利用混沌差分进化算法预测 RNA 二级结构*)

胡桂武[1] 彭 宏[2]

(广东商学院数学与计算科学系 广州 510320)[1] (华南理工大学计算机科学与工程学院 广州 510640)[2]

摘 要 RNA 二级结构预测在生物信息学中具有重要意义。本文针对 RNA 二级结构预测,提出了一种混沌差分进化算法。算法对种群进行混沌初始化,利用混沌扰动产生新的个体,缩小搜索空间;根据个体的适应值和种群密度自适应地对个体进行混沌更新,改善了种群的多样性。该算法充分利用了差分进化算法速度快以及混沌的遍历性、随机性和规律性等特点,有效克服了早熟现象,提高了算法的全局搜索能力。实验证明了算法的有效性。

关键词 RNA 二级结构,生物信息学,差分进化算法,混沌

An Algorithm-based Chaos Differential Evolution for Predicting RNA Secondary Structure

HU Gui-Wu[1] PENG Hong[2]

(Department of Mathematics & Computational Science, Guangdong University of Business Studies, Guangzhou 510320)[1]

(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640)[2]

Abstract The prediction of RNA secondary structure has important significance in bioinformatics. A chaos differential evolution (CDE) has been proposed for prediction of RNA secondary structure. The basic principle of CDE is that population is initialized by chaos, the chaos disturbance is used to get new individuals directly and reduce search space. The individuals is updated by chaos according to individual fitness and its density and improve the diversity of population. The algorithm not only sufficiently exerts the quick speed of differential evolution and chaotic characteristics-randomness, ergodicity and regularity, but also the global search capability of the algorithm has been enhanced badly, and the premature of algorithm is avoided effectively. The experiments show that the algorithm is effective.

Keywords RNA secondary structure, Bioinformatics, Differential evolution(DE), Chaos

1 引言

RNA 的结构分为一级结构、二级结构和三级结构。其中一级结构是由四种碱基 A、G、C 和 U 组成的一个有限线性序列;二级结构是由 RNA 单链自身回折而形成部分碱基配对和单链交替出现的茎环结构;三级结构是二级结构在空间的进一步折叠形式。RNA 许多功能的实现需借助一定的二级结构,甚至三级结构,因此二级结构预测是一项非常重要的工作。

由于 RNA 分子通过生化实验方法去测定 RNA 分子的立体结构很不容易,并且代价高昂,虽然测得的结果比较精确可靠,但是面对当前海量的生物序列,这种方法显然跟不上要求。故而像蛋白质结构研究一样,借助计算机、生物、化学、物理以及各种数学技术预测 RNA 的空间结构,是提高认识 RNA 空间结构的有效方法。目前具代表性的方法有比较序列分析方法(或者叫系统发育方法)、动态规划算法、组合优化算法以及一些智能化的启发式算法等[1~5]。由于问题本身的复杂性,所有的算法都没有达到令人满意的结果。本文尝试用差分进化算法(Differential Evolution, DE)[6]去预测 RNA 的二级结构。

差分进化算法是一种模拟生物进化的演化算法,最早由 Rainer Storn 和 Kenneth Price 在 1996 年为求解切比雪夫多项式而提出。差分进化算法的优点是速度快、鲁棒性好、在实数域上搜索能力强等。已经在许多经典优化问题上取得了比较好的结果,目前已经引起了国内外学者的广泛关注。但差分进化算法仍然存在许多缺陷,如无法保证收敛到全局最优解、计划过程中存在过早收敛等问题。为了避免出现这些问题,前人曾经做了许多局部改进的工作[7,8],但与混沌优化算法的结合,并且应用于 RNA 二级结构预测,据笔者所知还没有相关的研究。

混沌是自然界广泛存在的一种非线性现象,它看似混乱,却有着精致的内在结构,具有"随机性"、"遍历性"及"规律性"等特点[9],利用混沌运动的这些性质可以进行优化搜索。

本文把混沌特性融入到差分进化算法中,提出了混沌差分进化算法(Chaos Defferential Evolution,简称 CDE)。利用混沌初始化改善个体质量,利用混沌扰动避免搜索过程陷入局部极值,提高差分进化算法的全局搜索能力,克服差分进化算法易陷入局部极小的固有缺陷。

2 混沌差分进化算法

2.1 差分进化算法

差分进化是一种典型的演化算法,它的整体结构类似于遗传算法。与遗传算法的主要区别在变异操作上,其余操作和遗传算法类似。

设求解 n 维的优化问题:

*)国家自然科学基金重点项目(编号:30230350)资助、广东省自然科学基金(编号:06301003)资助。**胡桂武** 副教授,博士,主要研究方向:计算智能、生物信息学、数据挖掘;**彭 宏** 教授,博导,主要研究方向:人工智能、数据挖掘。

$$\min f(x_1,x_2,\cdots,x_n)$$
$$\text{s.t. } a_i\leqslant x_i\leqslant b_i\ i=1,2,\cdots,n \tag{1}$$

解优化问题的差分进化算法步骤如下[6]：

Step1：随机生成含有 N 个个体的初始种群，依次对每一个个体进行如下操作。

Step2：通过变异操作产生变异向量 $\vec{v}_{i,G}$，生成的方法如下：

$$\vec{v}_{i,G}=\vec{x}_{r1,G}+\eta\cdot(\vec{x}_{r2,G}-\vec{x}_{r3,G}) \tag{2}$$

其中 $r1,r2,r3\in\{1,\cdots,N\}$ 为随机整数，表示个体在种群中的序号；$\eta\in[0,2]$ 为收放因子，其中控制了序号为 $r1$ 和 $r2$ 两个向量的差异向量的放大量。

Step3：经交叉操作生成的探测向量为

$\vec{x}'_{iG}=[x'_{1G},x'_{2G},\cdots,x'_{nG}]$

生成的方法如：

$$x'_{ji}=\begin{cases}v_{ji}\ \text{if } rb(j)\leqslant CR \text{ or } j=rd(i)\\ x_{ji}\ \text{if } rb(j)>CR \text{ and } j\neq rd(i)\end{cases} \tag{3}$$

其中：v_{ji} 是 $\vec{v}_{i,G}$ 的第 j 个分量；x_{ji} 是 $\vec{x}_{i,G}$ 的第 j 个分量；$rd(j)$ 是在[0,1]之间的随机数，$rd(j)$ 是[1,n]之间的随机整数；CR 一般是[0,1]之间的随机数。

Step4：差分进化的选择操作对于最小化问题的定义如下：

$$\vec{x}_{i,G+1}=\begin{cases}\vec{x}'_{i,G},\text{if } Fit(\vec{x}'_{i,G})\leqslant Fit(\vec{x}_{i,G})\\ \vec{x}_{i,G},\text{otherwise}\end{cases} \tag{4}$$

其中 Fit(X)为向量的适应值函数。

Step5：通过以上差分进化的变异交叉和选择操作使种群进化到下一代，反复循环进化最后种群将达到最优。

从上述寻优过程可以看出，DE 进化的本质是利用了群体中向量(个体)的距离和方向信息，比较容易实现。目前 DE 发展很快，出现了不同的工作策略，一般用 DE/x/y/z 表示[6,7]，其中 DE 是指差分进化算法，x 表示 DE 变异时是使用"rand"(随机个体)"best"(最好个体)，y 为差异向量的个数，z 代表交叉操作方案是"bin"(二项式)还是"exp"(指数)。其中二项式交叉在交叉时对 D 维空间的每一个变量生成随机数判断，而指数交叉在交叉时只在 D 维空间生成随机数。下面是几种有代表性的模式：

DE/best/1/exp $\vec{v}_{i,G}=\vec{x}_{best,G}+F\cdot(\vec{x}_{r2,G}-\vec{x}_{r3,G})$ (5)

DE/rand/1/bin $\vec{v}_{i,G}=\vec{x}_{r1,G}+F\cdot(\vec{x}_{r2,G}-\vec{x}_{r3,G})$ (6)

DE/randtobest/1/bin $\vec{v}_{i,G}=\vec{x}_{i,G}+F\cdot(\vec{x}_{best,G}-\vec{x}_{i,G})+F\cdot(\vec{x}_{r1,G}-\vec{x}_{r2,G})$ (7)

DE/best/2/bin $\vec{v}_{i,G}=\vec{x}_{best,G}+F\cdot(\vec{x}_{r1,G}+\vec{x}_{r2,G}-\vec{x}_{r3,G}-\vec{x}_{r4,G})$ (8)

DE/rand/2/bin $\vec{v}_{i,G}=\vec{x}_{r5,G}+F\cdot(\vec{x}_{r1,G}+\vec{x}_{r2,G}-\vec{x}_{r3,G}-\vec{x}_{r4,G})$ (9)

表中 $\vec{x}_{best,G}$ 表示第 G 代种群中最好的个体；r_i 为随机整数，表示个体在种群中的序号。

2.2 混沌差分进化算法

目前混沌尚无严格的定义，通常将由确定性方程得到的具有随机性的运动状态称为混沌。Logistic 映射就是一个典型的混沌系统，迭代公式如下：

$$z_{i+1}=\mu z_i(1-z_i),i=0,1,2\cdots \tag{10}$$

式中 μ 为控制参量。利用混沌运动特性可以进行优化搜索，其基本思想是首先产生一组与优化量相同数目的混沌变量，用类似载波的方式将混沌引入优化变量使其呈现混沌状态，同时把混沌运动的遍历范围放大到优化变量的取值范围，然后直接利用混沌变量进行搜索。

混沌变量的"随机性"、"遍历性"，以及对初始条件的敏感性等特点，基于混沌搜索技术无疑会比其它随机搜索更具有优越性，本文的混沌差分进化算法是用混沌初始化，然后进化过程中加入了混沌扰动。

为了提高差分进化算法摆脱局部极值的能力，本文设计了混沌替换算子。如果差分进化算法连续一定的代数，群体中最优个体没有变得更优，即已经陷入局部最优，则选取一定数量的个体用混沌系统生成的个体替换，改善种群的多样性，跳出局部最优。替换的原则是：适应值越低、密度越高的个体被替换的概率越高。这样，既保证了群体中个体的多样性，又避免陷于局部极值，提高算法整体的收敛速度。这种思想可以与任何模式的差分进化算法结合。下面以基本差分进化算法为例，说明混沌差分进化算法。替换过程的关键是替换概率的定义。

定义 1(替换概率) 首先对种群中的个体按适应值从大到小排列，则第 i 个个体的替换概率 p_i 由个体本身的浓度概率 p_{id} 和第 $M-i$ 个个体的适应度概率 $p_{(M-i)f}$（M 是种群规模）决定，具体

$$p_i=\alpha\, p_{(M-i)f}+(1-\alpha)p_{id}\quad(0\leqslant\alpha\leqslant1) \tag{11}$$

$p_{if}=\dfrac{f_i}{\sum f_i}$，($f_i$ 是个体 i 的适应值) $p_{id}=\dfrac{m}{M}$，m 是与第 i 个个体距离小于固定值的个体总数，M 是群体规模。基于以上的设计，求解 n 维数优化问题(1)的混沌差分进化算法(Chaos Differential Evolution，简称 CDE)步骤如下：

Step1：混沌初始化种群。随机产生一个 n 维、每个分量数值在 0～1 之间的向量 $z_1=(z_{11},z_{12},\cdots,z_{1n})$。

根据(10)式 $z_{i+1j}=\mu z_{ij}(1-z_{ij})(j=1,2,\cdots,n;i=1,2,\cdots,N-1)$得到 N 个 $z_1,z_2,\cdots,z_N$。将 z_i 的个分量载波到优化变量的范围：$x_{ij}=a_j+(b_j-a_j)z_{ij}(j=1,2,\cdots n;i=1,2,\cdots,N)$。计算目标函数，从 N 个初始群体中选择性能较好的 m 个解作为初开解，依次对每一个个体进行差分进化算法中的如下操作。

Step2：执行变异操作；

Step3：执行交叉操作；

Step4：执行选择操作；

Step5：执行混沌扰动操作，通过(10)式，产生，$u_0=(u_{01},u_{02},\cdots,u_{0n})$，$u_{1j}=\mu u_{0j}(1-u_{0j})(j=1,2,\cdots,n)$，$u_1=(u_{11},u_{12},\cdots,u_{1n})$

将 u_1 的各个分量载波到优化变量的范围$[-\beta,\beta]$内，扰动量

$$\Delta x=(\Delta x_1,\Delta x_2,\cdots,\Delta x_n)\Delta x_i=-\beta+2\beta u_{1j},\vec{x}'_{iG+1}=\vec{x}_{iG+1}+\Delta x$$

计算 $\vec{x}_{iG+1}$，$\vec{x}'_{iG+1}$ 的适应值 f 和 f'。如果 $f'<f$，则 $\vec{x}_{iG+1}=\vec{x}'_{iG+1}$。

Step6：通过以上操作使种群进化到下一代，如果满足收敛条件，则转 Step7，否则如果连续一定的代数，群体中最优个体没有变得更优，则按照替换概率选择一定比例的个体由混沌系统生成的个体替换，转 Step2，否则直接转 Step2；

Step7：输出全局最优，算法结束。

3 混沌差分进化算法在 RNA 二级结构预测中的应用

3.1 RNA 二级结构预测的算法基础

基于论文的需要，在此引用了文[10]中 RNA 二级结构的相关概念。

定义 2 一个长度为 n 的 RNA 序列 $R=r_1, r_2, r_3, \cdots, r_n$ 的二级结构定义为碱基对集合 $S=\{(r_i, r_j)\}$，其中 (r_i, r_j) 满足如下几条：

(1) $(r_i, r_j) \in \{(A,U),(U,A),(G,C),(C,G),(G,U),(U,G)\}$，$1 \leqslant i < j \leqslant n$ 且 $j-i>3$。

(2) 若 $(r_i, r_j) \in S, (r_k, r_l) \in S$，则 $i=k$ 当且仅当 $j=l$。

(3) 若 $(r_i, r_j) \in S, (r_k, r_l) \in S, i<k$，则它们只有串联和并联两种位置关系，即 $i<k<l<j$，或 $i<j<k<l$。

该定义规定了本文算法所处理的 RNA 二级结构只能是常规碱基配对，发夹环环区单链至少为 2 个碱基长，并且不考虑三联碱基对以及假结等情况。

定义 3 设 R 是一长度为 n 的 RNA 序列，$R_1=r_i, r_{i+1}, \cdots r_{i+k-1}$ 和 $R_2=r_{j-k+1} r_{j-k+2} \cdots r_j$ 是 R 的两个子序列。如果 R_1 和 R_2 中的碱基依次互补配对(即 (r_{i+t}, r_{j-t})，$t=0,1,\cdots,k-1$，且满足定义 2 中性质，则称 R_1 和 R_2 在 R 的二级结构 S 中构成一个茎，记为 $S(i,j,k)$，并称 R_1 为茎的前段，R_2 为茎的后段. i 和 j 。

定义强调，只要是连续的碱基配对，则只算一个茎区。若要算两个茎，则中间至少隔着一个未配对碱基。这里设定茎的长度 $k \geqslant 3$，因为一般认为少于 3 对连续的碱基配对区是不稳定的。

定义 4 给定两个茎 $S_1(i_1, j_1, k_1)$ 和 $S_2(i_2, j_2, k_2)$，若 S_1, S_2 既不重叠又不交叉，即满足条件：$\{[i_1, i_1+k_1-1] \cup [j_1-k_1+1, j_1]\} \cap \{[i_2, i_2+k_2-1] \cup [j_2-k_2+1, j_2]\} = \Phi$，则称茎 $S_1(i_1, j_1, k_1)$ 和 $S_2(i_2, j_2, k_2)$ 是相容的。

给定一个 RNA 序列，其二级结构有许多折叠形式。而且可以证明，随着序列长度的增加，其二级结构的数目将按指数级增长，但其真实结构(或称自然结构)只有一个。从众多结构中确定哪一个是其真实结构的问题称为 RNA 二级结构预测问题。由上述定义和假设可以得到如下明显的结论：

命题 1 设 $\{s_1, s_2, \cdots, s_m\}$ 是 RNA 序列 R 中的一个茎序列，若其中任意两个茎 s_i，s_j $(1 \leqslant i, j \leqslant m)$ 都相容，则 $\{s_1, s_2, \cdots, s_m\}$ 可以惟一地确定 R 上的一个二级结构，记为 $S=s_1, s_2, \cdots, s_m$。

命题 2 设 R 是一长度为 n 的 RNA 分子，S 由 R 折叠而成的任一个二级结构，则 S 包含的茎数不超过 $(n-2)/7$。

现在普遍认为，RNA 的自然二级结构应该是稳定的。根据热力学理论，物体在稳定状态时其自由能最小。所以，现有的大多数预测算法都是基于这种假设来设计的。

下面给出 RNA 二级结构预测问题的一个组合优化描述。

定义 5 对于给定的 RNA 序列，假设茎区池(stem pool，SP)表示该序列所有可能的茎区集合，$E_{total}(S)$ 表示一个二级结构 S 的总体自由能，则问题描述如下：

求茎区子集 $\{s\,i1, s_{i2}, s_{i3}, \cdots, s_{ik}\} \subset SP$，使得由该子集构成的二级结构 $S^*=s_{i1}, s_{i2}, s_{i3}, \cdots, s_{ik}$ 有

$$E_{total}(S^*)=\min E_{total_s}(S) \tag{12}$$

s. t. 茎区子集 $\{s_{i1}, s_{i2}, s_{i3}, \cdots, s_{ik}\}$ 满足相容性条件，其中 $E_{total}=E_{stack}+E_{hairpin}+E_{bulge}+E_{internal}+E_{multi}$，即一个 RNA 二级结构 S 的总体自由能等于其各结构单元自由能之和。对于多分支环的情形，则采用如下线性近似：$E_{stack}=4.6+0.2n+0.1h$，其中 n, h 分别表示该多分支环中非配对碱基数目和茎区数目。

3.2 面向 RNA 二级结构预测的混沌差分进化算法

在实施算法之前利用螺旋区点阵作图法[10]找出所有可能的茎区，建立包含 RNA 序列的所有茎 $s(i,j,k)$ 的茎列表 B，并将茎按 i 升序、j 降序用自然数进行编号，编号的最大数目即茎的个数 M。

1)算法编码

在本文的算法中，个体采用实数串表示，实数串的长度根据命题 2 定义为 $(\mathrm{Len}-2)/7$，Len 为 RNA 的长度，实数串中的元素为 0 到 M 的整数。当元素为 0 时，表示没有选择任何茎，为非 0 的 K 时候，表示选择了第 K 个茎。当然个体其对应的茎序列必须是一个相容的茎序列。

2)适应值函数

即为个体所对应的茎序列自由能，适应值越小表示其个体越优，并将适应值并作为以后操作的依据。

定义 6(状态空间与位置) 状态空间中的位置，即本节算法编码中的序列，状态空间即搜索空间，记为：$\Omega=\{(x_1, x_2 \cdots x_{\mathrm{Length}}) \mid x_i \in \{0,1,\cdots,M\}\}$，$M$ 为被预测的 RNA 序列茎的个数。其中 Length 为个体长度。状态空间中的位置也就是差分进化算法中的个体，记 $X=(x_1 x_2 \cdots x_{\mathrm{Length}})$。为了定义方便，令 $X_{\max}$ 为 x_i 的最大值，其值为 M。

定义 7(个体与个体的加法) 假设两个个体 X 和 Y，令 $W=X+Y$，$X=(x_1, x_2, \cdots, x_k)$，$Y=(y_1, y_2, \cdots, y_k)$，

$$W=(w_1 w_2 \cdots w_k) w_i=\begin{cases} X_i+Y_{i, X_i+Y_i \leqslant X_{\max}} \\ [\widetilde{w}_i], X_i+Y_i \geqslant X_{\max} \end{cases} \tag{13}$$

其中 $0 \leqslant \widetilde{w}_i \leqslant x_{\max}$，$x_i+y_i=\widetilde{w}_i+lX_{\max}$，$[x]$ 为取整函数，l 为自然数。

定义 8(个体与个体的加法) 假设两个个体 X 和 Y 令 $W=X-Y$，$X=(x_1, x_2, \cdots x_k)$，$Y=(y_1, y_2, \cdots, y_k)$，

$$W=(w_1, w_2 \cdots w_k), w_i=\begin{cases} X_i-Y_{i, X-Y_i \geqslant 0} \\ [\widetilde{w}_i], X_i-Y_i \leqslant 0 \end{cases} \tag{14}$$

其中 $0 \leqslant [\widetilde{w}_i] \leqslant X_{\max}$，$\widetilde{w}_{ji}=(x_i-y_i)+lX_{\max}$，$[x]$ 为取整函数，l 为自然数。

定义 9(个体的数乘)

$X=(x_1, x_2, \cdots, x_k)$，令 $cX=(w_1, w_2, \cdots, w_k)$，

$$w_i=\begin{cases} [cx_i], 0 \leqslant cx_i \leqslant X_{\max} \\ [\widetilde{w}_i], cv_i \geqslant X_{\max} \end{cases} \tag{15}$$

$0 \leqslant [\widetilde{w}_i] \leqslant X_{\max}$，$cx_i=\widetilde{w}_i+lX_{\max}$，$l$ 为自然数。

预测算法采用本文第 2 部分的混沌差分进化算法(CDE)，详细过程如下：

Step0：建立被预测 RNA 序列的茎区池；

Step1：混沌初始化种群；

Step2：执行变异操作；

Step3：执行交叉操作；

Step4：执行选择操作；

Step5：执行混沌扰动操作；

Step6：通过以上操作使种群进化到下一代。如果满足收敛条件，则转 Step7，否则如果连续一定的代数，群体中最优个体没有变得更优，则按照替换概率选择一定比例的个体由混沌系统生成的个体替换，转 Step2，否则直接转 Step2；

Step7：输出最优 RNA 二级结构，算法结束。

4 实验与分析

为了评价本文提出的混沌差分进化算法(CDE)对组合优化问题的求解性能，分别将 CDE 中的差分进化模式使用本文第 2 部分的(6)(记为 DE_1)、(7)(记为 DE_2)，得到的混沌优化算法分别为分别为 CDE_1、CDE_2。测试的数据是长度为 359 的 PSTVd 碱基序列。

5′-CGGAACUAAA[10]CUCGUGGUUC[20]CUGUGGUUCA[30]
CACCU GACCU[40]CCU GAGCAGA[50]AAAGAAAAAA[60]
GAAGGCGGCU[70]CGGAGGAGCG[80]CUUCAGGGAU[90]
CCCCGGGGAA[100]ACCUGGAGCG[110]AACUGGCAAA[120]
AAAGGACGGU[130]GGGGAGUGCC[140]CAGCGGCCGA[150]
CAGGAGUAAU[160]UCCCGCCGAA[170]ACAGGGUUUU[180]
CACCCUUCCU[190]UUCUUCGGGU[200]GUCCUUCCUC[210]
GCGCCCGCAG[220]GACCACCCCU[230]CGCCCCCUUU[240]
GCGCUGUCGC[250]UUCGGCUACU[260]ACCCGGUGGA[270]
AACAACUGAA[280]GCUCCCGAGA[290]ACCGCUUUUU[300]
CUCUAUCUUA[310]CUUGCUUCGG[320]GGCGAGGGUG[330]
UUUAGCCCUU[340]GGAACCGCAG[350]UUGGUUCCU[359]—3′

其真实的二级结构匹配碱基总数为 246，茎总数为 25 个，具体如下：

s_1(3,357,7)，s_2(14,348,8)，s_3(25,337,4)，
s_4(30,331,6)，s_5(39,322,4)，s_6(44,317,6)，
s_7(52,309,4)，s_8(60,300,9)，s_9(69,289,5)，
s_{10}(80,282,7)，s_{11}(90,270,3)，s_{12}(93,266,5)，
s_{13}(103,255,8)，s_{14}(114,246,4)，s_{15}(121,240,5)，
s_{16}(128,234,7)，s_{17}(136,226,3)，s_{18}(140,221,2)，
s_{19}(143,218,4)，s_{20}(148,213,2)，s_{21}(152,209,5)，
s_{22}(159,202,2)，s_{23}(162,199,4)，s_{24}(168,193,4)，
s_{25}(173,186,5)。。

表 1 算法预测结果的比较

算法	碱基配对正确率(平均)	茎区正确率(平均)
DE_1	78.3%	70.1%
CDE_1	86.3%	81.8%
DE_2	80.91%	72.4%
CDE_2	89.93%	84.3%

表 1 是 4 种算法 DE_1、DE_2、CDE_1、CDE_2 独立运行 20 次，预测获得的 RNA 二级结构碱基配对正确率(平均)和茎区正确比率(平均)的统计结果

实验结果表明本文的混沌差分进化算法的精确度比差分进化算法要高，其中碱基配对预测正确率要高 10%左右、茎区预测正确率要高 16%左右，验证了算法的有效性。

结论 为了借助非线性混沌本质来有效改善差分进化算法，提高算法的性能，本文将混沌优化搜索技术融入到差分进化算法，提出了混沌差分进化算法，该算法不仅保持了差分进化算法简单的优点，而且充分利用了混沌的随机性、遍历性和规律性等特点，有效克服了算法的早熟的缺陷，提高了全局最优解的计算效率，是一种高效的差分进化算法，具有较大的使用价值。另外据笔者所知，本文是混沌差分进化算法在 RNA 二级结构预测中的首次应用，是一种很好的应用尝试。对混沌差分进化算法进行理论上的分析以及在其它方面的应用是我们未来研究的方向。

参 考 文 献

1 Cai L, Malmberg R L, Wu Y. Stochastic modeling of RNA pseudoknotted structures: a grammatical approach. Bioinformatics, 2003, 19: 66～73
2 TAN Guang-Ming, FENG Sheng-Zhong, SUN Ning-Hui. An optimized and efficiently parallelized dynamic programming for RNA secondary structure prediction. J Software, 2006, 17(7): 1501～1509
3 Hofacker I L, Schuster P. Combinatorics of RNA Secondary Structure. Discr Appl Math, 1998, 88: 207～237
4 Nebel M E. Identifying Good Predictions of RNA Secondary Structure. In: Proceedings of the Pacific Symposium on Biocomputing 2004, 2004. 423～434
5 Kirkpatrick S, Gelatt C D, Vecchi M P. Optimization by simulated annealing. Science, 1983, 220: 671～680
6 DE Homepage. Http: //www. icsi. Berkeley. Edu/storn/code. htm
7 Shi Yan-jun. Teng Hong-fei, Li Zi-qiang. Cooperative Co-evolutionary Differential Evolution for Function Optimization. Lecture Notes in Computer Science, 2005, 1075～1083
8 Ali M M, Fatti L P. A differential free point generation scheme in the differential evolution algorithm. Journal of Global Optimization, 2006, 35 (4): 551～572
9 Zhang Huanguang, Wang Zhiliang, Huang Wei. Control theory of chaos system[M]. Shen Yang: Publishing House of Northeast University, 2003. 1～45
10 王翼飞，等. 生物信息学—智能计算算法及其应用. 北京：化学工业出版社，2006. 172～210

(上接第 130 页)

一种基于模糊聚类的协同信息推荐算法，通过实验结果和数据分析，基于模糊聚类的协同推荐方法较一般的协同推荐(User -based 和 Item-based)其查全率有了很大的提高，提高了推荐的质量和精度，而又有相对较小的平均绝对偏差 MAE，实验中和传统的基于用户的和基于项目的协同推荐进行了比较，充分证实了实行模糊聚类推荐的有效性。用目标用户相对于聚类后的用户组群的兴趣隶属度来描述用户兴趣更能真实地反映金融用户的需求，实验中的聚类算法采用了模糊 c 均值聚类法进行用户聚类，对传统的协同推荐方法做了一步改进。

然而，随着 Internet 信息的不断增长，如何更为有效地组织信息资源，加上金融领域信息的时效性特点，虽然以用户组群的兴趣爱好为参考，能解决一点数据稀疏的问题，如何客观描述用户在某个特定领域的最小最全需求，进一步解决用户-项目矩阵的数据稀疏问题，提高系统的推荐效率仍然是协同推荐中一个必须解决的问题。

参 考 文 献

1 Ma Zhaofeng, Feng Boqin. Support Vector Machines Learning for Adaptive and Active Information Retrieval. Advanced Web Technologies and Applications (APWEB'04), Lecture Notes Computer Science, 2004, 3007: 89～99
2 马兆丰，冯博琴. 基于支撑向量机的自适应信息推荐算法. 小型微型计算机系统，2004，25(3)：384～387
3 Goldberg D, Nichols D, Oki B, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35 (12): 61～70
4 Konstan J, Miller B, Maltz D, et al. GroupLens: applying collaborative filtering to usenet news. Communications of the ACM, 1997, 40(3): 77～87
5 Shardanand U, Maes P. Social information filtering: algorithms for automating "Word of Mouth"[C]. In: Proceedings of ACM CH I'95 Conference on Human Factors in Computing System s, 1995. 210～217
6 Hill W, Stead L, Rosenstein M, et al. Recommending and evaluating choices in a virtual community of Use[C]. In: Proceedings of CH I'95, 1995. 194～201
7 张巍. 刘鲁，葛健. 一种基于粗糙集的协同过滤算法 小型微型计算机系统，2005，26(11)
8 曾艳，麦永浩. 基于内容预测和项目评分的协同过滤推荐. 计算机应用，2004，24(1)
9 O'ConnerM, Herlocker J. Clustering items for collaborative filtering[C]. In: Proceedings of the ACM SIGIR Workshop on Recommender System s. Berkeley, CA, 1999
10 林鸿飞，杨志豪，赵晶. 基于内容和合作模式的信息推荐机制. 中文信息学报，2005，19(11)：1003～0077
11 孙汝杰，张宇光. 基于时间序列的个性化信息协同过滤技术研究. 情报杂志，2006(8)
12 Mobasher B, Jin X, Zhou Y. Semantically enhanced collaborative filtering on the web[C]. In: Proceedings of the European Web Mining Forum, 2004
13 Hill F, Stead L, Rosenstein M, et al. Recommending and Evaluating Choices in a Virtual Community of Use[C]. In: Proceedings of ACM CHI'95Conference on human factors in computing systems, 1995. 210～217