DOI:10.13657/j.cnki.gxkxyxb.2012.01.001

广西科学院学报
Journal of Guangxi Academy of Sciences

2012,28(1):1～3
Vol. 28, No. 1 February 2012

文本特征抽取中基于基因集编码的遗传退火算法*

The Application of Genetic Annealing Algorithm Based on Gene-Set in the Feature Selection of Text Classification

符保龙
FU Bao-long

(柳州职业技术学院,广西柳州 545006)
(Liuzhou Vocational Technological College, Liuzhou, Guangxi, 545006, China)

摘要:采用基因集的形式对传统遗传算法的编码方式进行改进,再引入模拟退火的思想,提出一种基于基因集编码的遗传退火算法的文本特征抽取方法(GSGAA),并与遗传算法(GA)和模拟退火 GA 算法(SA-GA)进行比较实验。结果表明,GSGAA 算法用于文本分类的特征抽取所得出结果的正确率和执行时间都比采用单基因进行编码的 GA 算法和 GA-SA 算法好,具有一定的应用价值。

关键词:文本分类 特征抽取 基因集 遗传算法 退火算法
中图法分类号:TP301.6 **文献标识码**:A **文章编号**:1002-7378(2012)01-0001-03

Abstract: By usage of gene-set, the encoding of traditional genetic algorithm is improved. The improved encoding with the introduction of simulated annealing, a feature selection of text classification through the genetic annealing algorithm based on genetic-set (GSGAA) is illustrated. Compared with genetic algorithm (GA) and simulated annealing genetic algorithm (SAGA), the results show that GSGAA can significantly improve the accuracy and shorten the execution time, which indicate the application value of GSGAA.

Key words: text classification, feature selection, gene-set, genetic algorithm, annealing algorithm

面对互联网上海量的信息,想要迅速有效地提取出用户所需信息是当前信息科学和技术领域面临的一大挑战。文本分类技术是实现这一目标的有效途径,从而成为当前文本挖掘研究领域的一个重点。文本分类的任务是对需要进行分类的文本,根据其内容和属性自动地把它归入预先定义好的类别中。文本分类技术通过分析待分类文本,提取该文本的特征,再通过和原有预定义特征进行对比,从而将待分类对象进行归类。特征选择是文本自动分类中的一项关键技术,但是将待分类文本表示成特征向量时,所得到的文本特征集非常庞大,有必要对特征进行降维处理。通过特征选择一方面可以有效降低文本向量的特征数,另一方面通过选择合适的特征可以提高系统的分类精度。

目前,常用的特征选择法有文档频率(Document Frequency, DF)、信息增益(Information Gain, IG)和互信息(Mutual Information, MI)等。此外,基于生物演化技术的算法也被用于文本的特征选择。遗传算法(Genetic Algorithm, GA)是一种模拟生物进化的智能算法,具有全局搜索功能,是处理大规模数据项目集的有效方法。文献[1～3]已经采用遗传算法对文本的特征进行抽取,效果良好。虽然传统 GA 算法能快速、有效地进行全局优化搜索,但是在对染色体进行编码时无论是采用二进制还是十进制进行编码,都只是基于单基因形式。对于大型数据集来说,采用单基因形式进行编码,无疑增加了算法的运算长度和运算时间。而且采用单基因形式编码容易造成算法多样性保持能力不足,容易出现早熟和陷入局部最优的缺点。模拟退火算法

收稿日期:2011-04-02
修回日期:2011-05-13
作者简介:符保龙(1978-),男,副教授,主要从事数据挖掘、演化计算研究。
*广西教育厅科研项目(NO:200911LX486;201106LX745)资助。

(Simmualted Annealing,SA)是一种基于蒙特卡罗迭代求解的启发式随机搜索算法[4]。SA 算法与初始值无关,已在理论上被证明是一种以概率收敛于全局最优解的全局优化算法,把 SA 算法和 GA 算法相结合,能有效保证得到全局最优解。本文对传统 GA 算法的编码机制进行了改进,采用基因集的形式进行编码,并在此基础上引入模拟退火算法的思想,提出了一种基于基因集编码的遗传模拟退火算法(Gene-Set Genetic Annealing Algorithm,简称 GSGAA)的文本特征抽取方法。

1 重要技术指标的定义

要正确地对文本进行分类,必须先将文档表示成计算机能够处理的形式。目前,常用的向量空间模型(Vector Space Mode,即 VSM)[5,6]是简便有效的文本表示模式之一。在 VSM 模型中,每一对象模型化为空间中的点,两对象间的差异由多维空间中两点间的距离表示。为了更有效地分析和解决文本的特征抽取问题,需要对一些重要技术指标进行定义。

定义 1 文本的内容特征用它所含有的字、词、词组或短语等基本语言单位来表示,这些基本的语言单位被统称为文本的项,即文本可以用项集(Term List)表示为 $D(T_1,T_2,\cdots,T_n)$,其中 T_k 是项,$1\leqslant k\leqslant n$。

定义 2 项的权重 W_k 表示项 T_k 在文本 D 中的重要程度,即 $D(W(T_1),W(T_2),\cdots,W(T_n))$,$1\leqslant k\leqslant n$。

项的权重 W_k 一般采用 $TF-IDF$ 向量表示:

$$W_k(T_k)=\frac{tf_k(T_k)\times\lg(N/n_k)}{\sqrt{\sum_{i=1}^{n}tf_i(T_k)\times\lg(N/n_i))^2}},$$

其中,$tf_k(T_k)$ 为项 T_k 在文档 D 中出现的词频。N 为所有文档的数目,n_k 为出现了项 T_k 的文档数目。

定义 3 给定一个文本 $D(W(T_1),W(T_2),\cdots,W(T_n))$,当暂时不考虑 T_k 在文本中出现的先后顺序,并要求 T_k 互异时,可把 $T_1,T_2,\cdots,T_n$ 看作是一个 n 维的坐标,而 $W(T_1),W(T_2),\cdots,W(T_n)$ 就是 n 维坐标的对应值,m 个训练文档则可表示为矩阵 $A=(W(T_k))_{m\times n}$,该矩阵 A 称为向量空间模型。

2 GSGAA 算法描述

2.1 染色体编码及初始群体

为了能更好的发挥遗传退火算法的优势,保持基因的多样性,引入基因集的概念。

定义 4 在遗传编码中,以 m 个基因为单位参与遗传算法的各种操作,把 m 个基因组成的整体称为基因集。

以二进制编码来说,假设 n 为基因总数,m 为基因集的长度。设 $n=2^a$,$m=2^b$(a、b为自然数,且 $a\geqslant b$),则一个染色体里共有 2^{a-b} 个基因集组成。此时,参与遗传操作的长度可以等同地认为是整个基因集的长度,这比采用单基因编码方式的长度小很多。我们将每个基因表示成一种特征组合,基因集长度为 3。随机产生 N 个基因集为初始群体。

2.2 适应度函数

适应度函数的选取对遗传算法的收敛速度影响以及能否找到最优解至关重要。考虑到文本分类的复杂性,我们直接以项的权重计算公式作为适应度函数,即 $fit(T_k)=\lambda\times W_k(T_k)$,$\lambda$ 为常数。

2.3 选择操作

选择的本质是优胜劣汰,由适应度大小来决定。在当前群体 $F=\{x_1,x_2,\cdots,x_n\}$ 中,个体 x_i 被选取的概率为

$$P(x_i)=\frac{f(x_i)}{\sum_{i=1}^{n}f(x_i)},$$

其中,$f(x_i)$ 是个体 x_i 的适应度函数值。

2.4 交叉操作

交叉是指把两个父代个体的部分结构加以替换重组而生成新个体的操作,是产生新个体的主要方法。不同于传统 GA 算法的交叉算子,我们定义的交叉算子如下:在一定的交叉概率 P_c 下,参与交叉操作的两个染色体的父类必须是一样的,根据 P_c 从染色体中随机的选择基因集的边界来进行交叉操作,也就意味着本次交叉操作是以基因集为单位来进行的,如图 1 所示。

2.5 退火变异

变异操作能够有效改善遗传算法的局部搜索能力,保持种群多样性。设计的退火变异算子如下:(1)采用高频变异对群体进行变异操作。高频变异的计算公式[7]为

$$a_i'=a_i+\rho*\frac{f(a_i)}{\max(f(a_i))}*N(0,1),$$

其中,a_i、a_i' 分别表示变异前后的基因集的位值,$N(0,1)$ 是均值为 0、方差为 1 的正态分布的随机数,ρ 为变异常数,用来控制子群中个体进行局部搜索的范围,其大小根据种群中的个体数量选取,一般 $\rho=0.1$。

(2)计算变异前个体 x_i 和变异后个体 x_i' 的适应度分别为 $f(x_i')$,$f(x_i)$,如果 $\min\{1,\exp(-(f(x_i')-f(x_i))/T_k)\}>\text{random}[0,1]$,则按照退火接受准则,接受新解;否则,放弃变异后的个体。T_k 为第 k 次进化的温度。

(a)

010	110	101	001	000	011
101	001	111	110	100	000

(b)

101	001	101	001	000	011
010	110	111	110	100	000

图 1 交叉操作结果

(a)交叉前的染色体,(b)交叉后的染色体。

2.6 算法描述

(1)初始化参数:群体规模 N,交叉概率 P_c,初始温度 T_0,进化代数 $k=0$,基因集的长度 m,对基因按基因集形式进行编码,随机产生初始种群 S_0。

(2)评价个体的适应度。

(3)对群体分别进行选择、交叉和退火变异操作。

(4)当 $T_k=0$ 时,退火过程自然结束;否则:$T_{k+1}=\lambda\times T_k$,$0<\lambda<1$,$k=k+1$,并将步骤(3)得到的新群体,按适应度进行排序,截取前面的 N 个子代个体作为新的父代群体,算法返回到步骤(4)。

3 GSGAA 算法实验

为了验证本文所提出的观点和方法的有效性,在 Pentium(R) 4 CPU 2.80G Hz、1G 内存的 PC 机上,采用 MATLAB 7.0 进行编程,分别对 GA 算法、GA-SA 算法和 GSGAA 算法进行对比实验。实验用到的数据集全部来自(www.163.com)共 6000 个网页的样本集(2007 年 1 月到 2010 年 2 月的新闻语料),涵盖经济、军事、体育、汽车、生活、娱乐六大类。各类样本数均为 1000 个。取其中的 3/4 文档作为训练集,余下的 1/4 作为测试集。实验中设定的参数为:种群规模 $N=100$,交叉概率 $P_c=0.85$,初始温度 $T_0=100$。用分类正确率和执行时间作为衡量算法性能的标准,实验结果如图 2 和图 3 所示。

从图 2 中可以看出,GSGAA 算法执行时间比 GA 算法和 GA-SA 算法有很大的优势。这主要是因为 GSGAA 算法是以基因集的形式进行编码参与整个遗传退火操作,此时,参与运算的染色体长度要比传统 GA 算法和 GA-SA 算法短,所以执行时间就减少。从图 3 可以看出 GSGAA 算法在不同进化代数下所获得的分类准确率高于其它两种算法。因为 GSGAA 算法采用基因集进行编码,在变异的过程中比采用单基因的编码方式获得更大的种群多样性,有利于搜索到全局最优值。GSGAA 算法用于文本分类的特征抽取所得出结果的正确率和执行时间都比采用单基因进行编码的 GA 算法和 GA-SA 算法好,具有一定的应用价值。

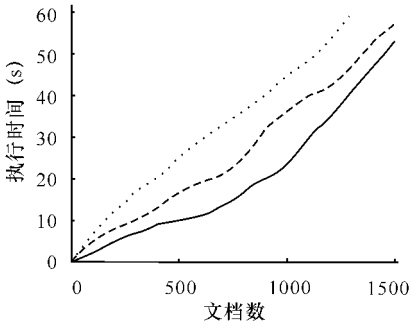

图 2 文档数和训练时间的变化情况

……:GA 算法;……:GA-SA 算法;—:GSGAA 算法。

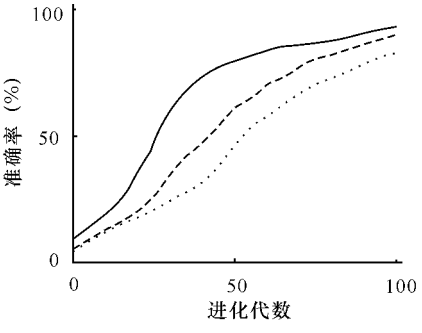

图 3 不同进化代数下各算法的分类准确率

……:GA 算法;……:GA-SA 算法;—:GSGAA 算法。

参考文献:

[1] Yang J, Honavar V. Feature subset selection using a genetic algorithm[J]. IEEE Trans on Intelligent Systems, 2003, 18(2): 44-49.

[2] Chakraborty B. Genetic algorithm with fuzzy fitness func-tion for feature selection[C]. Proc fo 2002 IEEE Int'lSymp on Industrial Electronics, 2005: 315-319.

[3] 刘勇国,李学明,张伟,等. 基于遗传算法的特征子集选择[J]. 计算机工程,2003,29(6):19-21.

[4] 王雪梅,王义和. 模拟退火算法与遗传算法的结合[J]. 计算机学报,1997,20(4):381-384.

[5] Doloca Adrian. Feature selection for texture analysis using genetic algorithms[J]. International Journal of Computer Mathem Atics, 2000, 74: 279-292.

[6] 高洁,吉根林. 文本分类技术研究[J]. 计算机应用研究,2004,7(30):28-30.

[7] Scinvivas M, Patnaik L M. Adaptive probabilities of crossover and mutation in genetic algorithm[J]. IEEE Trans SMC, 1994, 24(4): 656-666.

(责任编辑:尹 闯)