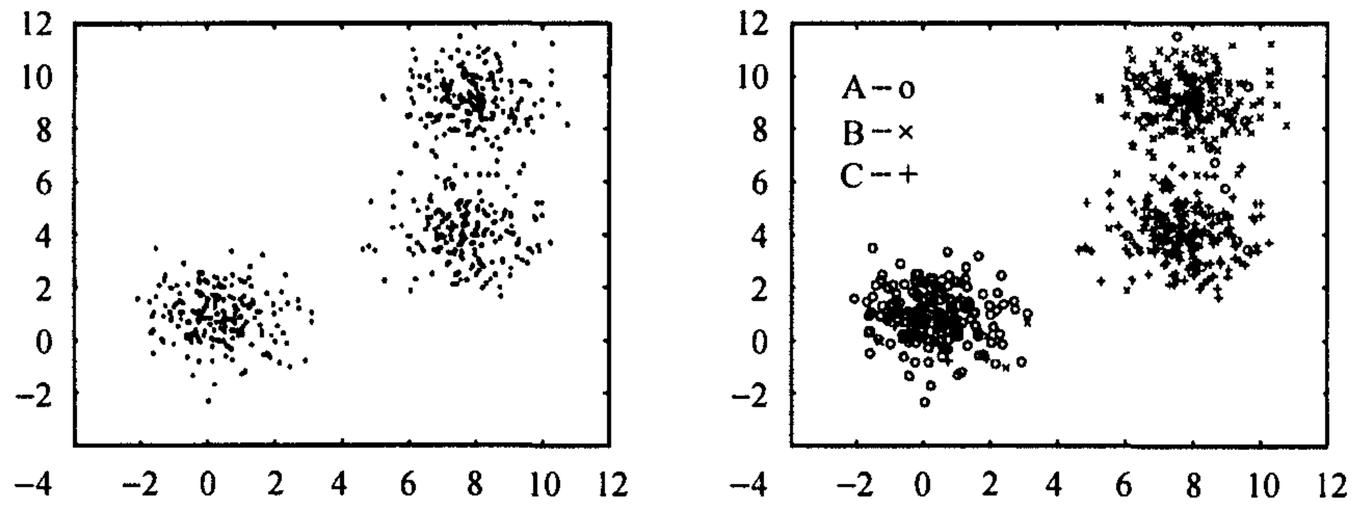
4 实验结果

为了测试基于 GA 的算法的有效性，我们给出一些初步的实验结果，并比较了增加一步迭代算子前后聚类算法的收敛性，将新算法和 k-原型算法的收敛速度和分类性能进行了比较，显示出新算法的优良性能．

4.1 数据集的构造

为了便于直观显示，我们构造的数据样本仅具有 3 个特征，两个数值型的和一个类属型的．首先产生一组具有 3 个正态分布的二维的数据点，共包含 600 个样本，如图 1(a) 所示．然后通过给每一个点叠加一个类属特征而扩展到三维 (如图 1(b))．对于类属特征的赋值是这样的：在每一部分中分配给大多数点相同的类属值，剩下的分配给其它的类属值．例如在图 1(b) 右上部分中大多数点的类属特征都是 B，在这一部分中剩下的指定为 A 或 C，并且所有的分派都是随机的．

注意每个点的类属值并不表示它的类别信息．实际上，这些点根本没有分类．类属值只是简单地在第三维表示目标，第三维是不连续且是无序的．我们也可以把这维看作是一个位面集，在位面上任何两点间的距离是 1 ．每个位面是由唯一的属性值确定的．所有的点依据其属性值投影到相应的位面上．



(a) 三组二维正态分布的点　　(b) 叠加了类属特征的点

图 1　具有数值和类属特征的人造检测数据集

4.2 聚类分析结果及一步迭代算子的影响

图 2 显示了用本文提出的算法对上述数据集进行聚类的结果 ($\lambda = 1$) ．如果一个样本具有类属特征 A ，但它接近的大多数点具有类属特征 B ，而且它距离大多数具有类属特征 A 的点很远，那么它就分配给大多数具有类属特征 B 的类．在这种情况下，两个数值特征决定了样本的类别，而不是类属特征．然而，如果一个点具有类属特征 A ，而且被具有类属特征 B 的点包围，但距离大多数具有类属特征 A 的点不太远，那么就把它分为具有类属特征 A 的类．在这种情况下，它的类别标记是由类属特征决定的，而不是它的空间位置．所以，在确定点的类别时，数值特征和类属特征具有同等的价值．

基于 GA 的聚类算法的在线和离线特性显示在图 3 中．在线特性和离线特性的定义如下：

$$P_{\text{online}}(t) = \frac{1}{t}\sum_{i=1}^{t}\left[\frac{1}{N}\sum_{k=1}^{N} f^{(i)}(g_k)\right] \tag{16}$$

$$P_{\text{offline}}(t) = \frac{1}{t}\sum_{i=1}^{t}\left[\max_{k=1}^{N} f^{(i)}(g_k)\right] \tag{17}$$

式中 t 表示进化代数．图 3 中所示的点划线和实线分别为我们提出的增加了一步迭代算子的新遗传算法的离线和在线特性，虚线是传统遗传算法的在线特性．该图清楚地显示出：在引入一步迭代算子的操作之后，我们新算法的收敛速度明显提高，能够很快地收敛到全局最优解。

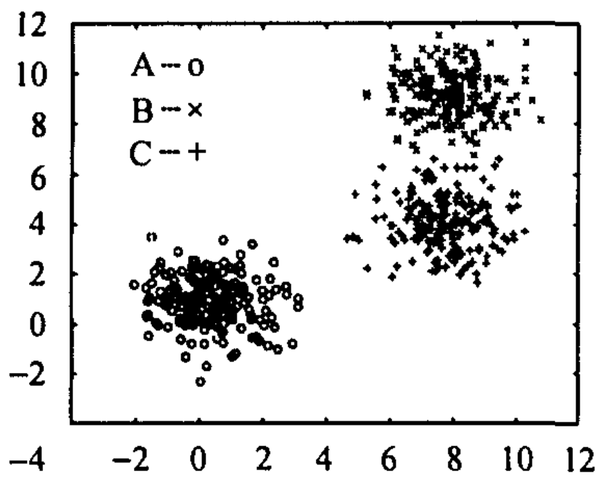


图 2　利用本文算法的聚类结果

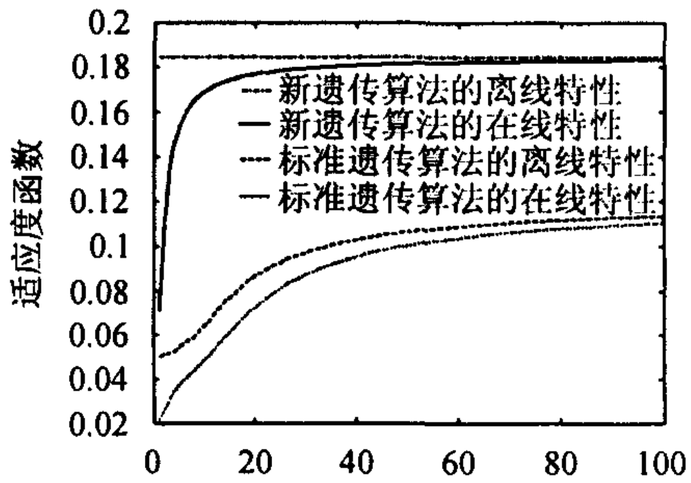


图 3　本文算法的收敛曲线

4.3 大数据集聚类分类结果

为了检验大数据集聚类分析的性能，本节给出了新算法与 k-原型算法的比较实验．该实验所利用的数据集包含 10000 个样本，每个样本具有 9 个数值特征和 11 个类属特征．

从图 4 我们看出，本文算法的收敛速度明显高于 k-原型算法，而且在每一次迭代过程中，基于 GA 的聚类算法也比 k-原型算法的目标函数值要小得多．说明其聚类效果明显优于 k-原型算法．图 5 显示了本文算法的在线和离线特性．

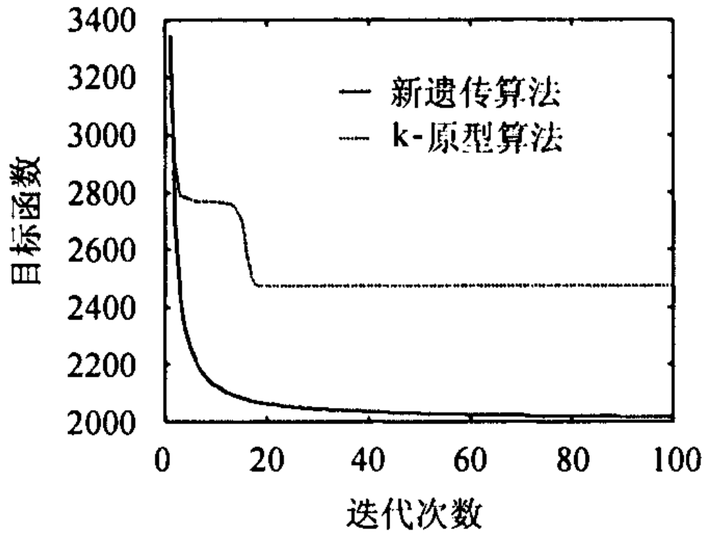


图 4 本文算法与 k-原型算法收敛曲线

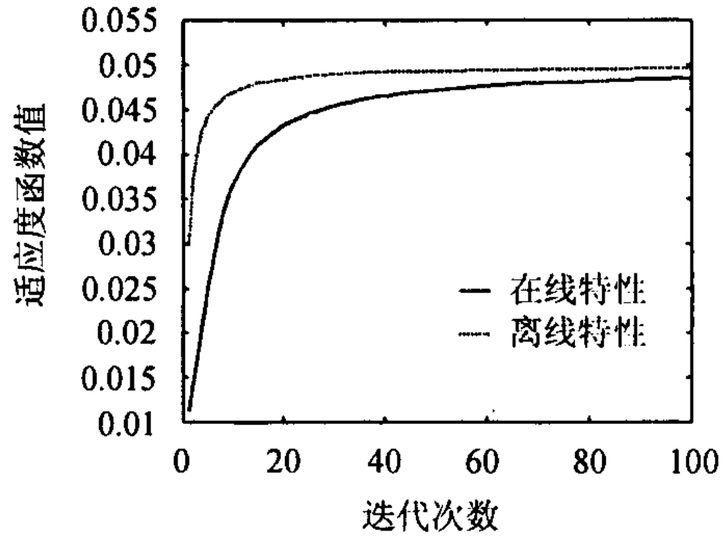


图 5 本文算法的在线和离线特性

为了验证本文算法的分类性能，我们在同一数据集上，随机运行了 10 次，将 GA 算法与 k-原型算法得到的分类正确率列在表 1 中．

表 1 GA 算法与 k-原型算法分类正确率

正确率	第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次	第 7 次	第 8 次	第 9 次	第 10 次
新算法	99.2%	99%	99.2%	99.2%	99.2%	99.2%	99.2%	99.2%	99%	99%
k-原型算法	40%	60.4%	79.6%	99.2%	59.7%	62.2%	99.2%	60.2%	99%	99.2%

从表中我们看出，本文提出的基于 GA 的聚类算法每次分类正确率都在 99% 以上，说明该算法得到的确是全局最优解，而 k-原型算法有时候正确率高，有时却很低，这表明 k-原型算法的有效性取决于初始原型的选择．因此，实验结果表明：本文的方法在收敛速度和分类性能两个方面都是十分有效的．

5 结论

本文提出将遗传算法用于大数据集聚类分析．并在大数据集上对聚类性能进行了评估，实验结果表明该方法能够有效地发现数据中隐含的结构．当对大量具有数值和类属混合特征的数据进行聚类分析时，我们发现基于 GA 的算法收敛速度快，且不依赖于初始原型的选择，能以极大的概率收敛到全局最优解．而这些特性在数据挖掘中是十分重要的．

本文的重点放在如何利用遗传算法解决具有混合属性特征数据的聚类问题．然而，在运用该方法解决实际的数据挖掘问题时，我们需要面对这样一个难题：数据集应该分为几类？即聚类有效性问题将是我们进一步研究的重要方向．

参 考 文 献

[1] Klosgen W, Zytkow J M. Knowledge Discovery in Databases Terminology. Advances in Knowledge Discovery and Data Mining, Fayyad U M, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. (Eds.), AAAI Press/ The MIT Press, MA, 1996: 573–592.

[2] Cormack R M. A review of classification. *J. Roy. Statist. Soc. Series A*, 1971, 134: 321–367.

[3] IBM. Data Management Solutions. IBM White Paper, IBM Corp. 1996.

[4] Anderberg M B. Cluster Analysis for Applications. New York: Academic Press. 1973: 79–90.

[5] Kaufman L, Rousseeuw P J. Finding Groups in Data—An Introduction to Cluster Analysis. New York: John Wiley, 1990: 98–110.

[6] Everitt B. Cluster Analysis. New York: Heinemann Educational Books Ltd., 1974: 45–60.

[7] Huang Zhexue, Michael K N. A fuzzy k-modes algorithm for clustering categorical data. *IEEE Trans. on Fuzzy Systems*, 1999, 7(4): 446–452.

[8] Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Dept. of Computer Science, The University of British Columbia, Canada, 1997: 1–8.

[9] Holland J H. Adoption in Natural and Artificial System. Ann Arbor, MI: Univ. Mich. Press, 1975: 83–90.

[10] Krovi R. Genetic algorithm for clustering: A preliminary investigation. Proceedings of the 25th Hawaii International Conf. on System Sciences, 4, Information Systems, Hawaii, 1992: 504–544.

李　洁：　女，1972 年生，讲师，硕士，主要研究方向为：人工智能、模式识别.

高新波：　男，1972 年生，教授，博士生导师，主要研究方向为：模式识别、图像处理、人工智能等.

焦李成：　男，1959 年生，教授，博士生导师，主要研究方向为：模式识别、图像处理、人工智能等.