一种基于 CSA 的混和属性特征大数据集聚类算法

李 洁,高新波,焦李成
(西安电子科技大学 电子工程学院,陕西西安 710071)

摘 要: 在数据挖掘中,我们经常会遇到和分析大量具有数值和类属特征的数据.然而,现有的大多数分类算法只能单独处理数值特征数据或类属特征数据,而不能分析具有两种混合属性的数据.为此,本文提出一种基于克隆选择的模糊聚类新算法,通过改进距离测度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;通过引入克隆选择算法(CSA)实现目标函数的全局优化.由于克隆算子能够将进化搜索与随机搜索、全局搜索和局部搜索相结合,因而通过对候选解进行克隆算子操作,能够快速得到全局最优解.实验结果表明,基于 CSA 的模糊聚类新算法对于处理具有混和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; 克隆选择算法
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2004) 03-0357-06

A CSA-Based Clustering Algorithm for Large Data Sets with Mixed Numeric and Categorical Values

LI Jie, GAO Xin-bo, JIAO Li-cheng
(School of Electronic Engineering, Xidian Univ., Xi'an, Shaanxi 710071, China)

Abstract: In the field of data mining, it is often encountered to perform cluster analysis on large data sets with mixed numeric and categorical values. However, most existing clustering algorithms are only efficient for the numeric 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 clonal selection algorithm (CSA) is used to optimize the new cost function, since the clonal operator can combine the evolutionary search and random search, and incorporate the global search with local search, by the clonal operation on candidate solutions; the new algorithm can quickly obtain the global optimum. Experimental result illustrates that the CSA-based new clustering algorithm is feasible for the large data sets with mixed numeric and categorical values.

Key words: cluster analysis; numeric data; categorical data; clonal selection algorithm

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)、聚合、划分或解剖[2]等.聚类[3]就是一种现在相当流行的近似划分方法.它把一组样本划分成若干类,使得在某一特定的标准下,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[3],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,在数据挖掘中遇到的数据通常既包含数值特征也包含类属特征.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较好的处理这些问题,例如 k-原型算法等[4,5].

为了处理具有混和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-原型算法类似,这种新算法也对原型初始化敏感,容易陷入局部极值点,因此人们提出了基于遗传算法(GA, Genetic Algorithm)的聚类方法,尽管该方法能以较高的概率收敛到全局最优点,但收敛速度较慢,而且还容易出现早熟现象[6].

为了解决全局优化问题,我们在聚类过程中引入克隆选择算法(CSA, Clonal Selection Algorithm). CSA 是一种新兴的人工免疫系统方法[7,8],它借助生物学免疫系统的抗体克隆选择机理,构造适用于人工智能的克隆算子.由于基于克隆算子

收稿日期:2003-04-14;修回日期:2003-10-24
基金项目:国家自然科学基金(No. 60073053, 60202004);863 计划基金

的克隆选择算法是群体搜索策略,本质上固有并行性和搜索变化的随机性,在搜索中不易陷入局部极值,最终能以较大的概率获得问题的全局最优解,且具有较快的收敛速度. 因此,与基于 GA 的聚类算法相比,基于 CSA 的模糊聚类新算法将会具有更高的效率,因而适合于大数据集的聚类分析.

下文的安排如下:下一节给出聚类新算法的目标函数的定义,第三节把克隆选择算法引入混和属性聚类算法中,第四节为实验结果,将本文提出的新方法与 k-原型算法及基于 GA 的聚类算法进行了性能比较,并讨论了参数 λ 取值对分类结果的影响. 最后总结全文,并指出进一步的研究方向.

2 目标函数的定义

令 $X=\{x_1, x_2, \cdots, x_n\}$ 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 X 进行聚类的目的就是要找到一个最优划分,将 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[3],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. 因此,通常的解决方法是选择一个聚类准则[3]来指导搜索划分. 下面,我们定义一个目标函数作为本文的聚类准则.

2.2 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.

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

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]. 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 $X\subset R^m$,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,称 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和特征时,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,混合类型样本 x_i 和 x_j 之间的相异性测度可由式(3)计算:

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

式中第一项是数值特征上的欧几里德距离平方,第二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.

对于混合类型的目标,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,这样对于模糊聚类问题,目标函数进一步修正为:

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

令

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

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

可将式(5)重写为:

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

对具有数值和类属混合特征的数据集进行模糊聚类分析时,式(8)就是其目标函数. 因为 C_i^r 和 C_i^c 都是非负的,所以可以通过分别极小化 C_i^r 和 C_i^c,来达到极小化 $C(W,P)$ 的目的. 需要指出的是,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 CSA 的混合特征数据聚类算法

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分析结果,我们采用克隆选择算法来最小化目标函数. 因为克隆算子将进化搜索与随机搜索、全局搜索和局部搜索相结合,因此它以概率 1 搜索到全局最优解[8]. 而且,克隆选择算法能够并行处理,所以基于 CSA 的聚类算法特别适合大数据集的分析处理.

3.1 克隆选择算法(CSA)

1958 年 Burnet 等提出了著名的克隆选择学说[9],其中心思想是:抗体是天然产物,以受体的形式存在于细胞表面,抗原可与之选择性的反应. 抗原与相应抗体受体的反应可导致细胞克隆性增殖,该群体具有相同的抗体特异性,其中某些细胞克隆分化为抗体生成细胞,另一些形成免疫记忆细胞,参加以后的二次免疫反应. 克隆性在细胞水平上表现出 TCR 和 BCR(T and B Cell Antigen Receptor)结构的极端多样性,因此,直接导致了抗体网络的多样性、记忆性和特异性. 克隆选择算法正是基于抗体克隆选择这一生物特性而形成的一种新的人工免疫系统方法.

与进化计算一样,人工免疫系统方法也能解决函数优化问题. 假设所需优化的函数为 ϕ: $\prod_{i=1}^{t}[d_i,u_i]\to R(d_i<u_i)$,其中 t 是优化变量的个数,变量 $x_i\in[d_i,u_i]$,则抗原就是被优化的函数 ϕ,抗体群 $\bar{A}=\{A_1, A_2, \cdots, A_N\}$ 为抗体 A 的 N 元组,抗体 A_i 是解空间 S^t 中的一个点.

抗体-抗原亲合度函数 f 一般是 ϕ 的函数,抗体-抗体亲合力定义为:

$$W_{ij}=\|A_i-A_j\|,\qquad i,j=1,2,\cdots,N \tag{9}$$

$\|\cdot\|$ 为任意范数,$W=(W_{ij})_{N\times N}$,$i,j=1,2,\cdots,N$ 为抗体-抗体亲合力矩阵,它反应了种群的多样性.

克隆选择算法可简单的描述为:

step1 $l=0$,初始化抗体群落 $\bar{A}=\{A_1(0), A_2(0), \cdots, A_N(0)\}$,设定算法参数,计算初始种群的亲合度;
step2 依据亲和度和设定的抗体克隆规模,进行克隆算子操作[8],获得新的抗体群 $\bar{A}(l+1)$;
step3 $l=l+1$,若满足终止条件,停止计算;否则,返回 step2.

假设 $M=\{\bar{A}\mid \max(f(\bar{A}))=f^*, \forall \bar{A}\in S^N\}$ 为满意种群,即满意种群集 M 中的任意抗体群 $\bar{A}$ 中至少包含一个最优解 f^*. 可以证明克隆选择算法生成的抗体种群序列 $\{\bar{A}(l), l\geq 0\}$ 是有限非齐次可约马尔可夫链,它以概率 1 收敛到满意种群 M[8].

3.2　基于 CSA 的模糊聚类算法

为了利用克隆选择算法求解数据集的模糊聚类问题,首先需要解决以下三个问题:(1)如何将聚类问题的解编码到抗体中;(2)如何构造抗体-抗原亲合度函数来度量每个抗体对聚类问题的响应程度;(3)如何选择各个克隆算子,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

由式(8)定义的聚类目标函数可知,聚类的目标就是要获得数据集 X 的一个模糊划分矩阵 W 和聚类的原型 P. 而 W 和 P 是相关的,即已知其一则可求得另一个的解,所以,在基于 CSA 的模糊聚类算法中,可令一组聚类原型 P 就是一个抗体,这样把原型中的组特征连接起来,根据各自的取值范围,就可以将其量化值 k(用二进制串表示)编码成一个抗体.

$$A_i(0)=\{\underbrace{\zeta_1,\zeta_2,\cdots,\zeta_m}_{\text{Encode}(p_1)},\cdots,\zeta_j,\cdots,\underbrace{\zeta_{(k-1)m+1},\zeta_{(k-1)m+2},\cdots,\zeta_{km}}_{\text{Encode}(p_k)}\},\quad i=1,2,\cdots,N \tag{10}$$

式中参数集依据每个原型 $p_i(1\leq i\leq k)$ 取值.

3.2.2　抗体-抗原亲合度函数构造

由聚类目标函数定义可知,目标函数越小,则聚类效果越好,而此时抗体-抗原亲合度应该越大. 因此我们借助目标函数来构造抗体-抗原亲合度函数如下式.

$$f(A)=\frac{1}{1+C(W,P)}=\frac{1}{1+\sum_{i=1}^{k}\sum_{l=1}^{n}w_{il}^2(d(x_l,p_i))^2} \tag{11}$$

3.2.3　克隆算子选择

在基于 CSA 的模糊聚类算法中,克隆算子包括克隆操作、免疫基因操作(克隆重组和克隆变异)以及克隆选择和克隆死亡四个部分.

(1)克隆操作 T_c^C

定义克隆操作为:

$$T_c^C(\bar{A}(l))=[T_c^C(A_1(l))\ T_c^C(A_2(l))\ \cdots T_c^C(A_N(l))] \tag{12}$$

其中,$T_c^C(A_i)=A_i\otimes I_i$, $i=1,2,\cdots,N$; I_i 为 q_i 维全 1 行向量,称抗体 A_i 的 q_i 克隆. q_i 的大小取决于 $f(A_i)$ 和 N_c, N_c 是与克隆规模有关的设定值,有 $N_c>N$. 一般情况下:

$$q_i=\mathrm{Int}\left(N_c\times\frac{f(A_i)}{\sum_{j=1}^{N}f(A_j)}\right),\quad i=1,2,\cdots,N \tag{13}$$

$\mathrm{Int}(x)$ 表示大于 x 的最小整数. 经克隆操作后,抗体种群变为:

$$\bar{A}'(l)=[\bar{A}(l)\ \bar{A}'_1(l)\ \bar{A}'_2(l)\cdots\bar{A}'_N(l)] \tag{14}$$

其中

$$A'_i(l)=[A_{i1}(l)\ A_{i2}(l)\cdots A_{iq_i-1}(l)], A_{ij}=A_i;\ j=1,2,\cdots,q_i-1 \tag{15}$$

(2)免疫基因操作 T_g^C

①克隆变异

依据概率 p_m^i 对克隆后的抗体群进行变异操作,为了保留抗体原始种群的信息,变异算子并不作用到 $\bar{A}(l)\in\bar{A}'(l)$,即:

$$p_m(A_{ij}(l)\to A'_{ij}(l))=\begin{cases}p_m^i, A_{ij}(l)\notin\bar{A}(l), & i=1,2,\cdots,N\\ 0, A_{ij}(l)\in\bar{A}(l), & j=1,2,\cdots,q_i-1\end{cases} \tag{16}$$

②克隆重组

克隆变异之后,对所有新的抗体群依据概率 p_c 进行克隆重组操作,同样为了保留抗体原始种群的信息,克隆重组算子并不作用到 $\bar{A}(l)\in\bar{A}'(l)$,即:

$$p_c((A_{ij}(l),A_{st}(l))\to(A'_{ij}(l),A'_{st}(l)))=\begin{cases}p_c, A_{ij}(l),A_{st}(l)\notin\bar{A}(l)\\ 0, A_{ij}(l),A_{st}(l)\in\bar{A}(l)\end{cases} \tag{17}$$

其中 $i,s=1,2,\cdots,N$; $j=1,2,\cdots,q_i-1$; $t=1,2,\cdots,q_s-1$.

(3)克隆选择操作 T_s^C

对 $\forall i=1,2,\cdots,N$,存在新抗体 $B=\{A'_{ij}(l)\mid \max f(A'_{ij}(l))\}$, $j=1,2,\cdots,q_i-1$,则 B 取代 $A_i(l)\in\bar{A}(l)$ 的概率为:

$$p_s(A_i\to B)=\begin{cases}1, & f(A_i)<f(B)\\ \exp\left(-\frac{f(A_i)-f(B)}{\beta}\right), & f(A_i)\geq f(B)\text{ 且 }A_i\text{ 不是目前种群的最优解}\\ 0, & f(A_i)\geq f(B)\text{ 且 }A_i\text{ 是目前种群的最优解}\end{cases} \tag{18}$$

式中,$\beta>0$ 是一个与抗体种群多样性有关的参数,一般 β 取值越大,多样性越好.

(4)克隆死亡操作 T_d^C

经克隆选择后,新的抗体群为:

$$\bar{A}(l+1)=[A_1(l+1)\ A_2(l+1)\cdots A_N(l+1)] \tag{19}$$

若在 $\bar{A}(l+1)$ 中存在 $A_i(l+1)$ 和 $A_j(l+1)$,满足 $f(A_i(l+1))=f(A_j(l+1))=\max f(\bar{A}(l+1))$, $i\neq j$,则可随机产生一个新的抗体,以概率 p_d 任意死亡 $A_i(l+1)$ 和 $A_j(l+1)$ 中的一个.

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. 对于每个抗体,将其解码到聚类原型 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:

$$w_{ij}=\left(\sum_{l=1}^{k}(d(x_j,p_l))^{-2}\right)^{-1}\Big/(d(x_j,p_i))^2,\ \forall i,j \tag{20}$$

$$p_{il}=\begin{cases} p_{il}^{r}=\sum_{j=1}^{n} w_{ij}^{2}x_{jl}^{r}\Big/\sum_{j=1}^{n} w_{ij}^{2}, & l=1,2,\cdots,t \\ p_{il}^{c}=c_{l}^{\max}, & l=t+1,\cdots,m \end{cases}, \quad \forall i \tag{21}$$

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(21)获得了新的聚类原型后,再将其编码到抗体中,并重新进行上述克隆算子的操作,直到聚类原型收敛到最优解.

本文提出的基于 CSA 的模糊聚类新算法的流程可用下图表示.

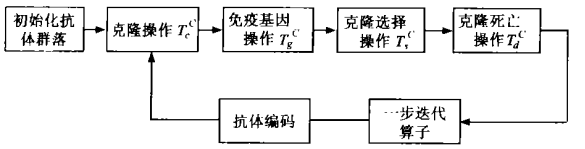


图 1 基于 CSA 的聚类新算法的流程框图

4 实验结果

为了测试基于 CSA 的模糊聚类算法的有效性,我们给出一些初步的实验结果,将新算法和 k-原型算法以及遗传算法的收敛速度和分类性能分别进行了比较,并比较了增加一步迭代算子前后聚类算法的收敛性,以及 λ 取值对混合属性特征数据集分类结果的影响.

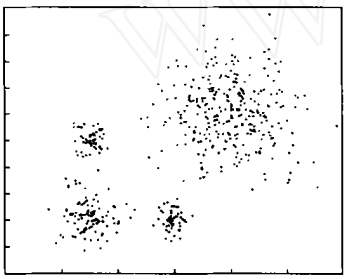
4.1 数据集的构造

为了便于直观显示,我们构造的数据样本仅具有三个特征,两个数值型的和一个类属型的. 首先产生四组不同方差的正态分布的二维数据点,共包含 500 个样本,如图 2(a)所示. 然后通过给每一个点迭加一个类属特征而扩展到三维(如图 2(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 2(b)左下部分中大多数点的类属特征都是 C,在这一部分中剩下的指定为 A、B 或 D,并且所有的分派都是随机的.

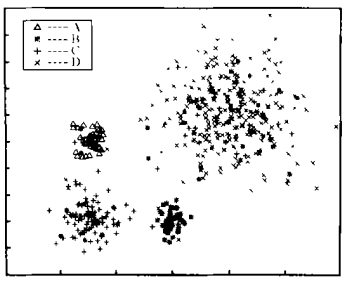
注意每个点的类属值并不表示它的类别信息. 实际上,这些点根本没有分类. 类属值只是简单的表示第 i 个样本在第三维上是不连续且是无序的. 也可以把这维看作是一个位面集,在位面上任何两点间的距离是 1. 每个位面是由唯一的属性值确定的. 所有的点依据其属性值投影到相应的位面上.

4.2 分类性能测试

图 3 显示了用本文提出的算法、k-原型算法以及遗传算法分别对上述数据集进行聚类的结果($\lambda=0.8$). 图 3(a)是 k-原型算法的分类结果,我们可以看到,由于样本集里,几类数据的分布方差不同,k-原型算法陷入局部极值点;图 3(b)和(c)都是用遗传算法分类的结果,由于其初始状态不同,(b)找到的是全局最优解,而(c)却产生早熟,图 3(d)是采用本文算法的聚类结果,图中显示基于 CSA 的聚类算法得到的是该数据集的最优划分.

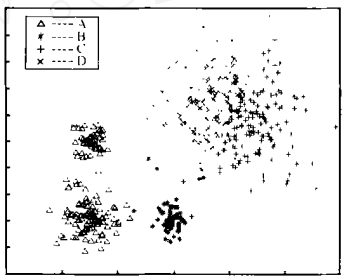


(a) 四组二维正态分布的点



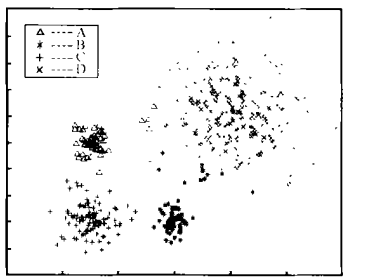
(b) 叠加了类属特征的点

图 2 具有数值和类属特征的人造检测数据集



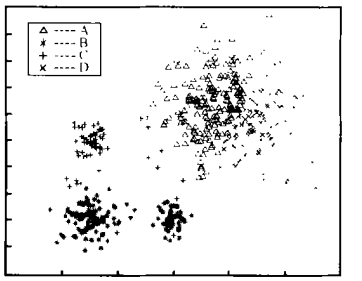
λ=0.8

(a) k-原型算法分类结果



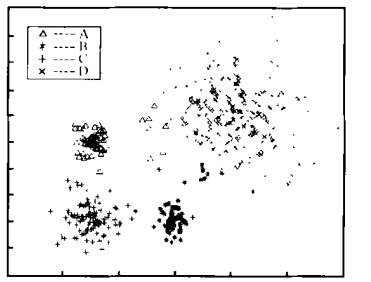
λ=0.8

(b) 遗传算法分类结果



λ=0.8

(c) 遗传算法分类结果



λ=0.8

(d) 克隆算法分类结果

图 3 测试结果(一)

4.3 收敛性能及一步迭代算子影响测试

为了检验新算法的收敛性以及一步迭代算子对收敛速度的影响,我们仍然采用图 2 所示的混和特征数据集进行测试实验. 得到的离线和在线特性显示在图 4 中,在线特性和离线特性的定义如下:

$$P_{\text{online}}=\frac{1}{l}\sum_{i=1}^{l}\left[\frac{1}{N}\sum_{l=1}^{N} f^{(i)}(A_l)\right] \tag{22}$$

$$P_{\text{offline}}=\frac{1}{l}\sum_{i=1}^{l}\left[\max_{l=1}^{N} f^{(i)}(A_l)\right] \tag{23}$$

式中 l 表示进化代数.

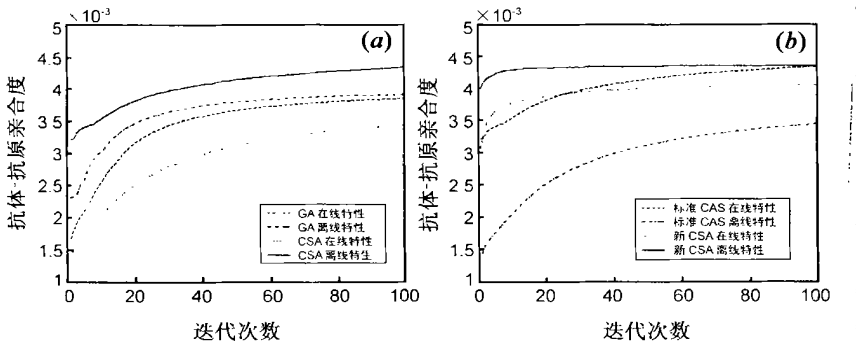


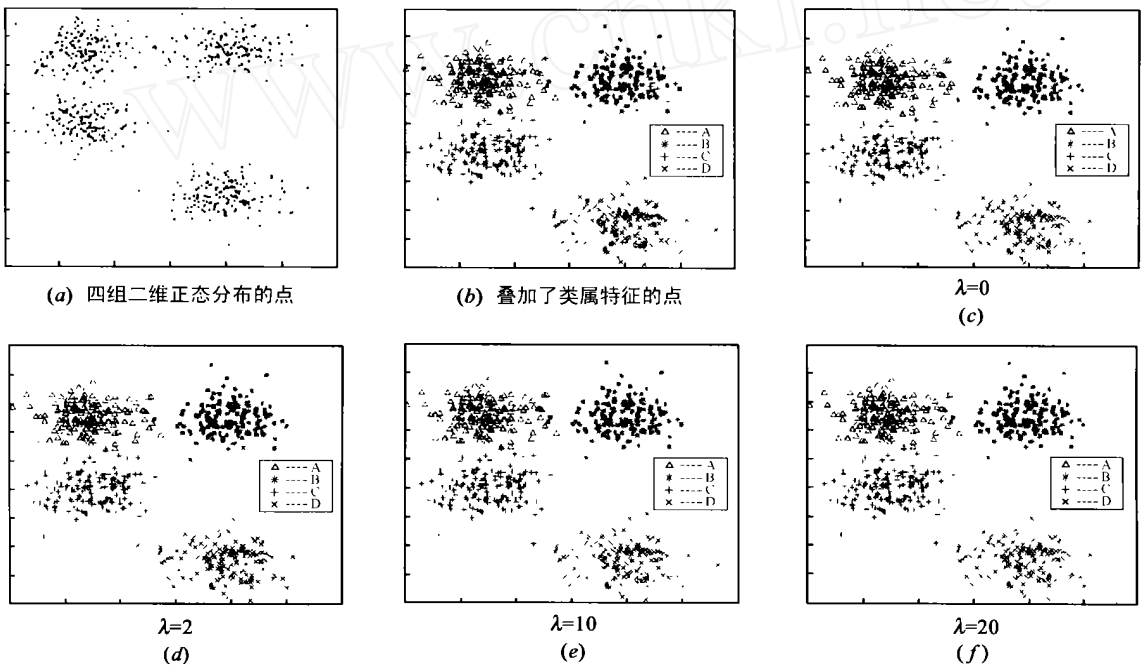
图 4　测试结果(二)

图 4(*a*)分别是基于 CSA 聚类算法和基于遗传算法的在线和离线特性. 从中可以看到:在每一代中,基于克隆算法得到的最优解的抗体-抗原亲合度比遗传算法的要好,而且收敛速度更快,说明在聚类过程中,克隆算法在保证候选解多样性的基础上,性能比遗传算法更优越. 图 4(*b*)是我们提出的增加了一步迭代算子的 CSA 聚类新算法以及标准 CSA 聚类算法的离线和在线特性,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有了明显的提高.

4.4　λ 取值对分类结果的影响

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.

为了便于直观显示,我们同样构造仅具有三个特征的数据样本集,如图 5(*a*)和(*b*)所示. 图 5(*c*)～(*f*)分别是取不同值时的 CSA 分类结果. 当 $\lambda = 0$ 时,聚类仅仅取决于数值特征,即目标的位置如图 5(*c*)所示. 当 $\lambda > 0$ 时,我们看到左上部分中一些样本由于具有类属特征 *B* 而被划分为右上边一类. 当逐渐增大,左上部分中具有类属特征 *B* 被划分为右上边一类 λ 的样本数就越多.



(*a*) 四组二维正态分布的点　(*b*) 叠加了类属特征的点　λ=0 (*c*)

λ=2 (*d*)　λ=10 (*e*)　λ=20 (*f*)

图 5　测试结果(三)

我们可以用图 6 来说明(三角和圆分别表示两种类属特征). 当 $\lambda = 0$ 时,聚类就仅仅取决于数值特征. 聚类结果由垂直虚线分开. 当 $\lambda > 0$ 时,则样本 *c* 可以变成右边一类,因为它接近右边一类,而且其属性特征与右边一类的大多数相同. 同样,样本 *d* 可以变为左边一类. 然而,样本 *a* 还是停留在左边一类,即使其属性特征与右边一类的大多数相同,因为它离右边太远了. 同样,样本 *e* 留在右边. 样本 *b* 是不确定的,它取决于 λ 是偏向数值特征还是偏向属性特征. 如果它是偏向属性特征的,样本 *b* 就分为右类. 相反,它就留在左边.

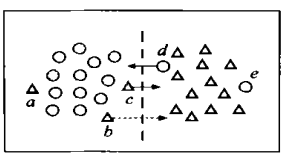


图 6　λ 对聚类结果的影响

4.5　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-原型算法及遗传算法的比较实验. 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 7 我们看出,本文算法的收敛速度明显高于遗传算法,而且在每一次迭代过程中,基于 CSA 的聚类算法也比 k-原型算法及遗传算法的目标函数值要小得多. 说明其聚类效果明显优于 k-原型算法及遗传算法. 同时当数据量逐渐增大时,其收敛速度会有所降低,而数据集的可分性越好,相应的收敛速度就会越快. 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.

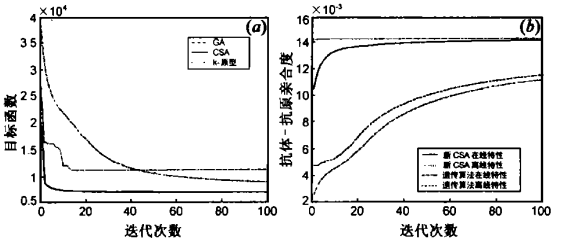


图 7 测试结果(四)

5 结论

本文提出将克隆选择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的聚类结构. 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 CSA 的算法收敛速度快,且不依赖于初始原型的选择,能以概率 1 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克隆选择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问题:数据集应该划分为几类才合适?属于聚类有效性问题以及 λ 的最优取值问题,这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 [A]. Advances in Knowledge Discovery and Data Mining [C]. USA: AAAI Press/ The MIT Press, 1996. 573 - 592.

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

[3] Anderberg M R. Cluster Analysis for Applications [M]. New York: Academic Press. 1973.

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

[5] Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data Mining [A]. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery [C]. USA: ACM Press, 1997. 1 - 8.

[6] 李洁,高新波,焦李成,基于 GA 的混和属性特征大数据集聚类算法研究[R]. 陕西西安:西安电子科技大学,2002.

[7] Jungwon Kim, P J Bentley. Towards an artificial immune system for network intrusion detection: An investigation of clonal selection with a negative selection operator [A]. Proceedings of the 2001 Congress on Evolutionary Computation [C]. USA: IEEE Press, 2001, 2: 1244 - 1252.

[8] Haifeng DU, Licheng JIAO, Sun'an Wang. Clonal operator and antibody clonal algorithm [A]. Proceedings of the First International Conference on Machine Learning and Cybernetics [C]. USA: IEEE Press, 2002. 506 - 510.

[9] 周光炎,免疫学原理(Principles of Immunology) [M]. 上海:上海科学技术出版社.

作者简介:

李 洁 女,1972 年生于陕西西安,工学硕士,西安电子科技大学讲师,现为西安电子科技大学电子工程学院博士研究生,主要从事人工智能、模式识别,数据挖掘等方面的研究.

高新波 男,1972 年生于山东莱芜,工学博士,西安电子科技大学教授,硕士生导师,IEEE 会员,中国电子学会高级会员,主要从事智能信息处理、计算机视觉、基于内容的图像与视频信息检索等领域的研究.

焦李成 男,1959 年 10 月生于陕西白水,1984 年和 1990 年在西安交通大学分别获得硕士和博士学位,IEEE 高级会员,现为西安电子科技大学教授,博士生导师,主要从事非线性科学和智能信号处理以及神经网络与大规模并行处理等领域的研究.