使用 PGA 的特征选择方法

马春华[1],朱颢东[2]
MA Chun-hua[1],ZHU Hao-dong[2]

1.绥化学院 计算机科学与技术系,黑龙江 绥化 152061
2.中国科学院 成都计算机应用研究所,成都 610041
1.Department of Computer Science and Technology,Suihua College,Suihua,Heilongjiang 152061,China
2.Chengdu Institute of Computer Application,Chinese Academy of Sciences,Chengdu 610041,China
E-mail:shzlm009@sina.com

MA Chun-hua,ZHU Hao-dong.Feature selection method applicated PGA.Computer Engineering and Applications,2009,45(22):107-110.

Abstract: Feature selection is one of the key steps in text classification system.However,most of existing feature selection methods are serial and are inefficient timely to be applied to Chinese massive text data sets,so it is a hotspot how to improve efficiency of feature selection by means of parallel strategy.It detailedly designs a Parallel Genetic Algorithm(PGA) which is used to select features.The algorithm uses genetic algorithm to search features and calculates fitness of feature subsets in multiple computing nodes at the same time,so can acquire quickly feature subsets which are more representative.Experimental results show that the method is effective.

Key words: text categorization;feature selection;Genetic Algorithm(GA);parallel strategy

摘 要:特征选择是文本分类系统的核心步骤之一。然而现有的特征选择方法都是串行化的,应用于中文海量文本数据时时间效率较低,因此利用并行策略来提高特征选择的效率,已经成为研究的热点。详细设计了一个用于特征选择的并行遗传算法,该算法采用遗传算法搜索特征,利用并行策略评价特征子集,即将种群中个体的适应度计算并行在多个计算节点上同时进行,从而较快地获得较具代表性的特征子集。实验结果表明该方法是有效的。

关键词:文本分类;特征选择;遗传算法;并行策略
DOI:10.3778/j.issn.1002-8331.2009.22.035 **文章编号**:1002-8331(2009)22-0107-04 **文献标识码**:A **中图分类号**:TP301

文本分类系统通常采用特征集来表示待学习的文档,这使得特征向量的维数非常大,有时会达到数十万维。如此高维的特征对后续的分类过程未必全是重要的、有益的,而且高维的特征可能会大大增加分类的计算开销,使整个处理过程的效率非常低下,而且可能产生与小得多的特征子集相似的分类结果[1]。因此,必须对文档的特征向量进一步净化处理,在保持原文含义的基础上,找出最能反映文本内容,又比较简洁的特征向量。特征选择就是为解决上述问题而产生的一个关键步骤。这个步骤不仅能够解决上述问题,而且它在一定程度上还能够消除噪声词语,使文本之间的相似度更加准确,即提高语义上相关的文本之间的相似度,同时降低语义上不相关的文本之间的相似度[2]。

现存诸多特征选择方法都是基于英文的,例如 WF、DF、IG、MI、CHI 等[3-7]。由于中文与英文的文本分类问题具有相当大的差别,体现在原始特征空间的维数更大,文章表示更加稀疏,词性变化更加灵活等多个方面。在英文文本分类中表现良好的特征抽取方法未必适合中文文本分类,并且这些特征抽取方法都是串行化的,遇到中文海量数据集时效率较低,因此利用并行策略来研究适合于中文的特征选择方法十分有必要。

详细设计了一个用于特征选择的并行遗传算法,该方法采用遗传算法搜索特征,利用并行策略评价特征子集,即将种群中个体的适应度计算并行在多个计算节点上同时进行,从而较快地获得较具代表性的特征子集。实验结果表明该方法是有效的。

1 粗糙集基本理论

这里要用到一些粗糙集知识,下面只简单介绍紧密相关的粗糙集知识,具体请参阅文献[8]。

定义 1[8] 设信息系统 $S=<U,C\cup D,V,f>$,$U/C=\{X_1,X_2,\cdots,X_n\}$,$U/D=\{Y_1,Y_2,\cdots,Y_m\}$,则决策集 D 关于特征集 C 的支持度定义为:

基金项目:四川省科技计划项目(No.2008GZ0003);四川省科技厅科技攻关项目(No.07GG006—014)。
作者简介:马春华(1971-),女,副教授,研究方向为计算机控制理论与控制工程;朱颢东(1980-),男,博士,主要研究领域:软件过程技术与方法、文本挖掘。
收稿日期:2009-04-08 **修回日期**:2009-06-05

$$\gamma_C(D)=\frac{1}{|U|}\sum_{i=1}^{m}|Pos_C(Y_i)|,\text{其中 } Y_i\in U/D \tag{1}$$

定理 1[8] 设信息系统 $S=<U,C\cup D,V,f>$,$B\subseteq C$,如果有 $\gamma_B(D)=\gamma_C(D)$,则 B 是 C 的一个特征约简集。

定义 2[5] 特征 a 加入 $R\subseteq C$,对于分类 U/D 的重要度定义为:

$$SGF(a,R,D)=\gamma_{R+(a)}(D)-\gamma_R(D) \tag{2}$$

$SGF(a,R,D)$的值越大,说明在已知特征集 R 的条件下,特征 a 对决策 D 就越重要。

2 并行遗传算法

遗传算法(GA)是一类基于自然选择和遗传学原理的有效搜索方法，许多领域成功地应用遗传算法得到了问题的满意解[9]。但随着问题规模和复杂程度的不断提高,GA 在求解中面临着两个主要困难:(1)存在“漂移”现象,使得计算、搜索时间过长;(2)存在“早敛”问题,即由于选择压力的影响,进化搜索过程中容易失去含关键基因的染色个体，使种群多样性下降,致使搜索过程滞留在局部最优解。对此,很多专家做了大量的研究工作,其主要方法包括:改进选种机制和进化算子、引用自调整技术、优化参数设置和采用并行实现[10]。随着计算机技术的飞速发展,应用并行实现被视为克服以上两点困难的最有效办法之一。

并行遗传算法(PGA)是 GA 的一个主要研究方向。目前,人们正不断地致力于把 GA 应用到各种并行计算机上。并行实现方案主要有:同步主从式、异步同时式、网络式[11-12]。采用网络式方案中的主从式控制网络并行 GA 来解决特征选择问题。

由于并行遗传算法主要以遗传算法为基础,下面详细介绍遗传算法的设计。

2.1 编码方案

在特征选择问题中,一个特征要么属于所选的特征集要么不属于所选的特征集,因此,采用 0 和 1 的二进制一维编码形式是合适的。其具体方法如下:

假设原始特征集 $C=\{c_1,c_2,\cdots,c_n\}$,其空间可以方便地影射为染色体。染色体是一个长度为 n 的 0 和 1 字串,每位对应一个特征。如果在某个位置为 1,则表明选择该特征,否则不选,这样每个染色体就对应一个特征集。

2.2 初始种群方案

随机选择一个随机数 m，产生 m 个长度为 n 的 0、1 字符串,作为初始种群。

2.3 适应度函数设计

根据实际情况,定义适应度函数如下:

$$F(x)=\begin{cases}\dfrac{Card(C-B(x))}{Card(C)}\dfrac{\beta}{1+e^{a(k_0-\gamma_{B(x)}(D))}}, & \gamma_C(D)-\gamma_{B(x)}(D)\geqslant\varepsilon\\ \dfrac{2\times Card(C-B(x))}{Card(C)}\dfrac{\beta}{1+e^{a(k_0-\gamma_{B(x)}(D))}}, & \gamma_C(D)-\gamma_{B(x)}(D)<\varepsilon\end{cases} \tag{3}$$

其中,ε 表示精度误差,$B(x)$表示个体 x 中对应位为 1 的特征组成的集合。很明显特征集 $B(x)$元素的数越少、对决策类 D 的支持度越大,适应度函数值也就越大。β 为惩罚因子,k_0 是预设的阈值。适当地选择 a 的值可使$\dfrac{1}{1+e^{a(k_0-\gamma_{B(x)}(D))}}$在 $\gamma_{B(x)}(D)>k_0$ 时取值近似为 1,而在 $\gamma_{B(x)}(D)<k_0$ 时其值迅速衰减为 0,也就是使精度不满足预设阈值 k_0 的个体适应度变得很小，而满足预设阈值 k_0 的个体适应度基本不受影响。由于求解原始特征集中的最小相对特征集，实际就是要在保持整体特征支持度不变的情况下寻找所含特征最少的特征集，而构造的适应度函数恰好从这两方面满足了问题的求解要求。在适应度函数 $F(x)$中,$Card(C-B(x))/Card(C)$的目的是要求 x 中所含特征的个数尽可能的少,$1/(1+e^{a(k_0-\gamma_{B(x)}(D))})$ 的目的是要求 x 中所含特征对决策类的支持度尽可能的大。如果 $\gamma_C(D)-\gamma_{B(x)}(D)<\varepsilon$,则说明此时特征集$B(x)$接近最优,应适当提高 X 的适应度。由此选择的适应度函数可获得特征集的最佳的搜索效果。

2.4 选择算子设计

选择算子采用轮盘赌方式实现。假设染色体规模为 m,表示为 $G(x)=\{x_1,x_2,\cdots,x_m\}$,个体 x_j 的适应度值为 $F(x_j)$,那么它被选择的概率为:

$$P(x_j)=\frac{F(x_j)}{\sum_{i=1}^{m}F(x_i)}\quad j=1,2,\cdots,m \tag{4}$$

经过选择操作生成用于繁殖的交配池,其中父代种群中个体生存的期望数目为:$G(x_j)=m\times P(x_j)$,$j=1,2,\cdots,m$。选择过程体现了生物进化过程中“适者生存,优胜劣汰”的思想,并可保证优良基因传给下一代。

2.5 自适应交叉算子设计

在特征选择时每个特征的重要性是不同的,选择重要性作为交叉算子的设计基础。对于两个个体:$X_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$,$X_j=\{x_{j1},x_{j2},\cdots,x_{jn}\}$,则操作描述如下:

$$O(P_{ij},s_{ij})=\begin{cases}x_{ik}=\begin{cases}x_{ik} & s_{ij}\geqslant P_{ij}\\ x_{jk} & s_{ij}<P_{ij}\end{cases}\\ x_{jk}=\begin{cases}x_{jk} & s_{ij}\geqslant P_{ij}\\ x_{ik} & s_{ij}<P_{ij}\end{cases}\end{cases},k=1,2,\cdots,n \tag{5}$$

其中 $s_{ij}\in[0,1]$为个体 X_i 和个体 X_j 交叉计算时的均匀随机概率变量,P_{ij} 为个体 X_i 和个体 X_j 交叉计算时的交叉概率，这里令 $P_{ij}=(r_{B(X_i)}(D)+r_{B(X_j)}(D))/2$。

2.6 自适应变异算子设计

变异运算是通过按变异概率 P_m 随机反转某位等位基因的二进制字符值来实现,对于给定的染色体位串 $X_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$,具体操作如下:

$$O(P_{ik},s_{ik}):x_{ik}=\begin{cases}x_{ik} & ,s_{ik}\leqslant P_{ik}\\ 1-x_{ik} & ,s_{ik}>P_{ik}\end{cases},k=1,2,\cdots,n \tag{6}$$

其中 $s_{ik}\in[0,1]$是个体 X_i 第 k 基因位产生变异的均匀随机概率变量,p_{ik} 为个体 X_i 第 k 基因位产生变异时的变异概率，这里令 $p_{ik}=1-SIG(B(x_{ik}),B(x_i),D)$,$B(x_{ik})$为个体 X_i 第 k 基因位的值对应的特征,如果值为 0,则 $B(x_{ik})=\varnothing$,否则 $B(x_{ik})$就为对应的特征。

2.7 中止条件的设计

以预先设定的最大繁殖代数 MAX 作为停止准则。

2.8 基于遗传算法的特征选择

为能有效地将粗糙集的属性约简的特性用于指导遗传算法寻优,在标准遗传算法基础上做了重点改进:一是在适应度函数中引入了惩罚函数，从而可使算法快速收敛到全局最优解;二是根据实际情况自动设置交叉概率和变异概率,避免了人工盲目设置这两个参数的缺陷,使得遗传算法有了一定的自

适应性。基于遗传算法的特征选择步骤如下:

步骤 1 首先由公式(1)计算出决策类 D 关于原始特征集 C 的支持度 $\gamma_C(D)$。

步骤 2 由随机产生的 m 个长度为 n(n 等于原始特征集的长度)的二进制串所代表的个体组成初始种群。

步骤 3 对每个个体 X,计算它的 $r_{B(x)}(D)$,然后根据公式(3)计算 X 的适应度 $F(X)$。

步骤 4 首先计算总体适应度 $F=\sum_{i=1}^{m}F(X_i)$,然后根据公式(4)计算出每个个体被选择的概率 $P(X_i)$以及累积选择概率 $P=\sum_{i=1}^{m}P(X_i)$,最后以轮盘赌方式选择个体。

步骤 5 按公式(5)进行自适应交叉计算。

步骤 6 按公式(6)进行自适应变异计算。

步骤 7 判断是否遗传代数等于预先设定的最大代数 MAX,如果是则终止计算;否则转步骤 3。

2.9 基于并行遗传算法的特征选择

采用主从式控制网络并行 GA,它将其中一个子群体设置成中心,其它子群体均与中心子群体通讯。中心子群体中始终保存当前最优个体,其它子群体通过"引进"中心子群体中的最优个体来加快收敛速度[13]。在网络环境下,通常采用最优个体迁移或最优个体广播的方法来实现粗粒度的并行 GA,这两种方法与主从式控制网络并行 GA 相比存在着不足:通信开销很大。同时实验结果表明[9],在相同的网络环境下,主从式控制网络并行 GA 优于最优个体迁移的并行 GA,也略优于最优个体广播的对等式并行 GA。那么,基于并行遗传算法的特征选择步骤如下:

步骤 1 产生一个进程,该进程为父进程,在进行基于 GA 的特征选择的同时用于存放和发送当前最优个体。

步骤 2 由父进程产生 $t-1$ 个子进程(每个子进程用于实现基于 GA 的特征选择)。

步骤 3 各子进程(包括父进程)进行基于 GA 的特征选择,当子进程中遗传代数被 10 整除,子进程发送最优个体至父进程。

步骤 4 父进程选择当前各子进程中最优个体(Best)发送给各子进程。

步骤 5 各子进程把 Best 替换各子进程当前代种群中适应度值最低个体。

步骤 6 若遗传代数小于设定最大繁殖代数,则转步骤 3。

步骤 7 算法终止。

3 实验例证

3.1 实验语料库

进行文本分类方面的实验,语料库的选择是非常重要的。选择的原则是国内外使用广泛、权威标准和规范。这样使得实验和国内外同行的实验结果具有可比性,同时也便于分析实验数据、分析算法的优劣。

在中文文本分类方面,经过分析和比较,选用的分类语料库是复旦大学中文文本分类语料库。该语料库由复旦大学计算机信息与技术系国际数据库中心自然语言处理小组构建,语料文档全部采自互联网,它可以从网上免费下载,网址为:http://www.nlp.org.cn/categories/default.php?cat_id=16。

复旦大学中文文本分类语料库中包含 20 个类别,分为训练文档集和测试文档集两个部分。每个部分都包括 20 个的子目录,相同类别的文档存放在一个对应的子目录下;每个存储文件只包含一篇文档,所有文档均以文件名作为唯一编号。共有 19 637 篇文档,其中训练文档 9 804 篇,测试文档 9 833 篇;训练文档和测试文档基本按照 1:1 的比例来划分。去除部分重复文档和损坏文档后,共保留有文档 14 378 篇,其中训练文档 8 214 篇,测试文档 6 164 篇,跨类别的重复文档没有考虑,即一篇文档只属于一个类别。该语料库中的文档的类别分布情况是不均匀的。其中,训练文档最多的类 Economy 有 1 369 篇训练文档,而训练文档最少的类 Communication 有 25 篇训练文档;同时,训练文档数少于 100 篇的稀有类别共有 11 个。训练文档集和测试文档集之间互不重叠。取前 10 个类的部分文档,其类别文档统计数如表 1 所示。

表 1　中文文本分类语料库各类别文档数统计

类别	训练文档数目	测试文档数目	类别	训练文档数目	测试文档数目
经济	480	419	环境	405	371
体育	584	489	艺术	510	286
计算机	628	591	太空	506	248
政治	573	482	历史	466	468
农业	547	435	军事	74	75

3.2 实验环境及参数设置

实验所用计算机配置如下:操作系统均为 Microsoft Windows XP Professional (SP2),CPU 规格均为 Intel® Celeron® CPU 2.40 GHz,内存 512 M,硬盘 80 G。

进行中文分词处理时,采用的是中科院计算所开源项目汉语词法分析系统 ICTCLAS 系统。

实验使用的软件工具是 Weka,这是纽西兰的 Waikato 大学开发的数据挖掘相关的一系列机器学习算法。实现语言是 Java。可以直接调用,也可以在代码中调用。Weka 包括数据预处理、分类、回归分析、聚类、关联规则、可视化等工具,对机器学习和数据挖掘的研究工作很有帮助,它是开源项目,网址为:http://www.cs.waikato.ac.nz/ml/weka/。实验使用的计算工具为 MATLAB 7.0。算法中各参数需要反复试验才能得到,经实验算法中各参数最后设置如下:$a=40$,$\beta=2$,$k_0=0.95$,$\varepsilon=0.095$,$m=100$,n 等于原始特征集的长度,MAX=400。

3.3 实验所用分类器及其评价标准

本实验旨在比较本方法与信息增益(IG)、x^2 统计量(CHI)、互信息(MI)等三种特征选择方法对后续文本分类精度的影响,因此本实验在各种特征选择方法后采用相同的分类器对文本进行分类。本实验中使用 KNN 分类器来比较这几种特征选择方法(K 设置为 10)。

为了评价分类效果,实验中选择分类准确率和召回率作为评价标准:准确率(Precision)=$a/(a+b)$,它是所判断的文本与人工分类文本吻合的文本所占的比率。召回率(Recall)=$a/(a+c)$,它是人工分类结果应有的文本与分类系统吻合的文本所占的比率。在实际中,查准率比查全率重要。其中 a、b、c 代表相应的文档数,它们的含义如表 2 所示。

表 2　二值联表

	真正属于此类	真正不属于此类
判断属于此类	a	b
判断不属于此类	c	d

表 3 准确率和召回率 (%)

类别	该文方法		IG		CHI		MI	
	准确率	召回率	准确率	召回率	准确率	召回率	准确率	召回率
经济	91.28	92.56	82.52	80.83	79.31	87.67	75.63	76.99
体育	89.17	91.67	83.88	82.93	81.71	85.60	79.54	80.78
计算机	90.07	91.56	87.64	88.43	82.41	83.51	80.71	77.91
政治	89.67	88.33	78.78	84.29	83.29	78.80	79.99	80.72
农业	88.33	90.18	83.27	89.67	79.56	77.23	72.48	79.45
环境	90.48	88.67	81.67	86.42	81.93	86.56	76.42	80.13
艺术	90.27	92.78	80.55	85.81	82.51	82.78	80.51	81.81
太空	89.67	93.88	82.46	87.47	80.84	79.23	78.57	78.47
历史	90.73	89.91	80.33	87.39	78.34	80.42	77.45	81.92
军事	88.81	92.95	75.53	79.73	60.94	87.67	63.67	74.71
平均率	89.95	91.25	81.66	85.30	79.08	82.95	76.50	79.29

为了评价各个特征选择方法的时间效果,串行算法采用特征选择过程所消耗的时间,而对于并行算法常采用加速比与效率分析。加速比 $S_p=T_s/T_p$,其中 T_s 是求解一个问题最快的串行算法在最坏情况下的运行时间, 而 T_p 是求解同一个问题的并行算法在最坏情况下的运行时间。可见加速比是评价算法的并行性对运行时间改进的程度。效率 $E_p=S_p/p$,其中 p 为处理器的个数,效率反映了并行系统中处理器的利用情况。

3.4 实验结果及其分析

表 3 表明了 4 种方法的准确率和召回率,可以看出它们从大到小的顺序依次为该文方法、IG、CHI、MI。由于该文方法在选择特征时,不但考查了特征的权重,而且还考查了它们之间的潜在的隐含关系, 对要选择的特征进行了较全面的考查,所以效果最佳;由于 IG 方法受样本分布影响,在样本分布不均匀的情况下,它的效果就会大大降低,但从整体上看本文所选样本分布相对均匀,只有极个别相差较大,所以总体效果次之;由于 MI 方法仅考虑了特征发生的概率,而 CHI 方法同时考虑了特征存在与不存在时的情况,所以 CHI 方法比 MI 方法效果要好。

表 4 和表 5 表明, 在选择特征子集过程所消耗的时间上,该文方法在一个 CPU 上处理的时间要劣于其他三种方法,但采用并行策略后所需时间要远远少于其他三种方法。

表 5 给出了该文方法在 CPU 个数变化时的运行时间、加速比和效率。从该表可看出,随着结点数的增加,速度有明显提高; 在子 CPU 个数是种群规模的约数时, 效率比较高, 达到 88%左右;当子 CPU 个数在 16 时,效率下降到 84%左右;当子 CPU 个数在 32 时,效率下降到 79%左右;整个效率趋势是下降的。产生这种情况的原因主在于:当子 CPU 个数是种群规模约数时,各个子 CPU 分配的种群个数相等,此时系统负载较平衡,各个子 CPU 可以较好地并行工作,主 CPU 不需要单独等待某个子 CPU 就可以工作;当子 CPU 个数不是种群规模约数时,此时负载不平衡,子 CPU 分配的工作量不同,因而完成的时间也不同,主 CPU 必须等待各个子 CPU 都结束工作后才能工作;由于随着 CPU 的个数不断增加,每个 CPU 的计算量在不断减小,这样数据传送时间与整个时间的比值就越大,导致效率逐渐降低。从这可以说明,在并行算法中,不要为了追求时间效率而无限增加 CPU 个数,那会造成资源的极大浪费,应该在加速比和效率之间做出一个权衡。

表 4 所用串行算法消耗的时间 s

IG	CHI	MI
1 427	1 538	1 496

表 5 不同处理器个数下时间、加速比、并行效率列表

CPU 个数	时间/s	加速比	效率
1	1 738	1.00	1.000 0
2	882	1.97	0.985 3
4	451	3.85	0.963 4
8	237	7.33	0.916 7
10	192	9.05	0.905 2
16	129	13.47	0.842 1
20	98	17.73	0.886 7
32	68	25.59	0.798 7

4 结束语

为了解决海量文本数据集上的特征选择问题,详细设计了一个基于并行遗传算法的特征选择方法,该方法采用遗传算法搜索特征,利用并行策略评价特征子集,即将种群中个体的适应度计算并行在多个计算节点上同时进行,从而较快地获得较具代表性的特征子集。实验证明,该文特征选择方法同三种特征选择方法:IG、CHI、MI 相比,有较高的准确率和召回率,而且花费的时间远远低于这三种方法所需的时间,从而使得该文方法在文本分类中有一定的使用价值,同时为中文文本特征选择提供一种思路,也为后续的知识发现算法减少了时间与空间复杂性。

参考文献:

[1] Delgado M,Martin-Bautista M J,Sanchez D,et al.Mining text data:Special features and patterns[C]//Proceedings of ESF Exploratory Workshop,London,UK,Sept 2002:32-38.

[2] 朱颢东,蔡乐才,刘忠英.一种改进的文本特征选择算法[J].现代电子技术,2008(8):97-99.

[3] Friedman N,Geiger D,Goldszmidt M.Bayesian network classifiers[J].Machine Learning,1997,29(2):131-163.

[4] 张海龙,王莲芝.自动文本分类特征选择方法研究[J].计算机工程与设计,2006,27(20):3838-3841.

[5] 周茜,赵明生,扈曼,等.中文文本分类中的特征选择研究[J].中文信息学报,2004,18(3):17-23.

[6] 胡佳妮,徐蔚然,郭军,等.中文文本分类中的特征选择算法研究[J].光通讯研究,2005(3):44-46.

(下转 217 页)