结合模拟退火算法的遗传 K-Means 聚类方法

凌　静,江凌云,赵　迎
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘　要:K-Means 算法是一种经典的基于划分的聚类方法。传统的 K-Means 算法中存在很明显的缺陷,它对初始聚类中心的依赖性很大,聚类结果很容易陷入局部最优值;而基于遗传算法改进的 K-Means 聚类方法,提高了聚类结果的稳定性,但因为个体的多样性不足,常常会出现早熟等现象,其局部寻优能力较弱。针对上述问题,文中提出一种结合模拟退火算法的遗传 K-Means 聚类方法。利用模拟退火算法改进遗传算法的变异操作,用 K-Means 操作取代遗传算法的交叉操作,改善早熟现象,避免聚类结果陷入局部最优,实现聚类方法性能的提升。实验结果表明,该方法的聚类准确度比一般 K-Means 方法和遗传 K-Means 方法都要高。

关键词:聚类;K-Means 算法;遗传算法;模拟退火算法
中图分类号:TP18　　**文献标识码**:A　　**文章编号**:1673-629X(2019)09-0061-05
doi:10.3969/j.issn.1673-629X.2019.09.012

A Genetic K-Means Clustering Method Combined with Simulated Annealing Algorithm

LING Jing,JIANG Ling-yun,ZHAO Ying
(School of Telecommunications & Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:K-Means algorithm is one of the most classical division-based clustering methods. In the traditional K-Means algorithm,there are obvious flaws like strong dependence on the initial clustering center and the clustering result is easy to fall into the local optimal value. The improved K-Means clustering method based on genetic algorithm improves the stability of clustering results. However,due to the insufficient diversity of individuals,prematurity and other phenomena often occur,and its local optimization is weak. For this,we present a genetic K-Means clustering method combined with simulated annealing algorithm. The simulated annealing algorithm is used to improve the mutation operation of genetic algorithm,the classical K-Means operation is used to replace the crossover operation of the genetic algorithm,so as to improve the premature phenomenon,avoid the clustering result falling into the local optimal,and improve the performance of the clustering method. The experiment shows that the clustering accuracy of the proposed method is higher than that of the general K-Means method and the genetic K-Means method.

Key words:clustering;K-Means algorithm;genetic algorithm;simulated annealing algorithm

0　引　言

迄今为止已经有了多种聚类算法,根据数据在聚类中的积聚规则,以及应用这些规则的方法,聚类算法主要可以分为[1]:基于划分、基于层次、基于网格、基于密度以及基于模型等类型。聚类算法被广泛应用于模式识别、图像处理、文本检索、网络入侵检测、生物信息学等领域。其中,K-Means 聚类算法是最为经典的一种聚类算法[2],优点是简单有效、收敛速度快、局部搜索能力强;但也存在难以克服的缺陷,如过度依赖初始聚类中心、聚类结果极易陷入局部最优解、全局搜索能力不强等。

针对 K-Means 聚类算法过度依赖初始聚类中心,全局搜索能力不强的问题,已经有了大量的研究成果,如文献[3-4]中提出的遗传 K-Means 算法、优化遗传 K-Means 算法等。这些结合遗传算法的改进方法,有效提高了 K-Means 聚类算法的稳定性和全局性,但遗

收稿日期:2018-10-17　　**修回日期**:2019-02-19　　**网络出版时间**:2019-04-24
基金项目:国家自然科学基金(6127123)
作者简介:凌　静(1993-),女,硕士研究生,研究方向为下一代通信网络技术与物联网技术;江凌云,副教授,硕导,研究方向为下一代网络技术。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190424.1047.036.html

传算法自身存在早熟现象,且其局部寻优能力较弱。

文中提出一种结合模拟退火算法的遗传 K-Means 聚类方法。配合局部寻优能力强的模拟退火算法改进遗传算法的缺陷,得到性能在两者之上的遗传模拟退火算法,再将其应用于 K-Means 算法,从而提高 K-Means 聚类算法的性能,实现聚类结果的优化。

1 相关工作

1.1 K-Means 算法、遗传算法与模拟退火算法

K-Means 算法是一种基于划分的经典聚类算法,是由 Mac Queen 于 1967 年提出的,他结合 Cox、Fisher、Sebestyen 等的研究成果,给出了 K-Means 算法的详细步骤,并用数学方法对 K-Means 算法进行了证明。K-Means 算法的主要思想[5]是:对 n 个给定的对象,给出 k 个划分,每个划分代表一个类,其中 $k \leqslant n$。首先,从给定的所有对象中任意选择 k 个对象,作为 k 个类的聚类中心。对剩余对象,分别计算它们与各个聚类中心的相似度,分到相似度最高的类中;分类完成后,计算新类的平均值作为新的聚类中心,再计算所有对象与新聚类中心的相似度,将对象分到最相似的类中。不断重复直到准则函数的值达到最小,准则函数定义如下:

$$J = \sum_{j=1}^{k} \sum_{x_i \in c_j} \| x_i - z_j \|^2 \tag{1}$$

其中,k 为类别数;x_i 为样本对象;z_j 为类 c_j 的聚类中心。

遗传算法(GA)是一种全局优化自适应概率搜索算法,由 Holland 于 1975 年提出的。该算法模拟了生物的繁衍、交配和变异现象[6],在初始种群的基础上产生新的更适应环境的种群,一代代繁衍进化,最终收敛到一个最适应环境的个体上。在搜索过程中,该算法能够自动获取搜索空间的相关知识,并积累获得的信息;通过对搜索过程的自适应控制,能够获得问题的最优解。遗传算法使用适应度函数作为度量标准,通过计算群体中的每个个体的适应度函数值,来判断个体的优劣程度。适应度函数值越高,个体越优秀,就越有可能被遗传到新种群中,成为最适应环境个体的概率也就越高。一般会根据实际情况设计相应的适应度函数。

模拟退火算法(SA)又被称为模拟冷却法、概率爬山法,是由 Kirpatrick 于 1982 年提出的。模拟退火算法是模拟了一个高温固体的退火过程[7],在搜索过程中,开始先设定一个温和的初始结果作为最优解,然后随机获得一个新解,当得到的新解优于当前最优解时,直接接受新解为最优解;当新解劣于最优解时,以一定的概率接受新解为最优解,随着温度的下降重复上述操作,最终得到全局最优解。模拟退火算法利用了概率的突跳特性,具有并行性和渐近收敛性,理论上能够证明,模拟退火算法是以概率 1 收敛于全局最优解的。

1.2 遗传模拟退火算法

在遗传算法运行过程中,早期种群的个体之间差异较大即个体适应度函数值差异较大,而在通过选择算子生成下一代新种群时,新种群的子个体出现概率与上一代种群中父个体的适应度函数值成正比,也就是说,适应度值越高的个体越容易遗传到下一代种群中,这就容易出现优秀个体占领整个种群,形成早熟现象。后期,整个种群中的个体适应度值基本一致,差异较小,这就导致优秀个体在生成下一代种群个体时的优势较小,造成整个种群的进化停滞。因此,在算法运行过程中可以对个体的适应度函数值进行适当拉伸。

模拟退火算法中按照 Metropolis 准则[8]接受新解,除了接受优于当前最优解的新解作为新的最优解,还能以一定的概率接受劣于当前最优解的新解。在算法早期,温度值 T 较大,能够接受较劣的新解。随着算法不断运行,T 值也在不断变小,当前最优解的值会越来越逼近整体最优解,当 T 值接近 0 时,当前最优解最接近整体最优解,能够避免算法陷入局部最优。

遗传模拟退火算法是一种优化算法[9],在算法前期,种群中个体的适应度相差较大即存在较为突出的优良个体,而此时温度较高,有较大可能接受较差的个体,避免种群过早集中于优良个体;而在算法后期,种群中个体的适应度函数值较为接近,此时温度较低,模拟退火算法能对遗传算法中这些个体的适应度函数值进行拉伸,放大这些个体之间的适应度差异,提高优秀个体在选择过程中的优势。遗传模拟退火算法能够更加快速有效地收敛到全局最优解。

已有许多研究尝试将遗传算法与 K-Means 聚类算法进行结合,以改善 K-Means 聚类算法的缺陷。文献[10]将基于准则函数的经典聚类算法 K-Means 引入到遗传算法,用 K-Means 算法的一步—K-means 操作(KMO)代替标准遗传算法中的交叉操作,这样既能利用遗传算法确保聚类结果的稳定性,又能借助 K-Means 算法提高混合算法的收敛速度。

该算法融合了遗传算法与 K-Means 算法,保证了算法的全局搜索能力,也保证了算法的简单有效,同时还具有爬山能力。然而遗传算法自身还存在早熟、局部寻优能力弱等缺点。为此,在文献[10]的基础上,文中提出一种结合模拟退火算法的遗传 K-Means 聚类方法。将模拟退火算法引入已有的遗传 K-Means 算法,保留 K-Means 操作取代交叉操作的方法,通过模拟退火算法增强聚类方法的局部搜索能力,实现聚类结果的进一步优化。

2　结合模拟退火算法的遗传 K-Means 聚类方法

标准遗传算法中包括选择操作、交叉操作以及变异操作。文献[10]提出的方法在遗传算法中引入 K-Means 操作代替交叉操作,文中在其基础上,引入模拟退火算法对遗传算法的变异操作进行改进,改善遗传算法的早熟缺点,避免结果陷入局部最优,提高原有遗传 K-Means 聚类算法的性能,从而实现聚类结果的优化。该方法的整体结构如图 1 所示。

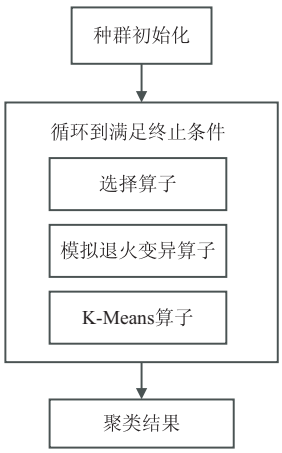

图 1　聚类方法整体结构

2.1　样本编码

样本编码[11]是遗传算法的基础操作,要将问题的解进行编码才能进行后续操作。遗传算法有多种编码方式,如符号编码、二进制编码、浮点数编码等。在聚类样本维度高、数量大时,如果采用传统的二级制编码方式,种群中的个体编码长度会随着维度的增加、精度的提高而出现显著增加的情形,从而导致整个搜索空间的增大,影响聚类方法的计算效率。因此文中采用的是一种基于聚类中心的十进制编码方式。

具体编码方式如下:设一个数据集中的样本个数为 n ,最终聚类的类别数目为 k 。有 k 个聚类中心,每个中心对应一个类别号,计算所有样本到各个聚类中心的距离,将其划分到相应的类中,编码值对应样本所属聚类的类别号,最终编码长度 $l = n$ 。如图 2 所示,S_n 为聚类的类别号,编码总长度为 n 。

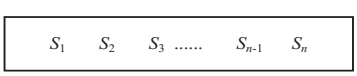

图 2　样本编码

举例:一个数据集为$\{x_1,x_2,x_3,x_4,x_5,x_6\}$,类别数目 $k=3$,分类模式为 1$\{x_1,x_4\}$ 2$\{x_3,x_6\}$ 3$\{x_2,x_5\}$,则编码为(1 3 2 1 3 2),编码长度 $l=6$。

文中采用的这种基于聚类中心的编码方式直观明确,相比二进制编码有效缩短了个体编码的长度,提高了整体的计算效率,对于大数据复杂问题的求解效果较好。

2.2　种群初始化

文中采用随机方式生成初始种群,具体方法为:从样本空间中随机选出 k 个样本作为聚类中心,将所有样本按其到各个聚类中心的距离分类到 k 个类中,得到一个个体,计算此时的个体编码 S_n ;设种群大小为 sizepop,将上述操作重复进行 sizepop 次,即可得到初始种群 P_0。

2.3　适应度函数

适应度函数[12]会影响整个聚类方法的收敛速度,以及对最优解的确定。一般使用适应度函数来衡量个体的适应度,判别该个体在种群中的优劣程度。某个体适应度的值越大,该个体在整个遗传过程中的存活概率也就越大。

K-Means 算法中判断聚类划分质量的标准是准则函数 J , J 的值与所有聚类中的点到相应聚类中心的距离总和相等。J 的值越小,表明该聚类划分的质量越好,反之表明该聚类划分的质量越差。

对于种群中的每个个体,根据准则函数 J 来构造适应度函数,适应度函数定义如下:

$$F(S_i) = 1.5 \times J_{\max} - J(S_i), i = 1,2,\cdots,\text{sizepop} \tag{2}$$

其中, $J_{\max}$ 是种群中所有个体的准则函数值的最大值; $J(S_i)$ 是当前个体的准则函数值。

根据函数定义可以看出,准则函数 J 的值越小,该个体代表的聚类划分的质量越好,适应度函数值越大,其存活概率也就越高。

2.4　选择操作

选择操作[13]遵循优胜劣汰原则,以个体的适应度函数值为基础,由父种群选出新种群。在进行选择操作时,适应度函数值越大的个体经过选择操作后,遗传到新种群中的概率就越高,反之被遗传到新种群中的概率就越小,经过多次选择操作得到的个体组成新种群。

选择操作常用的方法有轮盘赌选择法、最优个体保留法、锦标赛选择法[14],文中使用轮盘赌方法来进行选择操作。轮盘赌方法是将种群中所有个体适应度函数的总和,作为轮盘的整个圆周,按照每个个体的适应度值在总和中所占的比例,为其分配轮盘中相应大小的扇区。每选择一个个体就是随机转动一次轮盘,转动轮盘后选中哪个区域,就选择该区域对应的个体作为新种群的个体。在轮盘赌方法中,面积越大的区域越有可能被选中,反之被选中的概率就越低,而适应度函数值越大的个体其面积也就越大。

种群 P 中第 i 个个体的适应度函数值为 $F(S_i)$,则个体 i 被选中的概率为:

$$P_i = \frac{F(S_i)}{\sum_{i=1}^{k} F(S_i)}, i = 1,2,\cdots,\text{sizepop} \tag{3}$$

在父群体中进行 sizepop 次选择，即生成新种群 P_1。

2.5 模拟退火变异操作

变异操作[15]按位进行，在个体编码时每个样本都有多个可能的编码值，变异就是将指定位置的样本的现有编码值，按变异概率 P_i 用其余的可能值进行替换。

文中使用的是均匀变异操作，具体过程为：对个体编码上的每个样本点，依次进行变异操作，也就是按概率 P_i 从样本现有的类别号中选一个编码值替代原有值，最终得到新个体。变异概率 P_i 定义如下：

$$P_i = \frac{1.5 \times d_{\max}(x_i) - d(x_i - c_k) + 0.5}{\sum_{k=1}^{k} [1.5 \times d_{\max}(x_i) - d(x_i - c_k) + 0.5]} \tag{4}$$

$$d_{\max}(x_i) = \max_k \{ d(x_i - c_k) \} \tag{5}$$

其中，$d(x_i - c_k)$ 是样本 x_i 与第 k 个簇的质心 c_k 之间的欧几里得距离。引入偏差 0.5 是为了避免除 0 错误。这里采用的概率 P_i 不是固定值，使得个体上每个基因座的变异概率都不同，能够大幅度提高个体的变异概率，进一步避免遗传算法的早熟现象。

在均匀变异操作的基础上引入模拟退火算法。具体操作为：首先给定初始温度 T_0，终止温度 T_e 以及模拟退火算法内部最大迭代次数 N。将个体原有的准则函数值作为当前解 f，经过均匀变异操作后形成的新个体的准则函数值作为新解 f'，两者差值记为 $\Delta f = f' - f$。当 $\Delta f \leqslant 0$ 时，直接接受新解为最优解，即将新个体替代种群中的原有个体；当 $\Delta f > 0$ 时，以概率 $p = \exp\frac{-\Delta f}{KT}$ 接收新解为最优解，其中 K 为常数，T 为当前温度。将上述操作重复 N 次，判断当前温度 T 是否达到终止温度 T_e，没有达到就按照降温等式 $T(t) = T_0 \times a \times t$ 来降低当前温度值，其中 a 为降温速度，t 为当前 T 值，再重新进行模拟退火变异迭代；如果达到终止温度 T_e 则终止算法，得到新个体。

对父种群 P_1 中的每个个体都进行上述模拟退火变异操作，得到新群体 P_2。

2.6 K-Means 操作

为了加速聚类算法的收敛过程，使用 K-Means 算法中的一个步骤，即 K-Means 操作(KMO)代替遗传算法中的交叉操作。K-Means 操作的具体过程为：经过选择操作，模拟退火变异操作后得到新的种群 P_2。对群体 P_2 中的某个个体，根据其现有的聚类结果计算新的聚类中心，计算方法如下：

$$z_j^* = \frac{1}{n_j} \sum_{x_m \in z_j} x_m, j = 1,2,\cdots,k \tag{6}$$

然后计算数据集中所有样本到这些新的聚类中心的距离，并将样本分配到距离最近的类中，从而获得新个体。

对父种群 P_2 中所有个体都进行 KMO 操作，形成新的种群 P_3。然后再进行下一轮遗传操作。

2.7 聚类方法的具体过程

文中聚类方法主要包含 2 层循环：外层为遗传 K-Means 算法的进化循环，内层为模拟退火算法的降温循环。

算法的具体过程(见图 3)如下：

(1)初始化控制参数：聚类个数 k，种群个数 sizepop，最大迭代次数 MAXGEN；退火初始温度 T_0，温度冷却系数 a，模拟退火内部迭代次数 N，终止温度 T_e；

(2)随机初始化 k 个聚类中心，依照聚类中心对各个样本进行聚类得到一个个体，重复 sizepop 次生成初始种群 P_0；

(3)计算种群中每个个体的适应度值：$F(S_i)$，$i = 1,2,\cdots,\text{sizepop}$；

(4)对初始种群 P_0 依次进行选择操作、模拟退火变异操作、K-Means 操作，生成新种群；

(5)重复步骤 3 和步骤 4，直到达到最大迭代次数 MAXGEN；

(6)将最后生成的种群中适应度函数值最大的个体作为聚类结果输出。

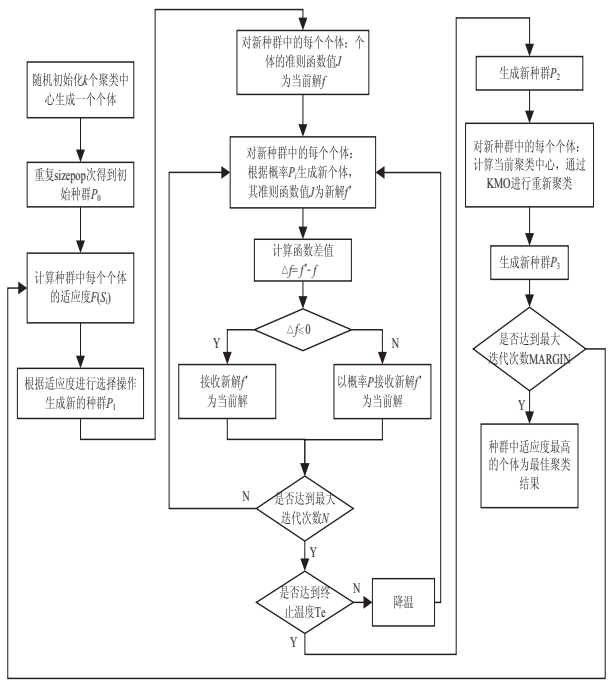

图 3 聚类方法流程

3　实验结果

对传统 K-Means 算法、文献[4]中提出的遗传 K-Means 算法以及文中提出的结合模拟退火算法的遗传 K-Means 聚类方法进行了对比实验。

实验工具为 MATLAB 软件,实验数据是来自 UCI Machine Learning Repository 的 iris 数据集和 wine 数据集。其中 iris 数据集包含 150 个数据,每个数据有 4 个属性,一共分为 3 类,每类各有 50 个数据;wine 数据集包含 178 个数据,每个数据有 13 个属性,一共分为 3 类,每类分别有 59、71、48 个数据。

分别编写 K-Means 算法、遗传 K-Means 算法以及文中的聚类方法,导入 iris 数据集和 wine 数据集进行测试。实验结果分别如表 1 和表 2 所示。

表 1　平均聚类准确度(iris 数据集)　%

运行次数	K-Means	GKAM	文中方法
1	52.67	84.66	88.53
2	64.67	84.32	90.99
3	56	84.78	89.23
4	67.22	84.59	89.58

表 2　平均聚类准确度(wine 数据集)　%

运行次数	K-Means	GKAM	文中方法
1	59.23	69.03	71.43
2	53.37	69.47	71.46
3	57.87	70.13	71.24
4	60.49	69.88	71.22

通过对表 1 和表 2 的实验结果进行分析,可以看出,与传统 K-Means 聚类算法相比,GAKM 算法的准确率有了明显的提升,而且传统 K-Means 聚类算法中会出现计算结果的浮动,每次的聚类结果都会存在较大的差异,而 GAKM 算法相对来说就比较稳定,结果基本不会发生太大的变化。

在 iris 数据集中,文中方法的平均聚类准确率最高能够达到 90.99%,最低能达到 88.53%,高于 GAKM 算法的 84.78%;在 wine 数据集中,文中方法的平均聚类准确率能达到 71.46%,同样高于 GKAM 算法中的 70.13%。通过数据对比可以发现,文中的聚类方法相对 GAKM 算法平均聚类准确度有了提升,而且能保证聚类结果的稳定。

标准 K-Means 聚类算法的计算结果受选取的初始聚类中心的影响较大,初始中心选择不当会导致结果陷入局部最优;基于遗传算法的 K-Means 聚类算法由于遗传算法自身的缺陷,容易出现早熟现象,其局部寻优能力较弱;文中提出的结合模拟退火算法的遗传 K-Means 聚类方法,充分利用模拟退火算法较强的局部寻优能力,改善遗传算法的缺陷,改善早熟现象,有效避免聚类结果陷入局部最优,最终获得的聚类结果要优于 K-Means 算法与 GKAM 算法。

4　结束语

提出一种结合遗传模拟退火算法的 K-Means 聚类方法,使用 K-Means 操作取代遗传算法的交叉操作,并引入模拟退火算法对遗传算法中的变异操作进行改进。该算法有效地解决了 K-Means 聚类算法过于依赖初始中心选择,易于陷入局部最优等问题,克服了遗传算法容易出现早熟现象以及局部搜索能力较弱的缺点。实验结果表明,该方法有效提高了 K-Means 聚类算法的聚类精度,聚类结果更加准确。

参考文献:

[1] 周　涛,陆惠玲. 数据挖掘中聚类算法研究进展[J]. 计算机工程与应用,2012,48(12):100-111.

[2] 郁启麟. K-means 算法初始聚类中心选择的优化[J]. 计算机系统应用,2017,26(5):170-174.

[3] KAPIL S, CHAWLA M, ANSARI M D. On K-means data clustering algorithm with genetic algorithm [C]//Fourth international conference on parallel, distributed and grid computing. Waknaghat, India: IEEE, 2016: 202-206.

[4] LU Bin, JU Fangyuan. An optimized genetic K-means clustering algorithm [C]//International conference on computer science & information processing. Xi'an, Shaanxi, China: IEEE, 2012: 1296-1299.

[5] 王　千,王　成,冯振元,等. K-means 聚类算法研究综述[J]. 电子设计工程,2012,20(7):21-24.

[6] 葛继科,邱玉辉,吴春明,等. 遗传算法研究综述[J]. 计算机应用研究,2008,25(10):2911-2916.

[7] 汪松泉,程家兴. 遗传算法和模拟退火算法求解 TSP 的性能分析[J]. 计算机技术与发展,2009,19(11):97-100.

[8] 康立山. 非数值并行算法(第一册):模拟退火算法[M]. 北京:科学出版社,2000:22-38.

[9] 齐　平,贾瑞玉,贾兆红,等. 用遗传模拟退火算法挖掘特征项权重的研究[J]. 计算机技术与发展,2007,17(2):143-145.

[10] LU Yi, LU Shiyong, FOTOUHI F, et al. FGKA: a fast genetic k-means clustering algorithm [C]//ACM symposium on applied computing. Nicosia, Cyprus: ACM, 2004: 622-623.

[11] 张超群,郑建国,钱　洁. 遗传算法编码方案比较[J]. 计算机应用研究,2011,28(3):819-822.

[12] 刘　英. 遗传算法中适应度函数的研究[J]. 兰州工业高等专科学校学报,2006,13(3):1-4.

[13] 张松艳. 选择算子与遗传算法的计算效率分析[J]. 宁波大学学报:理工版,2009,22(3):374-377.

[14] 凌有临,李　强,史　俊,等. 基于改进遗传算法的切削参数优化方法研究[J]. 机电一体化,2014(6):31-35.

[15] 贺永兴,杨　瑞,唐　伟,等. 基于重构变异算子遗传算法的研究[J]. 计算机技术与发展,2015,25(12):101-104.