

利用膜粒子群优化和信息熵的医学文本特征选择

豆增发, 高琳

(西安电子科技大学计算机学院, 710071, 西安)

摘要: 针对生物医学文本挖掘中有效特征选择困难的问题, 提出了一种新的基于膜粒子群优化和信息熵的文本特征选择方法. 该方法以文本信息熵总和为目标函数, 以膜系统的层次结构作为框架, 以膜系统的消息传递机制作为进化方向, 以粒子群优化进化作为进化规则. 作为进化规则的粒子群优化算法, 分别以局部搜索速率和全局搜索速率搜索得到文本信息熵最大的解, 并在不同膜区域之间传递搜索结果, 直到膜区域之间的消息传递结束或者达到限定的迭代次数. 实验结果表明, 利用提出的方法对医学文本特征进行选择后, 对其进行分类, 能使分类精确度和召回率分别提高2%和3%左右.

关键词: 膜系统; 粒子群优化; 生物医学文本; 特征选择; 信息熵

中图分类号: TP301 **文献标志码:** A **文章编号:** 0253-987X(2012)04-0045-07

Feature Selection of Biomedical Literature by Membrane Particle Swarm Optimizer and Information Entropy

DOU Zengfa, GAO Lin

(School of Computer Science, Xidian University, Xi'an 710071, China)

Abstract: To select the effective features of biomedical literature, a novel particle swarm optimizer based on membrane system is proposed. This scheme utilizes total text information entropy as the fitness function, employs the structure of membrane system as the frame, sets mechanism of message transferring as the propagation direction, and considers particle swarm optimizer as the iteration rule. The particle swarm optimizer searches the best solution maximizing the total text information entropy at local and global searching rates simultaneously, and exchanges solutions between membrane regions till the communication between regions stops or iteration reaches the limit times. The experiment shows that feature selection by the scheme improves the precession and heightens the classification recall of biomedical literature by 2% and 3%.

Keywords: membrane system; particle swarm optimizer; biomedical literature; feature selection; information entropy

生物医学文本中包含大量的蛋白质名称、基因名称、疾病名称、药物名称等各种专有名词, 这些专有名词构成生物医学本文的重要特征. 在对生物医学文本进行信息提取、自动分类、聚类分析等运算时, 随着要处理的生物医学文本数量的增长, 所包含

的蛋白质、基因、疾病、药物等相关的文本特征数量也越来越多, 这使得数据挖掘算法的运算量呈指数级增长, 从而导致算法性能下降. 同时, 在这些海量的生物医学文本特征中, 如果对所有特征不加区别同等对待, 不区分各个特征的重要性, 而过多地引入

收稿日期: 2011-07-03. 作者简介: 豆增发(1979—), 男, 博士生; 高琳(通信作者), 女, 教授, 博士生导师. 基金项目: 国家自然科学基金重点资助项目(60933009); 高等学校博士学科点专项科研基金资助项目(200807010013); 国家自然科学基金资助项目(60970065).

网络出版时间: 2012-03-09

网络出版地址: <http://www.cnki.net/kcms/detail/61.1069.T.20120309.1137.003.html>

与主题无关的特征,往往对挖掘工作造成不必要的干扰,严重影响算法的精确性,这种现象就是典型的噪声干扰。因此,生物医学文本特征的选择在生物信息学的数据挖掘领域是一个值得研究的课题。

本文提出一种新的特征选择算法,称为膜粒子群优化方法。该方法的主要特征是利用了膜系统的层次结构、运算规则以及消息传递机制这3个特点,将粒子群优化作为各个分层区域的子算法部署到子区域中。在这些区域中,最内层区域负责搜索全局最优解,而其他外层区域负责搜索局部最优解,与此相对应,将粒子群优化的搜索速率分解为全局搜索速率和局部搜索速率,将前者应用到最内层区域,将后者应用到其他外层区域。按照膜系统的消息传递机制,所有外部区域将最优解传递给相邻内部区域,内部区域将最差解传递给相邻外部区域,最内层区域向相邻外部区域传递最差解。当各个区域之间的最优解传递在一段时间内停止,或者算法迭代次数达到限定次数,算法收敛,取最内层区域的最优解为最终解。各个子算法以文本信息熵总和为目标函数,利用膜粒子群优化计算特征权重系数,最后选择那些权重系数较大的特征。将本文提出的方法应用到生物医学文本的特征选择上,实验结果表明,经本文方法做特征选择后,可以大大提高生物医学分类的精确度和召回率。

1 文本的向量空间模型和信息熵

文本特征一般用向量空间模型来表示。向量空间模型是一种用向量表示文本的代数模型,每个维度对应一个属性项。如果该属性项在文档中出现,它的值就是非0值,如果该属性项不在文档中出现,则该值为0。本文采用向量空间模型来表示生物医学文本,所有的属性项用特定单词在文档中的出现次数来表示,即每个独立的属性项 w_i 对应一个属性 t_i 。因此,一个生物医学文本可以用如下公式表示

$$\mathbf{c} = \{t_1, t_2, \dots, t_n\}^T \quad (1)$$

式中: $\mathbf{c}$ 代表一个生物医学文本; t_i 表示属性项 w_i 在文档 $\mathbf{c}$ 中出现的频率; n 为属性总个数,即

$$t_i = \frac{w_i}{n} \quad (2)$$

如果用向量空间模型表示文档中出现的所有的词汇,会造成维度灾难,这势必导致计算量的急剧膨胀,并且产生严重的噪声干扰。因此,对每个特征的重要性需要区分,适当放大重要特征的作用,减弱或者删除那些次要特征,以减少噪声干扰和计算量。据

此,我们给每个特征增加一个权重系数 λ ,新的文本特征向量为

$$\mathbf{c} = \{\lambda_1 t_1, \lambda_2 t_2, \dots, \lambda_n t_n\}^T \quad (3)$$

式中: λ_i 表示特征 t_i 的重要性。

为了选择出那些重要的属性,删减次要和冗余属性,我们选取文本的信息熵作为目标函数,即计算文本的权重系数向量 $\{\lambda_1, \lambda_2, \dots, \lambda_i, \dots, \lambda_n\}$,使文本的信息熵总和为最大。

信息熵是信息论中用于度量信息量的一个概念。香农信息熵定义如下

$$H(p) = - \sum_{i=1}^n p(x_i) \log(p(x_i)) \quad (4)$$

式中: $p(x_i)$ 表示特征 x_i 的概率值。

本文为了计算生物医学文本的信息熵,增加了每个特征的权重系数,因此将单个文本的信息熵定义为

$$H(t) = - \sum_{i=1}^n \lambda_i t_i \log(t_i) \quad (5)$$

2 膜系统介绍

膜计算基于DNA计算中生命细胞的分层结构时,将分层结构以处理化合物的方式抽象出来,形成一个膜系统(又称为 P 系统)^[1-2]。一般地,一个度为 m 的膜系统可表示为如下的一个多元组^[3-4]

$$\Pi = (V, T, C, \mu, \omega_1, \dots, \omega_m, (R_1, \rho_1), \dots, (R_m, \rho_m)) \quad (6)$$

式中: V 是字母表,其元素称为对象; T 是输出字母表; C 是催化剂,对象在进化过程中不发生变化,也不产生新字符,但某些进化规则必须有它的参与才能执行; μ 是包含 m 个膜的膜结构; ω_i 是膜结构 μ 中的区域 i 含有独享的多重集; (R_i, ρ_i) 是膜结构 μ 中的规则集合,其中 R_i 为规则, ρ_i 为规则的优先级。

膜系统由多个相互包含的膜组成,最外层的膜称为表层膜,最内层的膜称为内层膜,不含有其他的膜称为基本膜。每个膜所包围的部分称为区域,区域内包含着一个对象或多个对象以及相应的进化规则。膜系统从初始态开始,运用膜结构中的进化规则完成计算。一个典型的膜系统如图1所示。

图1所示的膜系统由4个膜组成,分别标号为膜1、膜2、膜3和膜4,各个膜之间形成4个区域,分别为区域1、区域2、区域3和区域4。区域1包含3个对象 a, b, c ,区域4有一个对象 e 。区域1包含运算规则 $c \rightarrow ca | c$ 和 $c \rightarrow c3$,区域3包含运算规则 $c \rightarrow c4$,区域4包含运算规则 $e \rightarrow ee | \rightarrow c$ 。其中: $c \rightarrow ca, e \rightarrow$

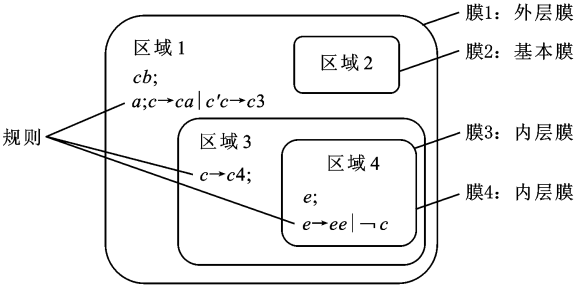


图 1 膜系统的层次结构

ee 表示 c 进化为 ca 、 e 进化为 ee ； $|$ 表示进化条件； $|c$ 表示 c 作为催化剂条件导致进化发生； $|\neg c$ 表示 c 作为抑制剂条件导致进化终止； $c\rightarrow c3$ 表示将 c 送入区域 3。图 1 所示的膜系统的计算过程如下。

- (1) 区域 1 中存在催化剂 c ，规则 $c\rightarrow ca | c$ 表示在催化剂 c 的作用下将对象 a 送入到区域 2 中，直到对象不可用为止；规则 $c\rightarrow c3$ 表示将对象 c 送入区域 3。
- (2) 当对象 c 到达区域 3 后，将会根据其中的规则 $c\rightarrow c4$ 继续到达区域 4。
- (3) 当对象 c 到达区域 4 后，其中对象 e 的产生将停止，因为 c 是规则 $e\rightarrow ee | \neg c$ 中的抑制剂。

3 基于膜粒子群优化的特征选择方法

粒子群优化算法^[5-6]是进化算法的一种，它通过初始化一个粒子群(x_i)并使其按一定规律更新、进化，以搜索最优值。每个粒子的速率和位置根据下式更新

$$v_{id(t+1)} = \omega v_{id(t)} + c_1 \text{rand}() (P_{id} - x_{id(t)}) + c_2 \text{rand}() (P_g - x_{id(t)}) \quad (7)$$

$$x_{id(t+1)} = x_{id(t)} + v_{id(t+1)} \quad (8)$$

式中： $v_{id(t)}$ 表示粒子 i 在维度 d 上的速率； $x_{id(t)}$ 表示粒子 i 在维度 d 上的当前位置； $x_{id(t+1)}$ 表示粒子 i 在维度 d 上的下一个位置； ω 是惯性因子； c_1 是局部学习因子； c_2 是全局学习因子； P_{id} 表示粒子 i 在维度 d 上迄今所得到的最优值； P_g 表示所有粒子在整个搜索空间中迄今所得到的最优值； $\text{rand}()$ 是值域为 $[0,1]$ 的随机函数。

原始的粒子群优化算法在搜索初期很容易陷入局部收敛，而在搜索后期，会很容易在最佳位置陷入无限迭代。为了避免出现这两种问题，一般采用控制惯性因子、局部学习因子和全局学习因子的方法，这些方法一般依靠经验来获得，需要做大量的实验才会得到一组比较合理的值。本文基于膜系统框架，提

出了一种完全不同的粒子群优化算法，即膜粒子群优化算法。膜粒子群优化算法将原来的速率计算公式拆分为两部分，即局部搜索速率和全局搜索速率。局部搜索速率定义为

$$v_{l(t+1)} = \omega_1 v_{l_t} + \phi_1 \text{rand}() (P_1 - x_t) \quad (9)$$

式中： $v_{l(t+1)}$ 表示下一代中粒子群的局部搜索速率； v_{l_t} 为当前迭代的局部搜索速率； ω_1 为局部搜索的惯性因子； ϕ_1 为局部学习因子； P_1 为粒子迄今为止所得到的局部最优值。

全局搜索速率定义为

$$v_{g(t+1)} = \omega_2 v_{g_t} + \phi_2 \text{rand}() (P_g - x_t) \quad (10)$$

式中： $v_{g(t+1)}$ 为全局搜索速率； ω_2 为全局搜索的惯性因子； v_{g_t} 为当前迭代的全局搜索速率； ϕ_2 为全局学习因子。

粒子按下式进化

$$x_{t+1} = x_t + v_{t+1} \quad (11)$$

如图 2 所示，膜粒子群优化算法由多个独立的分区、子算法，以及每个分区的计算解组成^[7]。

- (1) 有层层嵌套的膜分隔开来组成的很多分区。
- (2) 在每个区域，有粒子群优化子算法和该算法获得的试探性解。其中，最内层区域用全局搜索速率迭代，其他外层区域用局部搜索速率迭代。
- (3) 邻接区域之间的解传递机制，即所有外层区域将最优解向相邻内层区域传递，将最差解向相邻外层区域传递。通过不断重复这个过程，最优解将汇集到最内部区域。当满足终止条件时，算法终止最内层区域的最佳解并作为算法的最终输出。

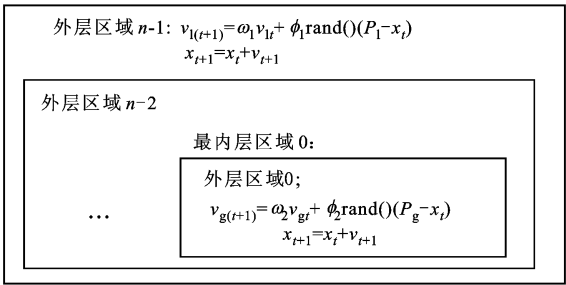


图 2 膜粒子群优化算法的层次结构

基于膜粒子群优化算法的文本特征选择方法的具体步骤如下。

第 1 步：设有 m 个测试文本，文本特征向量的维度为 n ，给定这些测试文本的信息熵总和(膜粒子群优化的目标函数)为

$$H = - \sum_{j=1}^m \left(\sum_{i=1}^n \lambda_i t_i \log(t_i) \right) \quad (12)$$

第 2 步：设定膜系统有 n 个区域，最内层为区域

0,外层为区域 1 到区域 $n-1$. 给定区域 0 一个随机解,给区域 1 到 $n-1$ 各两个随机解. 设权重系数向量 λ 为维度为 n 的粒子群

$$\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_i, \dots, \lambda_n\}$$

设局部搜索惯性因子和全局搜索惯性因子为 0.9,局部学习因子和全局学习因子为 1. 设 2 维数组 $P[d]$ 用来保存每个区域的最优解和最差解,即 $P[i,0]$ 为最差解, $P[i,1]$ 为最优解.

第 3 步:设最大迭代次数为 k ,重复以下步骤.
(1)每个区域同时用各自的子算法修改现有解. 最内层区域 0 用以下公式搜索最优解

$$\begin{aligned} v_{t+1} &= \omega_2 v_t + \phi_2 \text{rand}() (P_g - x_t) \\ x_{t+1} &= x_t + v_{t+1} \end{aligned}$$

外层区域 1 到 $d-1$ 用以下公式搜索最优解

$$\begin{aligned} v_{t+1} &= \omega_1 v_t + \phi_1 \text{rand}() (P_1 - x_t) \\ x_{t+1} &= x_t + v_{t+1} \end{aligned}$$

(2)每个区域将自己最优解发往邻接内层区域,将最差解发往邻接外层区域. 如果从外层区域收到的最优解优于自己的最优解,则用外层最优解代替自己的最优解. 如果内层区域发来的最差解比自己的最差解更差,则用内层区域的最差解代替自己的最差解. 最内层区域 0 仅将最差解发往外层 1,不发送自己的最优解.

(3)如果所有区域之间没有发生最优解的替代,转第 4 步,否则转第 3 步中的(1). 如果算法迭代次数超过了 k 次,则转向第 4 步.

第 4 步:输出最内层区域 0 的最优解为算法最终解.

第 5 步:对 λ 的各个分量从大到小排序,删除小于阈值 β 的分量以及对应的文本特征,输出剩余的

分量为最终权重系数向量 λ .

4 实验结果

4.1 膜粒子群优化算法和其他算法的性能比较

将膜粒子群优化算法和传统粒子群优化算法、GA 算法做了比较,结果表明,膜粒子群优化算法比其他算法收敛速度快,而且精确度更高. 传统粒子群优化的维度取 30,膜粒子群优化的膜区域个数取 30,每种算法各运行 50 次,得出平均迭代次数和平均收敛值,结果如表 1 所示.

4.2 文本特征选择比较

4.2.1 实验数据 为了评价膜粒子群优化算法做特征选择的有效性^[8-10],本文从 MEDLINE 中选择了 12 类生物医学文本数据作为分类数据,表 2 显示了所选择的文档类别以及数量.

4.2.2 实验结果 以信息熵为目标函数,用传统粒子群优化算法和膜粒子群优化算法分别对表 1 所示的文本数据做特征选择^[11],然后利用 K-NN,贝叶斯分类器、支持向量机算法^[12]对表 1 所示文本数据进行分类测试,来评价特征选择算法的性能. 每种算法分类两次,第一次用传统粒子群优化算法和信息熵做特征选择后进行分类,第二次用膜粒子群优化算法和信息熵做特征选择后进行分类,然后将两次的分类结果进行比较. 结果表明,经膜粒子群优化算法和信息熵方法做文本特征选择后,K-NN、贝叶斯分类器、支持向量机算法的分类精确度和召回率比传统粒子群优化算法和信息熵方法要高. 我们用精确度 p 、召回率 r 、 F 值对分类结果进行分析比较

表 1 膜粒子群优化算法与其他算法的性能比较

基准函数	算法	平均迭代次数	平均收敛值
$f_1(x) = \sum_{i=1}^n x_i^2$	GA	356	0.002 34
	传统粒子群优化	210	0.002 45
	膜粒子群优化	150	0.000 33
$f_2(x) = \sum_{i=1}^n (x_i^2 - 10\cos(2\pi x_i) + 10)$	GA	421	0.012 31
	传统粒子群优化	223	0.003 91
	膜粒子群优化	177	0.002 34
$f_3(x) = \sum_{i=1}^n (100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2)$	GA	351	0.012 12
	传统粒子群优化	198	0.009 00
	膜粒子群优化	140	0.005 62

表 2 用于测试可拓分类器的实验文本

文本集	ID	文本数量
Chickenpox	Ghk	1 283
Raynaud Disease	RD	1 153
Insomnia	Ins	1 352
Jaundice	Jn	1 486
Hepatitis B	Hpt	1 815
Hay Fever	HF	1 632
Kidney Calculi	KS	1 071
Impotence	Imp	1 092
Age-related Macular Degeneration	AMD	1 277
Migraine	Mg	1 174
Otitis	Ot	1 233
Osteoporosis	Ost	1 754

$$p = m_c / m_{all} \tag{13}$$

$$r = m_i / m_{all} \tag{14}$$

$$F = 2pr / (p + r) \tag{15}$$

式中: m_c 表示正确分类的文本个数; m_i 表示可以被明确分类的文本个数; m_{all} 表示参与实验的总的文本个数.

实验结果表明,经本文方法对原始数据做了特

征选择后,K-NN、贝叶斯分类器以及支持向量机算法的分类性能大大提高.表 3~表 5 分别展示了对 12 类生物医学文本进行分类时特征选择前后的精确度、召回率以及 F 值.

在实验中我们注意到,阈值 β 和膜系统区域数对特征选择有至关重要的影响.阈值 β 如果取得过小,将无法完全过滤噪声干扰,如果取得过大,会影响分类精确度.阈值 β 与分类精确度之间的关系如图 3 所示.一般来说,膜系统区域数越多,特征的权重向量的取值越优,但是过多的膜系统区域数会增加计算量,而且到了一定的数量,膜系统区域数对权重向量的影响会降低.本文为了计算简单,分别取膜系统区域数为 40、80、160.图 4 所示为膜系统区域数对文本分类的影响.

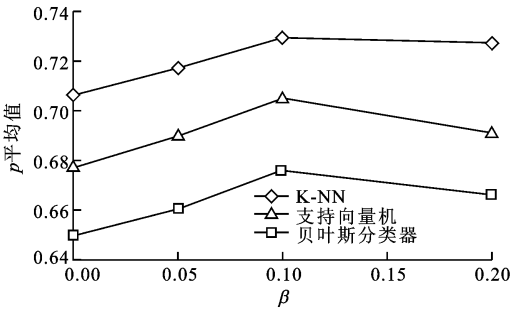


图 3 分类精确度与阈值 β 的关系

表 3 K-NN、贝叶斯分类器以及支持向量机算法的精确度比较

文本集	p					
	传统粒子群优化和信息熵方法			膜粒子群优化和信息熵方法		
	K-NN	贝叶斯分类器	支持向量机	K-NN	贝叶斯分类器	支持向量机
Chickenpox	0.712	0.662	0.712	0.734	0.676	0.731
Raynaud Disease	0.712	0.632	0.702	0.735	0.665	0.733
Insomnia	0.702	0.622	0.692	0.712	0.644	0.714
Jaundice	0.692	0.692	0.642	0.7	0.719	0.671
Hepatitis B	0.682	0.632	0.732	0.699	0.691	0.742
Hay Fever	0.702	0.612	0.632	0.771	0.633	0.666
Kidney Calculi	0.712	0.622	0.672	0.735	0.634	0.701
Impotence	0.732	0.672	0.682	0.743	0.681	0.711
Age-related Macular Degeneration	0.642	0.662	0.622	0.676	0.675	0.654
Migraine	0.752	0.602	0.692	0.763	0.641	0.721
Otitis	0.712	0.702	0.712	0.722	0.733	0.738
Osteoporosis	0.732	0.692	0.642	0.764	0.721	0.686

表 4 K-NN、贝叶斯分类器以及支持向量机算法的召回率比较

文本集	<i>r</i>					
	传统粒子群优化和信息熵方法			膜粒子群优化和信息熵方法		
	K-NN	贝叶斯分类器	支持向量机	K-NN	贝叶斯分类器	支持向量机
Chickenpox	0.713	0.663	0.713	0.722	0.677	0.725
Raynaud Disease	0.713	0.623	0.703	0.715	0.633	0.733
Insomnia	0.703	0.633	0.683	0.703	0.649	0.714
Jaundice	0.693	0.683	0.643	0.699	0.702	0.702
Hepatitis B	0.683	0.633	0.733	0.698	0.692	0.783
Hay Fever	0.703	0.613	0.633	0.711	0.632	0.698
Kidney Calculi	0.713	0.623	0.673	0.722	0.699	0.701
Impotence	0.723	0.663	0.683	0.739	0.683	0.711
Age-related Macular Degeneration	0.683	0.653	0.623	0.671	0.703	0.697
Migraine	0.713	0.623	0.693	0.749	0.701	0.738
Otitis	0.713	0.703	0.713	0.752	0.754	0.786
Osteoporosis	0.743	0.693	0.663	0.766	0.711	0.712

表 5 平均精确度、平均召回率以及平均 F 值的比较

参数	传统粒子群优化和信息熵方法			膜粒子群优化和信息熵方法		
	K-NN	贝叶斯分类器	支持向量机	K-NN	贝叶斯分类器	支持向量机
<i>p</i> 平均值	0.707	0.650	0.678	0.730	0.676	0.706
<i>r</i> 平均值	0.708	0.651	0.680	0.721	0.686	0.725
<i>F</i> 平均值	0.708	0.650	0.679	0.725	0.681	0.715

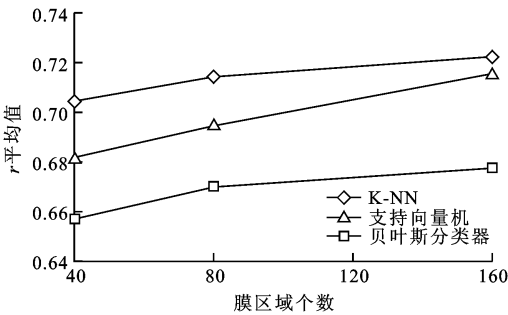


图 4 分类精确度与膜系统区域个数的关系

5 结 论

针对生物医学文本处理中有效特征选择困难的问题,提出了一种新的基于膜粒子群优化和信息熵的文本特征选择方法. 该方法利用膜系统的分层结

构,将传统粒子群优化的搜索分解为局部搜索和全局搜索. 膜系统的所有外层区域采用局部搜索速率,搜索局部最优解,最内层区域采用全局搜索速率,搜索全局最优解. 所有外部区域将最优解传递给相邻内部区域,内部区域将最差解传递给相邻外部区域,最内区域向相邻外部区域传递最差解. 当各个区域之间的解传递在一段时间内停止,或者算法迭代次数达到限定次数,算法收敛,取最内层区域的解为最优解. 本文以文本信息熵总和为目标函数,利用膜粒子群优化计算特征选择系数,最后选择那些系数较大的特征. 实验结果表明,利用本文提出的膜粒子群优化方法对生物文本进行特征选择,使文本分类的精确度和召回率较传统粒子群优化算法大大提高.

在膜粒子群优化中,阈值 β 和膜系统区域个数在实践中根据经验确定. 阈值 β 的选取关系到冗余

特征的过滤,取值过小会引入冗余特征,而选取过大会过滤掉部分有效特征.膜系统区域数关系到特征选择算法的计算量和性能,一般来说,膜系统区域数越多,算法的性能相对较好,但是计算量也会增大.所以,要根据实际问题,综合考虑区域数和计算量问题.如何能够快速确定阈值 β 和合适的膜系统区域个数,还需在下一步工作中进一步研究.

参考文献:

- [1] PAUN G. Computing with membranes [J]. Journal of Computer and System Sciences, 2000, 61(1): 108-143.
- [2] PAUN G, ROZENBERG G. A guide to membrane computing [J]. Theoretical Computer Science, 2002, 287(1): 73-100.
- [3] 张葛祥,潘林强.自然计算的新分支:膜计算[J]. 计算机学报, 2010, 33(2): 208-214.
ZHANG Gexiang, PAN Linqiang. A survey of membrane computing as a new branch of natural computing [J]. Chinese Journal of Computers, 2010, 33(2): 208-214.
- [4] 黄亮. 膜计算优化方法研究[D]. 杭州: 浙江大学, 2007.
- [5] KENNEDY J, EBERHART R. Particle swarm optimization[C]// Proceedings of the Fourth IEEE International Conference on Neural Networks. Piscataway, NJ, USA: IEEE, 1995: 1942-1948.
- [6] EBERHART R, KENNEDY J. A new optimizer using particle swarm theory [C]// Proceedings of the Sixth International Symposium on Micro Machine and Human Science. Piscataway, NJ, USA: IEEE, 1995: 39-43.
- [7] NISHIDA T Y. An application of P system: a new algorithm for np-complete optimization problems [C]// Proceedings of the 8th World Multi-Conference on Systemics, Cybematics and Informatics. Orlando, FL, USA: Int. Inst. of Informatics and Systemics, 2004: 109-112.
- [8] 朱虎明,焦李成. 基于免疫记忆克隆的特征选择[J]. 西安交通大学学报, 2008, 42(6): 679-682.
ZHU Huming, JIAO Licheng. Feature selection via immune memory clone [J]. Journal of Xi'an Jiaotong University, 2008, 42(6): 679-682.
- [9] YANG Yiming, PEDERSON J O. A comparative study on feature selection in text categorization [C]//

Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 1997: 412-42.

- [10] WONG J W T, KAN W K, YOUNG G. Action: automatic classification for full-text documents [J]. SIGIR Forum 30, 1996, 30(1): 26-41.
- [11] SCOTT S, MATWIN S. Feature engineering for text classification [C]// Proceedings of 16th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 1999: 379-388.
- [12] TAIRA H, HARUNO M. Feature selection in SVM text categorization [C]// Proceedings of 16th Conference of the American Association for Artificial Intelligence. Piscataway, NJ, USA: IEEE, 1999: 480-486.

[本刊相关文献链接]

- 应用粒子群优化-条件随机场的文本生物实体识别. 2010, 44(12): 38-42.
- 采用矩谱分析和支持向量机的地磁导航基准图构建方法. 2010, 44(10): 47-51.
- 柔性测量臂运动学建模及参数标定方法. 2010, 44(8): 122-126.
- 结合分割逼近和粒子群法的燃气轮机叶片轮廓度误差计算. 2010, 44(7): 42-45.
- 利用粒子群优化的人脸特征提取识别算法. 2010, 44(4): 48-51.
- 系统辨识的粒子群优化方法. 2009, 43(2): 116-120.
- 一种医学 X 射线图像动态范围扩展方法. 2008, 42(12): 1468-1471.
- 一种基于种群熵的混沌小世界优化算法. 2008, 42(9): 1137-1141.
- CT 模体投影解析计算及其在射束硬化校正中的应用. 2008, 42(6): 664-668.
- 离散粒子群优化在垂直分层空时系统检测中的应用. 2008, 42(2): 166-170.
- 结合对比度塔和克隆选择加权 Wedgelets 的图像融合. 2007, 41(10): 1170-1174.
- 医学可视化网格平台的设计与实现. 2007, 41(8): 1000-1002.
- 发射调频信号提高医学超声图像的信号噪声比. 2007, 41(6): 741-745.
- 改进粒子群算法及其在超导电缆参数优化中的应用. 2007, 41(2): 219-222.
- 粒子群优化算法的惯性权值递减策略研究. 2006, 40(1): 53-56.

(编辑 杜秀杰)