基于混合克隆量子遗传策略的文本特征选择方法 *

符保龙
(柳州职业技术学院, 广西 柳州 545006)

摘 要: 引入向量约简率和分类准确率的度量标准,采用量子比特对遗传算法进行编码,结合克隆算子,提出一种基于混合克隆量子遗传策略的文本特征选择方法。实验结果显示,该方法能有效地降低文本特征向量的维度,所提取的特征向量子集能有效提高文本分类的精度。

关键词: 特征选择; 文本分类; 量子遗传; 克隆算法
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2012)02-0485-02
doi:10.3969/j.issn.1001-3695.2012.02.022

Text feature selection method based on hybrid clone quantum genetic strategy

FU Bao-long
(*Liuzhou Vocational Technological College*, *Liuzhou Guangxi* 545006, *China*)

Abstract: The metrics of vector reduction rate and classification accuracy, and to use of the qubits encoded on the genetic algorithm, combined with the cloning operator, this paper proposed a strategy based on hybrid genetic quantum cloning text feature selection method. Experimental results show that the method can effectively reduce the dimension of feature vector text, set of extracted features can improve the quantum accuracy of text classification.

Key words: feature selection; text classification; quantum genetic; clonal algorithm

0 引言

文本分类技术是当前数据挖掘研究的一个热点。作为信息处理领域的一个重要分支,文本分类技术在信息发现中有重要的应用。其主要的任务是分析待分类文本的内容和属性,把它们归入到预定义的类别中。文本特征空间具有高维度和文档特征向量稀疏性的特点,为了提高文本分类的精度和效率,有限的特征集是必需的。文本特征选择是文本分类的一项关键技术,它是指从大量的、高维度的文本特征空间 D 中,选择一个数量尽可能最小、又能准确和有效表达原内容的特征子集 d 的过程。

近年来,作为文本分类领域的热点问题,国内外很多学者对文本的特征选择算法作了大量研究,提出了多种方法,如文档频数(document frequency, DF)、信息增益(information gain, IG)、CHI 统计量(chi-squared)等。这些特征选择方法都各有优点,无论是基于统计分析还是基于机器学习的方法,其基本原理都是构造一个权重函数,从待分类的特征集中选取每一个特征向量进行独立评估,按评估分值的高低排序,选取预定数目的向量作为待分类的特征子集。其实该过程是一个全局搜索的过程。遗传算法(genetic algorithm, GA)是一种具有全局搜索功能的智能优化算法。文献[1,2]把遗传算法引入到文本特征子集的选取中,有效地解决了文本特征高维度和向量稀疏性的问题;文献[3]设计了一种自适应的遗传算法用于文本特征抽取;文献[4]给出了一种自适应遗传算法与模拟退火算法相结合的特征选择方法,该方法针对 GA 在局部搜索能力方面的不足,引入热力学的退火平衡模型,以利于寻找全局最优解;文献[5]提出了一种基于量子遗传算法(quantum genetic algorithm, QGA)的文本特征选择方法,该方法用量子位对特征向量进行编码,通过量子门旋转更新来完成进化搜索;文献[6]把免疫克隆算子引入到文本特征选择中,在一定程度上优化了文本的特征提取,提高了文本的分类效果。本文在文献[16]的启发下,采用量子比特进行编码,引入克隆选择策略,提出一种基于混合克隆量子遗传策略(hybrid clonal quantum genetic strategy, HCQGS)的文本特征选择方法。

1 问题的定义

向量空间模型(vector space model, VSM)是使用较多的、最简便的且效果较好的文本表示方法之一。在该模型中,每一个文档对象被转换为空间中的点,两点之间的空间距离就表示两个对象之间的差异。为了能更有效地分析和解决问题,本文对一些重要技术指标进行如下定义:

定义 1 特征项。文档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特征项,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 特征项的权重。特征项的权重 W_k 表示项 T_k 在

收稿日期: 2011-08-04; **修回日期**: 2011-09-15 **基金项目**: 广西教育厅科研项目(200911LX486,201106LX745)
作者简介: 符保龙(1978-),男(壮族),副教授,硕士,主要研究方向为数据挖掘、演化计算(gxsony@163.com).

文本 D 中的重要程度，即 $D(W(T_1),W(T_2),\cdots,W(T_n))$，$1\leqslant k\leqslant n$。

特征项的权重 W_k 一般采用 TF-IDF 向量表示法计算，其定义如下：

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

其中：$tf_k(T_k)$ 为词条 T_k 在文档 D 中出现的频率；N 为所有文档的数目；n_k 为出现了词条 T_k 的文档数；分母为归一化因子。

定义 3 向量空间模型(VSM)。给定一个文本 $D(W(T_1),W(T_2),\cdots,W(T_n))$，当暂不考虑 T_k 在文本中出现的先后顺序，并要求 T_k 互异时，可把 $T_1,T_2,\cdots,T_n$ 看做是一个 n 维的坐标，而 $W(T_1),W(T_2),\cdots,W(T_n)$ 就是 n 维坐标所对应的值，m 个训练文档则可表示为矩阵 $\boldsymbol{A}=(W(T_k))_{m\times n}$，称 $\boldsymbol{A}$ 为向量空间模型。

2 HCQGS 算法的具体应用

2.1 量子编码

本文采用量子比特对染色体进行编码。一个量子比特可表示为 $[\alpha,\beta]^{\mathrm{T}}$，其状态可用公式 $|\Phi\rangle=\alpha|0\rangle+\beta|1\rangle$ 来表示。其中 α 和 β 为复数且满足条件 $|\alpha|^2+|\beta|^2=1$，$|\alpha|^2$ 和 $|\beta|^2$ 分别表示量子比特处于状态 $|0\rangle$ 和状态 $|1\rangle$ 的概率。因此，一个量子比特不仅可以表示 0 或 1 这两种状态，还可以同时表示它们之间的任意叠加态。一个具有 m 个量子比特的染色体可表示为 $\begin{bmatrix}\alpha_1,\alpha_2\cdots\alpha_m\\ \beta_1,\beta_1\cdots\beta_m\end{bmatrix}$，因此，一个采用 k 个量子比特编码的染色体可同时表示 2^k 个状态[7]。

相比传统的二进制编码和实数编码，采用量子比特对染色体进行编码。由于一个量子比特能够表示多个状态，因而可以有效减少染色体的数量，实现在种群规模较小的情况下，保持基因个体的多样性。随着 $|\alpha|^2$、$|\beta|^2$ 趋于 0 或 1，此时种群多样性趋向单态，算法收敛。

2.2 亲和度函数

所选择的特征子集应该在一定程序上能代表该文本类别，具体而言，其特征项的权重越大越能代表该文本类别。因此，本文以特征项的平均权重公式作为亲和度函数，即

$$\mathrm{fit}(T_i)=\sqrt[n]{\prod_{i=1}^{n}W_i(T_i)} \tag{2}$$

2.3 量子操作

在量子遗传算法中，用量子门来改变量子比特的状态。此时，将已经构造的量子门作用于量子比特的叠加态或其纠缠态的基态，使它们相互影响、干涉，以此来改变各量子比特基态的概率幅。因为量子状态 $|\Phi\rangle=\alpha|0\rangle+\beta|1\rangle$ 满足归一化条件 $|\alpha|^2+|\beta|^2=1$，因此量子门作用后的状态也必须满足这个条件。本文采用量子旋转门，即

$$U(\Delta\theta_i)\begin{bmatrix}\cos\Delta\theta_i & -\sin\Delta\theta_i\\ \sin\Delta\theta_i & \cos\Delta\theta_i\end{bmatrix} \tag{3}$$

其中：$\Delta\theta_i$ 为量子门的旋转角，其取值可通过查表获得。

2.4 克隆

在 HCQGS 算法中，为了防止进化早熟，采用克隆算子[8]扩大群体规模，增大搜索空间，保持解的多样性。设 $a_i(k)$ 为抗体，$A(k)$ 为抗体群，且 $a_i(k)\in A(k)$。把 $a_i(k)$ 按规则 $\mathrm{Int}\left[\lambda\cdot\frac{\mathrm{fit}(a_i(k))}{\sum_{j=0}^{n-1}f(a_j(k))}\right]$ 进行克隆，对 $A(k)$ 中的所有抗体执行相同的操作，从而得到新的抗体群 $A'(k)$。其中：λ 为克隆系数，$\mathrm{Int}\lceil*\rceil$ 表示上取整。根据克隆规则可知，抗体亲和度越大，则抗体的克隆规模也随之增大[9]。

2.5 HCQGS 算法的基本流程

HCQGS 算法的基本步骤如下：

a) 令 $t=0$，初始化种群 $Q(t)=\{q_1^t,q_2^t,\cdots,q_n^t\}$。其中：$n$ 是种群规模；$q_i^t=\begin{bmatrix}\alpha_{i1}^t & \alpha_{i2}^t & \cdots & \alpha_{im}^t\\ \beta_{i1}^t & \beta_{i2}^t & \cdots & \beta_{im}^t\end{bmatrix}$；$m$ 为量子比特数目。初始化时，所有的 $q_i^t(i=1,2,\cdots,m)$ 均取值为 $1\sqrt{2}$，它代表量子位以等几率的形式线性叠加。

b) 克隆 $Q(t)$ 生成 $Q'(t)$。

c) 根据 $Q'(t)$ 中各量子位的概率幅构造出观测态的集合 $P(t)$。

d) 评价 $P(t)$ 中的适应值，把最优解存入 Best(t) 中，判断终止条件。若满足，则算法结束；否则，转下一步执行。

e) 根据式(3)更新 $P(t)$。

f) $t=t+1$，算法转至 b) 继续执行，直到算法结束。

3 实验结果及其分析

3.1 实验参数

本文实验所采用的硬件环境为 Pentium® 4 2.40 GHz CPU 和 2 GB 内存的 PC 机，软件环境为 MATLAB 7.0。分别对 GA[1]、QGA[5] 和 HCQGS 算法进行对比实验。实验用到的数据集全部来自新浪网(www.sina.com.cn)共 660 个网页的样本集，涵盖经济、军事、体育、汽车、生活、娱乐六大类。文本过滤阶段，采用本文提出的特征提取算法对文本进行去低频词、停用词的基本处理，保留了 12 033 个特征关键词，通过人工分类，将所有的文档分为六大类，如表 1 所示。

表 1 文档分类

	经济	军事	体育	汽车	生活	娱乐
篇数	251	117	93	59	83	57

实验的参数设定为：种群规模 $N=150$，交叉率为 0.85，变异率为 0.1。为了更好地衡量本文提出的算法，引入了向量约简率和分类准确率两个评价指标[5]。

定义 4 向量约简率 E。约简后特征向量维数 N' 和原始特征维数 N 的比值，即

$$E=\frac{N'}{N}\times100\% \tag{4}$$

定义 5 分类准确率 F。分类正确的文本数 Ψ' 和测试集中总文本数 Ψ 的比值，即

$$F=\frac{\Psi'}{\Psi}\times100\% \tag{5}$$

3.2 结果分析

实验结果如图 1、2 所示。从图 1 和 2 中可以明显看出，本文提出的 HCQGS 算法在向量约简率和分类准确率上都优于传统的 GA 和 QGA 算法。这主要是因为采用量子比特编码的方式下，一个量子比特能表示多个量子状 (下转第 496 页)

4　结束语

由实验结果可知,本算法在优选项目邻居个数和用户邻居个数的情况下,可将 MAE 值始终控制在[0.80,0.82]内。改进的协同过滤算法不仅降低了数据稀疏性对推荐系统的影响,提高了系统效率及可扩展性,而且还能准确找到用户的兴趣最近邻,减小推荐误差。

为进一步提高本算法的推荐质量,在下一步研究中考虑增加项目类别的相似度计算和解决冷启动问题。

参考文献:

[1] 杨芳, 潘一飞, 李杰. 一种改进的协同过滤推荐算法[J]. 河北工业大学学报, 2010,39(3):82-87.

[2] 李聪, 梁昌勇, 董珂. 基于项目类别相似性的协同过滤推荐算法[J]. 合肥工业大学学报:自然科学版, 2008,31(3):360-363.

[3] 李春, 朱珍敏, 高晓芳. 基于邻居决策的协同过滤推荐算法[J]. 计算机工程, 2010,36(13):34-36.

[4] 周军锋, 汤显, 郭景峰. 一种优化的协同过滤推荐算法[J]. 计算机研究与发展, 2004,41(10):1842-1847.

[5] 邓爱林, 朱扬勇, 施伯乐. 基于项目评分预测的协同过滤算法推荐算法[J]. 软件学报, 2003,14(9):1621-1628.

[6] SARWAR B, KARYPIS G, KONSTAN J, *et al*. Item-based collaborative filtering recommendation algorithm[C]//Proc of the 10 th Intering World Wide Web Conference. New York: ACM Press, 2001: 285-295.

[7] 张海鹏, 离烈彪, 李仙,等. 基于项目分类预测的协同过滤推荐算法[J]. 情报学报, 2009,19(6):218-223.

[8] 嵇晓声, 刘宴兵, 罗来明. 协同过滤中基于用户兴趣度的相似性度量方法[J]. 计算机应用, 2010,30(10):2618-2610.

[9] 李大学, 谢名亮, 赵学斌. 结合项目类别信息的协同过滤推荐算法[J]. 重庆邮电大学学报:自然科学版,2010,22(6):823-827.

[10] GONG Song-jie, YE Hong-wu. Joining user clustering and item based collaborative filtering in personalized recommendation services[C]//Proc of International Conference on Industrial and Information Systems. Washington DC: IEEE Computer Society,2009:149-151.

[11] 黄裕洋, 金远平. 一种综合用户和项目因素的协同过滤推荐算法[J]. 东南大学学报:自然科学版,2010,40(5):917-921.

[12] THIESSON B, MEEK C, CHICKERING D M, *et al*. Learning mixtures of DAG models[C]//Proc of the 14 th Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 1998:504-513.

[13] SARWAR B M, KARYPIS G, KONSTAN J A, *et al*. Application of dimensionality reduction in recommender system: a case study[C]//Proc of ACM WebKDD 2000 Workshop. [S.l.]: ACM Press,2000: 264-268.

[14] RESNICK P, IACOVOU N, SUCHAK M, *et al*. GroupLens: an open architecture for collaborative filtering of Netnews[C]//Proc of ACM Conference on Computer Supported Cooperative Work. New York: ACM Press,1994:175-186.

[15] LEICHT E A, HOLME P, NEWMAN M E J. Vertex similarity in networks[J]. Physical Review E,2006,73(2):026120.

[16] SHEN Lei, ZHOU Yi-ming. A new user similarity measure for collaborative filtering algorithm[C]//Proc of the 2nd International Conference on Computer Modeling and Simulation. Washington, DC:IEEE Computer Society, 2010:375-379.

[17] 贺银慧, 陈端兵, 陈勇,等. 一种结合共同邻居和用户评分信息的相似度算法[J]. 计算机科学, 2010, 37(9):184-186, 204.

[18] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. IEEE Trans on Knowledge and Data Engineering, 2005,17(6):734-749.

(上接第 486 页)态的叠加,在较小的种群规模情况下,仍然保持种群中个体的多样性。同时,HCQGS 算法采用克隆算子,在进化中的过程中,根据亲合度的大小,在候选解的附近产生一个变异解的群体,扩大解的搜索范围,有助于防止进化早熟和陷于局部极小值的问题。从图中还可以看出,采用量子比特进行编码的 QGA 要优于传统的 GA 算法。

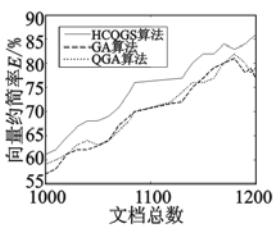

图1　向量约简率对比

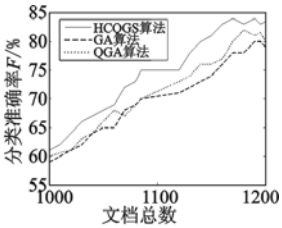

图2　分类准确率对比

4　结束语

文本的特征抽取是一个重要的研究方向。如何把表示文本的高维数特征向量转换为合适的特征子集是实现文本正确分类的关键。本文采用量子比特进行编码,引入人工免疫中的克隆选择策略,提出一种基于混合克隆量子遗传退火策略的文本特征选择方法。实验结果表明,该方法比传统的 GA 和 QGA 算法更能有效地降低文本特征向量的维度,所提取的特征子集能有效提高文本分类的精度和效率,具有一定的应用前景。

参考文献:

[1] DOLOCA A. Feature selection for texture analysis using genetic algorithms[J]. International Journal of Computer Mathematics, 2000,74(3):279-292.

[2] 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.

[3] 赵丽娜,刘培玉,朱振方. 自适应遗传算法在特征选择中的改进及应用[J]. 计算机工程与应用,2009,45(7):39-41.

[4] 张昊,陶然,李志勇,等. 基于自适应模拟退火遗传算法的特征选择方法[J]. 兵工学报,2009,30(1):82-85.

[5] 邱烨,刘培玉. 基于量子遗传算法的文本特征选择方法研究[J]. 计算机工程与应用,2008,44(25):140-142.

[6] 陈绯,郑华. 一种免疫克隆特征选择算法在文本分类中的应用[J]. 计算机工程与科学,2009,31(9):119-121.

[7] XIONG Yan, CHEN Huan-huan, MIAO Fu-you, *et al*. A quantum genetic algorithm to solve combinatorial optimization problem[J]. Acta Electronica Sinica,2004,32(11):1855-1858.

[8] SHANG Rong-hua, JIAO Li-cheng. An immune clonal algorithm for dynamic multi-objective optimization[J]. Journal of Software, 2007,18(11):2700-2711.

[9] 杨咚咚,焦李成,公茂果,等. 求解偏好多目标优化的克隆选择算法[J]. 软件学报,2010,21(1):14-33.