

# 模糊遗传算法在敏感词分类优化中的应用\*

叶 蕾, 邹国奇, 肖 健  
(重庆大学 通信工程学院, 重庆 400044)

**摘 要:** 针对以往敏感词分类优化的不足, 提出一种基于模糊遗传算法的敏感词分类优化方法, 该方法把模糊逻辑理论用于遗传算法, 模拟生物进化过程和机制来求解实际的敏感词定性结构优化问题。研究表明, 对于敏感词词性以及结构的变化有很好的分类优化效果, 从而保证了整体的分类质量、快速的分类效率、鲁棒和可靠的分类性能。

**关键词:** 敏感词; 遗传算法; 模糊逻辑; 分类优化

**中图分类号:** TP391      **文献标志码:** A      **文章编号:** 1001-3695(2012)07-2549-03

**doi:** 10.3969/j.issn.1001-3695.2012.07.039

## Application of fuzzy genetic algorithm in sensitive word classification optimization

YEI Lei, ZOU Guo-qi, XIAO Jian  
(College of Communication Engineering, Chongqing University, Chongqing 400044, China)

**Abstract:** In allusion to the shortcoming existed in previous sensitive words classification optimization, this paper proposed an optimization method which put the classification on the basis of fuzzy genetic algorithm. The method applied fuzzy logic theory to genetic algorithm, simulated the process and mechanism of biological evolution to figure out the actual problem of sensitive word qualitative structure optimization. Studies show that the method has a good effect on optimization classification for the change of part-of-speech and structure of sensitive words, thereby it ensures a holistic classification quality, rapid classification efficiency, robust and reliable classification performance.

**Key words:** sensitive word; genetic algorithm; fuzzy logic; classification optimization

随着互联网的高速发展, 网络上的信息也在飞速增长, 人们获取信息也变得越来越方便、快捷。但给人们带来巨大便利的同时, 许多不良信息内容也接踵而至, 这些不良信息已成为危害极大的一个社会问题, 所以阻止不良信息的非法传播, 确保网上信息内容的安全, 已成为重要的研究课题。

以最速下降法<sup>[1,2]</sup>、牛顿法<sup>[1]</sup>和共轭方向法<sup>[1]</sup>等为代表的传统优化算法是一类应用广泛的优化算法。最速下降法以线性速度收敛, 其方法简单、计算量小、存储量小, 但它要求目标函数连续可微, 且在函数极小点附近收敛速度很慢, 因为梯度是函数的局部性质, 从局部上看是下降得快, 对于整个求最优解的过程不一定是最快的。牛顿法要求目标函数二阶连续可微, 在迭代过程中, 要计算海赛矩阵的逆矩阵, 这是比较困难的一步; 同时, 所选择初始点不能离极小点太远, 否则迭代过程可能不收敛。共轭方向法是一类介于最速下降法和牛顿法之间的算法, 它比前者的收敛速度快, 同时避免了牛顿法计算海赛矩阵的困难, 但仍需目标函数的一阶导数信息, 只能在  $n$  步之内有效,  $n$  步之后的计算也没有意义。

这类优化算法普遍要求目标函数导数连续, 具有计算复杂、串行求解等特点, 在面对离散、不连续、无导数、高度病态的

优化问题时无能为力, 其建立在局部下降的基础上, 仅能对问题作简化的近似处理, 常常无法有效地求得全局最优解。

遗传算法是一种在全局范围内产生最优解的搜索算法, 它具有自组织<sup>[3]</sup>、自适应<sup>[4,5]</sup>、自学习<sup>[6]</sup>的特性, 而且不受搜索空间限制、不需要其他辅助信息, 因此被广泛应用于解决优化问题。而敏感词分类优化就是一个不断寻找能够代表相关类别的特征集的过程。为此, 本文在研究学习基于遗传算法的文本分类与聚类<sup>[7]</sup>的基础上, 把遗传算法引入到敏感词分类优化上来用于训练敏感词分类器。

### 1 模糊遗传算法的应用实现

模糊遗传算法<sup>[8]</sup>是一种非线性的智能算法, 利用模糊机理应用于种群的选择、交叉、变异的进化, 筛选适应性更强的种群或个体。可以根据环境的变化不断调整遗传因子和代数, 动态地适应环境, 而且种群的优化可以避免局部最优, 进而整体上提高利用率。模糊遗传算法在敏感词定性结构优化处理方法中的实现步骤如下:

a) 初始化群体。根据个体的特征, 按一定的规则初始化  $N$  个作为种群。将待优化的敏感词随机分配至各个岛屿中, 成为

**收稿日期:** 2011-12-28; **修回日期:** 2012-01-28      **基金项目:** 国家科技重大专项资助项目(2011ZX002-4); 重庆市科委自然科学基金计划面上资助项目/重庆市自然科学基金资助项目(CSTC, 2009BB2041); 中央高校基本科研业务费面上资助项目(CDJZR11160006)

**作者简介:** 叶蕾(1979-), 女, 重庆云阳人, 研究生导师, 博士, 主要研究方向为无线通信、下一代移动通信系统(yelei@ccee.cqu.edu.cn); 邹国奇(1986-), 男, 硕士, 主要研究方向为移动互联网与宽带无线多媒体通信、计算机视觉及图像处理; 肖健(1986-), 男, 硕士, 主要研究方向为互联网与宽带多媒体通信、新一代移动宽带无线通信。

不同的个体,也可限定不同的岛屿中包含规定的定额敏感词。

b) 个体选择。选择操作建立在对个体的适应度评价的基础上,用来将父代中的优秀个体保存到下一代,根据个体的适应度函数值来决定应选择个体数。加入全程控制的模糊选择机制,进行选择操作,按一定适应度函数值选择出若干敏感词作为父个体。

c) 个体交叉。父个体染色体之间按规则随机选择出染色体中的一个基因,按照经模糊化规则确定的交叉概率  $P_c$ ,交换被选基因及之后的所有基因序列。本文给被选择的敏感词按其适应值的大小赋给一个繁殖概率  $P_i (0 < i < N+1)$  后,按概率大小选择出两个或者多个个体作为父个体,概率相同或者相近的两个或者多个父基个体按经模糊化确定的交叉概率  $P_c$  进行交叉,交叉的方式选择均匀交叉。

d) 个体变异。在两个子个体的染色体编码中随机选择一个基因,按照经模糊化规则确定的概率  $P_m$ ,确定是否进行变异,即1变0或0变1。将选择出来的敏感词按照经模糊化规则确定的概率  $P_m$  在不同环境下进行变异操作,变异操作的方式选择非均匀变异。

e) 优化准则判断。求种群的整体适应值并判断是否满足优化规则,如果满足,结束计算;如果不满足将从步骤 b) 重新开始,继续循环直到规定的条件。

计算经过选择、交叉、变异操作之后的敏感词适应值,在不同语境的情况下,判断其是否满足优化准则;否则,重复选择、交叉、变异操作直至达到优化规则。

## 2 具体实现方式

### 2.1 模糊遗传算法构思

图1为本方法的模糊遗传算法构思图。

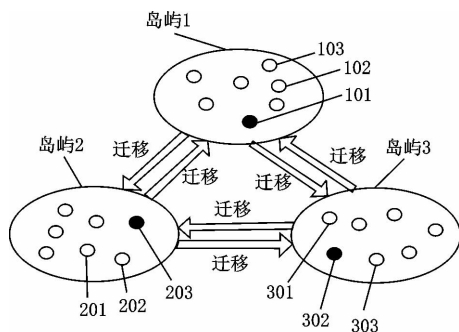


图1 算法构思

图1中存在的个体为:岛屿1中的个体101、102、103等;岛屿2中的个体201、202、203等;以及岛屿3中的个体301、302、303等。分别选择岛屿1中的个体101、岛屿2中的个体203和岛屿3中的个体302,并将个体101迁移至岛屿2,将个体203迁移至岛屿3,将个体302迁移至岛屿1。此后,在每个岛屿中进行交叉、变异、评估和判断所有个体,并重复该流程,直到获得优化值。算法具体过程如下所示:

初始化  $P_m$ 、 $P_c$ 、 $M$ 、 $G$ 、 $Tf$  等参数。产生第一代种群  $P_{op}$

do {

计算种群  $P_{op}$  中每一个体的适应度  $F(i)$ ;

初始化空种群  $newP_{op}$ ;

do { 根据适应度以比例选择算法从种群  $P_{op}$  中选出若干个体;

if (random (0,1) <  $P_c$ )

{ 对选择出来的个体按交叉概率  $P_c$  执行交叉操作; }

if (random (0,1) <  $P_m$ )

{ 对选择出来的个体按变异概率  $P_m$  执行变异操作; }

将2个新个体加入种群  $newP_{op}$  中;

} until ( N 个子代被创建 )

用  $newP_{op}$  取代  $P_{op}$ ;

} until ( 任何染色体得分超过  $Tf$ , 或繁殖代数超过  $G$ , 或满足优化规则)

其中:  $P_c$  为交叉发生的概率;  $P_m$  为变异发生的概率;  $N$  为种群规模;  $G$  为终止进化的代数;  $Tf$  为进化产生的任何一个个体的适应度函数,超过  $Tf$ , 或繁殖代数超过  $G$ , 或满足优化条件,则可以终止进化过程。

### 2.2 敏感词词性的定性

敏感词词性的定性实现过程如图2所示。

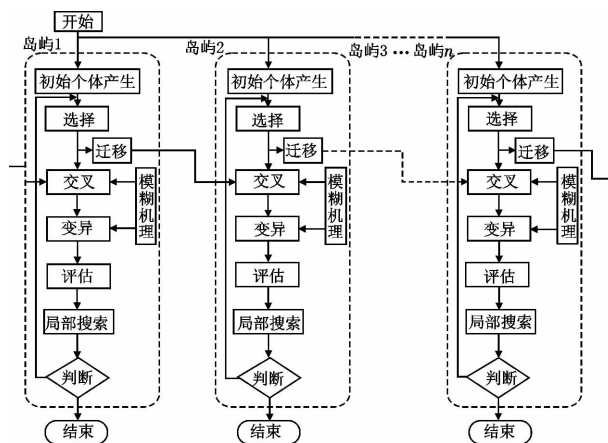


图2 运行流程

#### 2.2.1 初始化群体

根据敏感词的词性特征,按一定的规则初始化  $N$  个作为种群。将待优化的敏感词随机分配至各个岛屿中,成为不同的个体,也可限定不同的岛屿中包含规定数量的敏感词。在这里,所有的含有单独某个敏感词的文件名映射为这个敏感词表示成为种群,可以理解为所有的敏感词是所有的单独含有某个敏感词的文件名的间接代替。下文所有出现的个体均表示敏感词,且所有敏感词为包含此敏感词的文件名的映射。

#### 2.2.2 个体选择

在岛屿1中产生预定数目的个体,根据适应度函数进行适应度检测,并选择出要参与岛屿内部的遗传操作以及迁移至其他岛屿(在这种情况下,是岛屿2)的个体。选择操作的方法采用轮盘赌选择法,算法具体过程如下:

//  $p[x_i]$  表示第  $i$  个个体被选中的概率

int RWS( )

{  $m = 0$ ;

$\varepsilon = \text{random}(0,1)$ ; //  $\varepsilon$  为 0 至 1 的随机数

for ( $i = 1$ ;  $i \leq N$ ;  $i++$ )

{ // 产生的随机数在  $m \sim m + p[x_i]$  间则认为选中了  $i$

// 因此  $i$  被选中的概率是  $p[x_i]$

$m = m + p[x_i]$ ;

if ( $\varepsilon < m$ )

return  $i$ ; }

a) 计算每个个体的适应值,个体适应度函数为

$$\text{Fit}(f(x)) = \begin{cases} \frac{1-0.5}{1} \times \left[ \left| \frac{f(x)-b}{a} \right| \right]^\alpha & |f(x)-b| < a \\ 1 + \left[ \left| \frac{f(x)-b}{a} \right| \right]^\beta & |f(x)-b| \geq a \end{cases} \quad (1)$$

其中:  $a = \max \left[ 0.5, \frac{|\max \hat{f}_i - \hat{f}_i^*|}{30} \right]$ ;  $b$  的值取当前第  $i$  代中的最

小值,即  $b=f(x) \times i$ ;  $\alpha, \beta$  均为常数值,这里取  $\alpha=0.5, \beta=2$ ;

b) 计算此适应值在群体适应值总和中所占比例,表示该个体在选择过程中被选中的概率:

$$p(x_i) = f(x_i) / \sum_{n=1}^i f(x_n) \quad (2)$$

c) 通过各个个体的选择概率,计算其累计概率,第  $i$  个个体的累计概率为

$$p_x(x_i) = \sum_{n=1}^i p(x_n) \quad (3)$$

d) 产生  $0 \sim 1$  之间的随机数  $\varepsilon$  与  $p_x(x_i)$  进行比较来决定选择的个体。若  $p_x(x_{i-1}) < \varepsilon < p_x(x_i)$ , 则选择第  $i$  个个体。

复制所选择的个体,参与岛屿 1 内部的遗传操作,同时将其迁移至岛屿 2,参与岛屿 2 内的遗传操作。

### 2.2.3 个体交叉与变异

图 3 示出了个体  $A = (A_1, A_2, A_3, A_4, \dots, A_n)$  与个体  $B = (B_1, B_2, B_3, B_4, \dots, B_n)$  的交叉和个体  $A$  变异的原始说明。例如,第三基因与第四基因的交叉表示对个体  $A$  和  $B$  的每一个中的第四基因以及之后的基因进行替代。因此,从  $A$  产生新个体  $AA = (A_1, A_2, A_3, B_4, \dots, B_n)$ , 同时从  $B$  产生新个体  $BB = (B_1, B_2, B_3, A_4, \dots, A_n)$ 。此外,变异是用另一基因替代基因的操作。因此,作为用  $C_4$  ( $C_4$  不同于  $A_4$ ) 替代个体  $A$  中的第四基因  $A_4$  的变异结果,产生新个体  $AAA = (A_1, A_2, A_3, C_4, A_5, \dots, A_n)$ 。

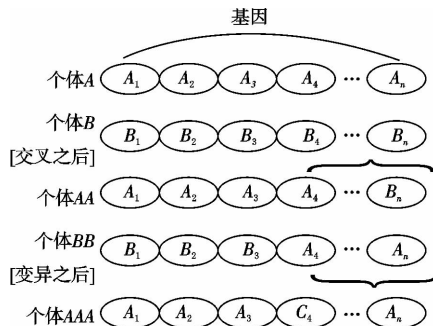


图3 个体交叉与变异说明

在岛屿 2 中,从岛屿 1 迁移来的个体与岛屿 2 中经选择操作筛选的个体进行交叉,组合成为语境下的新个体。这里所产生的新个体为:经过词间组合,包含两个或者多个敏感词的文件名映射为两个或者多个敏感词表示成为一个新个体,可以理解为所产生的新个体是所有的包含两个或者多个敏感词的文件名的间接代替。采用基于模糊机理的交叉方式,不仅考虑当前这一代种群的适值情况(横向比较),而且将参考若干代过程中适值的变化情况(纵向分析),很容易判断目前局部最优解和全局最优解。对不同的个体采用不同的交叉概率,对于适应度值远高于群体平均适应度值的个体,对应于较低的交叉概率,使它得以保护进入满足优化规则的行列,而相对较低适应度的个体,给予较高的交叉概率,增加其参与词间组合的几率。

方法如下:

$$p_c = \begin{cases} p_{c1} - \frac{(p_{c1} - 0.6)(f_{\max} - f)}{f_{\max} - \bar{f}} & f < \bar{f} \\ p_{c1} & f \geq \bar{f} \end{cases} \quad (4)$$

其中:  $p_{c1} = 0.9$ ;  $f_{\max}$  为群体中的最大适应度值;  $\bar{f}$  为每代群体的平均适应度值;  $f$  是进入交叉的个体的适应度值。

在岛屿 2 中,对岛中的所有个体按周期性模糊变异的方式进行变异,产生变异的新个体。这里的新个体可以理解为,某个敏感词在不同的文件名(区别于作为种群时的映射文件名)里出现,并产生了不同的语境意义,或者两个或者多个敏感词

在不同的文件名(区别于交叉操作后所映射的文件名)里出现,并产生了不同的语境意义。在相对长的时期内维持比较低的变异概率,较完全地搜索群体所覆盖的空间域,并在特定的时间段出现高变异概率的情况,使算法能够跳出局部搜索。公式为

$$p_m(t) = \frac{\alpha[t - (k + 0.5)T_c]^2}{T_c^2} \quad (5)$$

其中:  $t$  为进化代数;  $k$  为周期数;  $T_c$  为变化周期,一般为期望的最小进化代数;  $\alpha$  为概率调整系数,为变异概率  $p_m$  的极值情况。

在岛屿 2 中也产生预定数目的初始个体;然后经选择操作选择要迁移至岛屿 3 的个体;接着,复制个体,并将其迁移至岛屿 3。在岛屿 3 中,将从岛屿 2 迁移来的个体与岛屿 3 中存在的个体进行交叉或变异。此外,也在岛屿 3 中进一步执行相似操作,以此类推直至岛屿  $n$ 。

针对每个岛屿,计算岛屿中所有个体的评估值,所有个体包括父辈个体以及交叉、变异操作产生的新个体。这里新个体指的是词间组合以及发生变异而成为的新个体。评估值计算方法沿用 2.2.2 节中的方法。

在计算每个个体的评估值之后,针对岛屿中被选择操作所筛选的每个个体,执行局部区域搜索。局部区域搜索表示通过略微改变参数对通过应用模糊遗传算法而获得的结果执行搜索,该结果是对于特定个体的特定参数是否存在更好结果有关的。这里所指的参数是个体参与词间组合的几率,参与词间组合的几率与个体适应度的高低成正比。

经过应用模糊遗传算法而获得的新个体会出现重复现象,因为产生的新个体可能是两个或多个父辈个体交叉组合出现在了多个不同语境意义下的文件名内,或产生的新个体可能是父辈个体经变异后出现在了多个不同语境意义下的文件名内。针对选择的个体执行局部搜索,在这里表示搜索是否有任何更好的评估值。对于两个或者多个父辈个体交叉组合出现在了多个不同语境意义下的文件名内所映射出的个体执行评估,在上或下单元的评估值比人为选定单元的评估值好的情况下,保留较好的评估值。此外,人为选定单元的值比上或下单元的评估值都好的情况下,保留人为选定单元的评估值,并完成这一代中的局部区域搜索。当完成局部区域搜索时,在每个岛中选择个体并进行逐一判断,以完成一代模糊遗传算法。

### 3 优化准则

本研究方法中所实施的优化条件由敏感词所映射的文件名集合产生的模糊规则来决策。

在每个岛屿中选择完成局部区域搜索后的个体评估值作为输入的样本数据,给定一个标准评估值,其中有两项输入数据:  $e$  和  $\Delta e$ 。  $e$  为评估值误差,  $\Delta e$  为评估值误差的变化,输出为  $\mu$ ,  $\mu$  为个体优化后所属的模糊子集。

对于  $e$  和  $\Delta e$  将其划分为五个模糊子集,  $NB$  (大负),  $NS$  (小负),  $Z$  (零),  $PS$  (小正),  $PB$  (大正)。  $e$  和  $\Delta e$  规范在区间  $[-1, 1]$ , 则五个模糊子集定义为区间  $[-1, 1]$  上的任意模糊子集。

敏感词所映射的文件名集合产生的模糊规则,通过人为审核大量文件名,确定其属性并进行模糊隶属区域划分,规则构成如下,其中第  $i$  个规则为: 如果  $e \in E_i$  和  $\Delta e \in \Delta E_i$ , 那么

(下转第 2578 页)

$$u^* = \left[ \sum_{i=1}^n u_{E_i} \wedge u_{\Delta E_i} \times c_3(i) \right] / \sum_{i=1}^n u_{E_i} \wedge u_{\Delta E_i} \quad (6)$$

其中:  $E_i$  为  $e$  的隶属函数;  $\Delta E_i$  为  $\Delta e$  的隶属函数;  $n$  为模糊规则的数目;  $u^*$  为总输出,  $c_3(i)$  为输出隶属函数值为 1 的中心位置。

在这种情况下, 判断评估值是否满足优化规则, 如果判断为未完成, 则流程返回前一选择操作步骤, 并重复相同处理。然后, 当判断个体的评估值已满足优化规则时, 则认为该评估值是优化的, 并结束对敏感词后续的词性优化控制。

对每个岛屿执行该一系列的循环, 并在每个岛屿上执行个体迁移。在本实施方式的优化处理方法中, 执行以下优化处理: 对岛屿中个体数目均为 100 的 10 个岛屿进行评估, 即对于一代, 个体数目是 1 000。

如上所述, 根据本优化处理方法, 可以较容易地获得对敏感词的词性定性优化, 该值具有与目标特性更加接近的特性。

## 4 结束语

针对于敏感词分类优化上的组合优化问题和复杂的调度决策问题, 克服现有技术的缺点, 提出一种适于大规模问题的具有智能特征的基于模糊遗传算法的敏感词分类优化方法, 该方法在一定程度上改进了敏感词的定性分类。对于本领域技术人员, 还可以根据实际工程问题的不同以及本研究方法的核心理念设计和构造自己的多目标动态优化<sup>[9]</sup>控制算法, 在实际工程中达到最好的效果, 从而更好地解决实际工程的优化问题。本文以敏感词定性结构优化为例, 研究了模糊遗传算法的

应用, 如对本方法进行适当的调整后, 它同样适用于其他类似分类优化问题的寻优求解。

## 参考文献:

- [1] 袁亚湘, 孙文瑜. 最优化理论与方法[M]. 北京: 科学出版社, 2006.
- [2] 吴锋, 李秀梅, 朱旭辉, 等. 最速下降法的若干重要改进[J]. 广西大学学报: 自然科学版, 2010, 35(4): 596-600.
- [3] QI Zhi-dong, PENG Chun-ming. An improved fuzzy genetic algorithm with fuzzy adjusted crossover and mutation probabilities[C]//Proc of the 3rd International Conference on Advanced Computer Theory and Engineering. [S. l.]: IEEE Press, 2010: 581-585.
- [4] 张忠诚, 王淳, 李静. 基于模糊控制器的自适应遗传算法[J]. 兵工自动化, 2006, 25(4): 66-67.
- [5] ZHANG Lei, ZHANG Bo-feng, YU Jian-feng. Improvement of adaptive genetic algorithm and its application in examination timetabling optimization problem[C]//Proc of International Conference on Computer Science and Service System. [S. l.]: IEEE Press, 2011: 1326-1329.
- [6] ZHANG Huai-xiang, ZHANG Bo, WANG Feng. Automatic fuzzy rules generation using fuzzy genetic algorithm[C]//Proc of the 6th International Conference on Fuzzy Systems and Knowledge Discovery. [S. l.]: IEEE Press, 2009: 107-112.
- [7] 戴文华. 基于遗传算法的文本分类及聚类研究[M]. 北京: 科学出版社, 2008.
- [8] 张筱磊, 赵培庆. 模糊遗传算法研究与实现[D]. 合肥: 中国科学技术大学, 2005.
- [9] 公茂焦, 巢李成, 杨咚咚, 等. 进化多目标优化算法研究[J]. 软件学报, 2009, 20(2): 271-289.