基于融合信息的癌症相关基因选择方法

张树波[1,2,4] 赖剑煌[3,4]

(广州航海高等专科学校计算机系 广州 510725)[1] (中山大学数学与计算科学学院 广州 510275)[2]
(中山大学信息技术与科学学院 广州 510275)[3] (广东省信息安全技术重点实验室 广州 510725)[4]

摘 要 基因表达数据的出现,为人类从分子生物学的角度研究和探索癌症的发病机理提供了广阔的前景,利用基因表达数据发现与癌症相关的基因对于癌症的诊断和治疗具有重要的意义。在过去的十几年里,已经有很多种计算方法被成功地用于从基因表达数据中找出与癌症相关的关键基因,然而,不同的方法从不同的角度刻画基因对不同类型样本的区分能力,它们选择出来的关键基因可能不一致,这将给医学解释和应用带来困扰。现提出一种融合的方法,即将基因在不同方面对样本的判别能力结合起来,首先计算每个基因的信息增益、全局判别能力和局部判别能力,再用它们的识别率进行加权,进而计算每个基因的综合判别能力,最后筛选出判别能力最高的基因子集作为关键基因子集。实验结果表明,此方法得到了比采用单独一种评价标准更好的识别效果。

关键词 基因表达数据,基因选择,信息增益,全局可分性,局部可分性,融合信息

中图法分类号 TP3-05 **文献标识码** A

Cancer Relevant Genes Selection Approach from Integrated Information

ZHANG Shu-bo[1,2,4] LAI Jian-huang[3,4]

(Department of Computer Science, Guangzhou Marine College, Guangzhou 510725, China)[1]
(School of Mathmatics and Computational Science, Sun Yat-sen University, Guangzhou 510275, China)[2]
(School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510275, China)[3]
(Province Key Laboratory for Technology of Security Information, Guangzhou 510725, China)[4]

Abstract With the advent of gene expression data, it has shown great promise to explore cancer pathogenesis at the level of molecular biology. Exploring the key genes related with carcinoma is of great importance to the cancer diagnosis and treatment. During the past few decades, many approaches were successfully proposed to select the key genes related with carcinoma. However, the gene sets selected by different methods are not always consistent with each other, this will cause difficulty to biological interpretation and practical application. A novel approach based on integrated information was proposed for the selection of key genes from gene expression data in this study, the three kinds of information, namely information gain, global discriminate ability and local discriminate ability were derived firstly, then they were weighted and integrated into a score to characterize the discriminate ability of each gene, and the gene set with the highest classification accuracy was selected as key gene set. The experimental results showed that our approach can get better performance than those based on single information.

Keywords Gene expression data, Gene selection, Information gain, Global separability, Local separability, Integrated information

1 引言

癌症是人类生命的一种严重威胁,尽早发现癌症,并对可能发生转移的癌症采用特定的辅助治疗措施,可以降低死亡率或者延长患者的寿命,同时也能够降低癌症治疗的成本和负担。因此,可靠并准确地对癌症进行分类和判别,对于成功地进行诊断和治疗癌症具有十分重要的意义。基因芯片技术的出现,使得一次可以在一块很小的胶片上同时检测出成千上万基因的表达水平。通过检测基因在不同情况下表达量的变化情况,能够从分子水平了解细胞的生命过程和机理。大量基因表达数据的出现为癌症研究提供了新的思路,近年来,采用生物信息学方法对基因芯片数据进行分析,在预测癌症发生与否、癌症发病、转移机理、药物设计和癌症治疗等方面的研究越来越引起人们的关注。

从数据分析的角度来看,基因表达数据具有如下特点:样本数量少、基因数量(样本维数)大、基因之间存在高度的相关性,这就容易导致通常所说的"维数灾难"问题,一方面降低了计算的可靠性,另一方面降低了运算的效率。因此,在对基因

到稿日期:2010-01-27 返修日期:2010-04-15 本文受国家自然科学基金(No. 60675016, 60633030)资助。

张树波 博士,主要研究方向为模式识别、生物信息学,E-mail:treaton@tom.com;赖剑煌 教授,博士生导师,主要研究方向为模式识别、数字图像处理。

表达数据进一步分析之前，必须对数据进行预处理，包括数据的降维和冗余性、相关性的消除，找出与研究问题相关的、非冗余的基因，剔除与研究问题无关的基因，这个过程称为基因选择。从生物学的角度看，基因选择的目标就是从基因表达数据中找出与特定研究问题关联的、具有特定生物意义的关键基因，比如找出与癌症密切相关的基因。

由计算学习的理论[1]可知，基因选择问题是一个复杂程度为 $O(2^N)$ 的 NP-难题，在过去的十几年中，人们为解决这类问题提出了各种各样的近似算法，在一定程度上促进了基因芯片数据分析技术的发展，Saeys 等人曾经就生物信息学领域中的基因选择问题做了一个较为全面的综述[2]。总结起来，基因选择的方法包括 3 大类：(1)Filter 基因选择方法。这类方法首先按照某个给定的度量标准计算每个基因重要性的得分值，然后对它们进行排序，再选择那些重要性得分高的基因作为目标基因，常用的度量有相关系数[3]、方差[4]、Fisher Score[5]、信息熵[6]、t 值[7]和 Wilcoxon 值[8]等，这类方法的优点是直观、简单，而且计算的复杂度也较低(复杂度为 $O(N)$)，但是它们没有考虑基因之间的相关性，被选择出来的关键基因之间存在冗余性问题，需要进一步将其中冗余的基因排除掉。(2)Wrapper 基因选择方法[9]。这类方法的经典做法是从特征空间搜索所有基因子集，对每个子集分别构造分类器，以分类器的分类效果(如正确识别率)作为每个子集的得分，将得分最高的基因子集作为目标子集。Wrapper 方法将基因子集的选择与相应分类器的性能结合起来，因此一旦确定了基因子集，就不需要进一步设计分类器，而且也将基因之间的相关性考虑进基因选择的过程，能够有效地消除基因之间的冗余性。但是其计算的复杂度非常高，算法的复杂程度明显高于 Filter 方法[10]，对维数高达上万维的基因选择问题是不可取的，因此人们提出了各种启发式搜索算法，比如顺序搜索算法[11]、分支界限法[12]、遗传算法[13]和束搜索算法[14]等。(3)Embedded 基因选择方法。这类方法将基因选择问题嵌入到分类器的设计中，而且计算的效率比较高。比如文献[15]就是将基因选择问题嵌入到支持向量机分类器的设计过程中。

不同的基因选择标准从不同的角度对数据进行考察，比如基于方差的基因选择方法关注的是数据的离散程度，而基于 Fisher Score 的选择方法同时关注数据类内差异与类间差异，基于信息熵的选择标准关注的是基因与类标之间的关联程度。因此对相同的数据集，采用不同的标准选择出来的基因不一定相同，比如 xiong 等人[9]在一个乳腺癌的基因表达数据集上，以分类误差为标准找到的 17 个关键基因，每个基因的预测准确性都超过 90%，而 Hedenfalk 等人用 Wilcoxon-Mann-Whitney 检验方法只找到 6 个预测准确率超过 90%的关键基因[16]。另外，这两种方法找到的关键基因只有 4 个是相同的。这就给实际应用带来了困难，因为当医生拿到这两个结果时，不知道哪个是可靠的，无从下手，这样的信息对他们来说应该不具有什么实际参考价值。因此研究稳定的、可靠的基因选择方法是生物信息学面临的重要任务。

本文提出了一种基于融合信息的癌症相关基因选择方法。首先提出了 3 种度量基因重要性的标准：信息增益、全局可分性和局部可分性，接着根据这 3 种标准分别对每个基因的重要性进行排序，再用其样本判别准确率对各个重要性得分进行加权，进而计算出每个基因重要性的综合得分，最后筛选掉那些得分低的基因，得到一个包含重要性得分较高的基因子集。本文方法在 3 个癌症相关的基因表达数据集上得到了比较理想的实验效果。

2 数据集

2.1 直肠癌数据集[17]

这个数据集由 Alon 等人于 1999 年提供，它包含 2000 个基因，62 个样本中有 40 个来自人体的直肠癌组织，22 个来自正常的组织。Alon 等人采用 Bi-clustering 技术将样本聚成两类，从样本的维度看，一类主要包含正常样本，另一类主要包含非正常样本。从基因的维度看，具有相互调控关系的基因也被有效地聚类到一起。这个数据集的芯片平台是 Affymetrix 公司的寡核苷酸阵列，可以从如下网址下载该数据集的原始数据：http://microarray. princeton. edu/oncology/affydata。

2.2 弥漫型大型 B 细胞淋巴瘤数据集[18]

这个数据集由 Alizadeh 等人提供，包含 4026 个基因，42 个样本：增殖中心弥漫型大型 B 细胞淋巴瘤样本(GC B-like DLBCL)和活性弥漫型大型 B 细胞淋巴瘤样本(activated B-like DLBCL)各 21 个，Alizadeh 等人用 Pearson 相关系数作为基因之间的距离度量，然后采用层次聚类技术较好地将这两个类型分开。该数据集的芯片平台是 GenePix 4000，下载地址为：http://llmpp. nci. nih. gov/lymphoma。

2.3 乳腺癌数据集[19]

这个数据集由 van't Veer 等人于 2002 年提供，他们用这个数据集研究乳腺癌患者在 5 年时间内是否发生复发现象，该数据集包含 24481 个基因，共有 97 个乳腺癌早期患者的样本，其中 46 个样本在 5 年内复发，51 个样本在 5 年内不会复发。Van't Veer 等人用基因与类标之间的相关系数作为基因选择标准，先挑选出 231 个相关系数较大的基因，然后用 wrapper 方法找出一个包含 70 个基因的最优子集。采用留一法验证的预测准确率为 83%。该数据集使用的芯片平台是 Agilent Oligonucleotide，该数据集的下载地址为：http://www. rii. com/publications/2002/ vantveer. html。

3 方法

3.1 基于信息增益的基因选择标准

熵是信息论中的一个重要概念，它表示随机变量取值的“随机性“程度，数据的取值越是随机的，那么一次实验观察结果的不确定性程度就越高，它的熵就越大。对于一个离散型随机变量 Y，其熵可以定义为：

$$H(Y)=-\sum_i P(Y=y_i)\log P(Y=y_i) \tag{1}$$

有了信息熵的概念，我们就可以定义随机变量的信息增益：给定 X 条件下随机变量 Y 的信息熵减少的量：

$$IG(Y|X)=H(Y)-H(Y|X) \tag{2}$$

式中，$H(Y|X)=H(X,Y)-H(X)$，因此有：

$$IG(Y|X)=H(Y)-H(Y|X)=H(Y)+H(X)-H(X,Y) \tag{3}$$

计算信息增益时需要对每个特征的值进行离散化，我们采用 Tourassi 等人提出的离散化方法[20]。一个基因对类标的信息增益越大，说明这个基因提供的与类标相关的信息越

多，越有利于对样本进行分类，因此它越可能是一个关键基因。

3.2 基于全局判别能力的基因选择标准

在统计分析中，不同类样本之间的差异可以用类中心之间的距离来刻画，而同一类样本内部的差异则用方差来表示。1936 年著名统计学家 Fisher 提出将原始空间的样本线性投影到某个子空间，使得在新的空间中同类样本之间的差异尽量小，而不同类样本之间的差异尽量大，这就是后来被广泛应用的 Fisher 线性判别分析方法。在模式分类中，如果不同类样本中心之间的距离大，而且同类样本内部的方差小，那么我们就认为这样的样本可分性高，这就是一种经典的样本可分性原则。Weston 等人基于这种思想，提出了一种 Fisher Score[21] 准则来度量特征对样本的判别能力，进而用于特征选择，其定义形式如下：

$$fs=\frac{(\bar{x}_1-\bar{x}_2)^2}{\sum_{k=1}^{n_1}(x_1^k-\bar{x}_1)^2+\sum_{k=1}^{n_2}(x_2^k-\bar{x}_2)^2}=\frac{(\mu_1-\mu_2)^2}{\sigma_{(+)}^2+\sigma_{(-)}^2} \tag{4}$$

式中，fs 的值越大，表示特征 x 对两类样本的区分能力强，反之，则认为它对两类样本的区分能力小。本文用 Fisher Score 来度量基因对不同类别样本的全局判别能力。

3.3 基于局部判别能力的基因选择标准

除了可以用上面的 Fisher Score 方法从全局的角度衡量基因对样本的判别能力，也可以从局部的角度来衡量它们对样本的判别能力，即在一定的局部范围内，对于某个基因来说，如果同类样本之间距离与不同类样本之间距离的差异越大，则说明这个基因对样本的区分能力越大，越有利于将不同类样本分开。本文提出一种基于模糊近邻的度量，用样本点到它的 k(k 值取 3)邻域内同类样本和不同类样本点之间平均距离的差异来刻画样本的可分性。对于某个样本 i，首先找出与它距离最近的 k 个样本，然后将这 k 个样本分为两类：一类是与它同类，记为 A_i，另一类与它不同类，记为 B_i，分别用 $|A_i|$ 和 $|B_i|$ 表示两个集合中的样本数，并且分别定义样本 i 到集合 A_i 和 B_i 的距离如下：

$$d_{iA_i}=\frac{1}{|A_i|}\sum_{y_j\subset A_i}|x_i-y_j| \tag{5}$$

$$d_{iB_i}=\frac{1}{|B_i|}\sum_{y_j\subset B_i}|x_i-y_j| \tag{6}$$

则对所有样本点，可以计算在 k 近邻内它们与同类样本之间的平均距离以及与不同类样本之间的平均距离如下：

$$\bar{d}_A=(\sum_{i=1}^{N}\sum_{y_j\subset A_i}|x_i-y_j|)/\sum_{i=1}^{N}|A_i| \tag{7}$$

$$\bar{d}_B=(\sum_{i=1}^{N}\sum_{y_j\subset B_i}|x_i-y_j|)/\sum_{i=1}^{N}|B_i| \tag{8}$$

式中，N 表示样本的数量，$\bar{d}_A$ 越小说明各样本在 k 邻域范围内与同类样本之间的距离越近，$\bar{d}_B$ 越大说明样本在 k 邻域范围内与不同类样本之间的距离越远，由此可知，越小的 $\bar{d}_A$ 和越大的 $\bar{d}_B$ 越有利于将不同类的样本分开，于是可以根据这种思想定义特征的局部可分性度量：

$$Lss=\bar{d}_B-\bar{d}_A \tag{9}$$

对于基因选择问题，Lss 的值越大，说明它的分类能力越强，这样的基因很可能就是一个关键基因。

3.4 基于多种选择标准融合的基因选择方法

由于不同的基因选择标准从不同角度刻画基因对样本的分类能力，因此用不同的评价标准来选择关键基因通常会得到不同的结果，这就会产生基因选择的鲁棒性问题。本文尝试将不同的基因选择标准结合起来，构造一个综合的评价标准。我们首先按照上面的 3 种标准分别对每个基因的重要性进行排序，得到各个基因的次序统计量，然后用每个基因对样本的正确识别率分别对这些次序统计量进行加权，得到加权的次序统计量，用于表示每个基因的重要性。记 r_{ij} 为第 i 种标准中第 j 个基因的次序统计量，w_{ij} 为第 i 种标准中第 j 个基因的识别率，那么可以根据下面的公式计算这个基因的重要性综合得分：

$$R_j=\sum_{i=1}^{3}(1-\frac{r_{ij}}{m})w_{ij}\quad j=1,2,\cdots,m \tag{10}$$

式中，m 表示基因的总数。由于 r_{ij} 越小表示第 j 个基因在第 i 种标准中对样本具有越好的判别能力，因此越大的 R_j 值说明第 j 个基因对样本具有越好的判别能力。

4 实验结果与讨论

为了比较不同基因选择标准的效果，本文采用 Wrapper 方法构造候选基因子集，对每个子集分别构造分类器，并以分类器在基因子集上对样本的分类效果作为衡量标准。基因子集的构造采用 Foreward 策略，首先选择重要性得分最高的那个基因，然后依次从剩下的基因中将重要性得分最高的基因包括进来，形成一个基因子集序列 $S_1,S_2,\cdots,S_n$，这些集合之间具有相互包含的关系 $S_1\subset S_2\subset\cdots\subset S_n$，接着以每个基因子集 S_i，$i=1,2,\cdots,n$ 为特征，设计 k 近邻分类器(k 值取 3)。由于各数据集中的样本数量比较少，我们采用留一法对实验结果进行验证。

图 1 至图 3 显示了本文方法在各数据集上的实验结果。从图中可以看出，采用融合方法选出的基因子集比采用单独一种基因评价标准具有更好的识别效果。图 1 是在直肠癌数据集上的实验结果。从该图可以看出，基于融合方法筛选出来的基因的识别效果明显优于各种单独的评价标准。当选择的基因数量增加到 5 个时，本文方法的识别率曲线到达第一个高峰，这些基因对样本的正确识别率为 98.1%，之后随着基因数量的增加，分类器的正确识别率出现多次下降和上升交叉的趋势，整条曲线一共出现 9 个高峰。尽管随着候选基因的增加，分类器的识别效果会发生一些波动，但是总体上看，基于信息融合的基因选择方法还是取得了最好的识别效果。

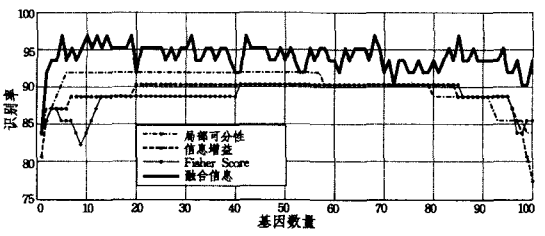

图 1 不同基因评价标准在直肠癌数据集上的分类效果

图 2 是在急性淋巴癌数据集上的实验结果比较。从该图可以看出，各种基因筛选方法具有相似的规律，即随着候选基因的增加，分类器的识别效果曲线出现多次的波动。对于基于融合的方法来说，虽然其识别率在某些时候还没有其他方法的高，但是从整体来看，该方法得到的正确识别率最高，分别在基因数量为 16，25，42，66，71 的时候取得最高的正确识

别率 99.1%。

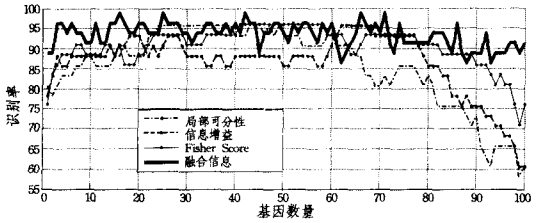

图 2 不同基因评价标准在急性淋巴癌数据集上的分类效果

图 3 是在乳腺癌数据集上的实验结果比较。从该图可以看出，各种方法随着基因数的增加，分类器的识别率出现了上升→下降→上升→下降的趋势。而对于基于融合方法的识别率曲线来说，分类器的识别率随着基因数量的增加呈逐渐上升趋势，当基因的数量为 17 时到达第一个高峰，这 17 个基因对样本的识别率为 93.8%；之后随着基因数量的增加，分类器的识别效果开始出现下降趋势，当基因的数量为 25 个时，分类器的识别率曲线到达第一个低谷，这时的正确识别率只有 86.3%；接着分类器的识别率又随着基因的增加而逐渐上升，当基因数量增加到 35 时，分类器的识别率曲线又到达另一个高峰，其正确识别率为 94.6%；在第 35 个基因之后，分类器的识别率总体呈下降趋势，一直到基因数为 100 时下降到 64.8%。尽管有时候融合方法的识别率不如其他方法，但是它能够找到具有最高识别率的基因子集。

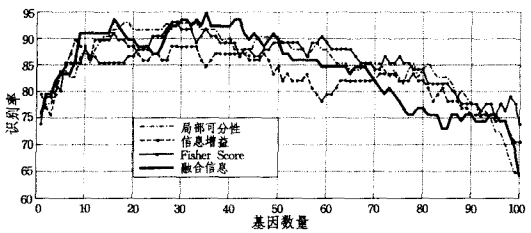

图 3 不同基因评价标准在乳腺癌数据集上的分类效果

生物体是一个非常复杂的系统，基因之间存在各种各样的调控关系，这些调控关系在基因表达量方面的反应就是一个(或者)一些基因表达量的变化将导致另外一些基因表达量的变化，也就是基因表达数据之间存在相关性。这种相关性从计算的角度来看，就是数据之间存在冗余性，这些冗余性的存在将对分类效果产生不利的影响，正如上面各实验结果显示那样，随着新基因的加入，分类器的识别效果发生波动，这种波动在很大程度上就是由于加进来的基因与原来的基因之间存在信息冗余，影响了分类的效果，所以应该在分类器的设计过程中尽可能地删除基因之间的冗余性。然而，这种冗余性具有非常重要的生物学意义，基因表达量之间存在相关性说明这些基因之间在生物功能方面是相关的，它们很可能就是某个生物通路上具有相互调控关系的基因，或者是协同完成某种生物功能的关键基因，不能被简单地筛选掉。所以在利用基因表达数据筛选与癌症相关基因的过程中，既要考虑提高基因对样本的判别能力，也要考虑具体的生物学含义。

结束语 基因表达数据的出现，为人类从分子生物学的角度研究和探索癌症的发生和发展机理、癌症的预测、药物设计、癌症防治等提供了广阔的前景，利用生物信息学的方法对基因芯片数据进行分析，进而筛选出与癌症相关的基因，已经成为生物研究的一个热点问题。然而，由于不同的计算方法从不同的角度刻画基因对不同类型样本的区分能力，它们选择出来的关键基因可能不一致。本文提出了一种基于融合信息的基因选择方法，其将基因在不同方面对样本的判别能力结合起来。首先计算每个基因的信息增益、全局判别能力和局部判别能力，然后用它们对样本的识别率进行加权，计算出每个基因的综合得分，用于表示基因对不同类型样本的区分能力。在 3 个癌症相关的基因表达数据集上的实验结果表明，基于多种信息融合选择出来的基因子集比采用单独一种信息选出的基因具有更好的识别效果。

由于基因表达数据具有生物意义的相关性，如何在找出所有与癌症相关的基因的同时也提高选出基因的判别能力成为今后基因表达数据分析工作的重要研究课题。其中一种研究的思路是先找出具有最佳判别能力的基因子集，然后利用基因表达数据之间的相关性和其他方面的知识进一步找出与这些判别能力高的基因高度相关的其他基因，进而研究它们之间的关系，更好地了解癌症的发生机理，为疾病的诊断和治疗提供有用的参考信息。

参 考 文 献

[1] Mitchell T M. Machine learning[M]. New York: McGraw-Hill, 1997

[2] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 23(19): 2507-2517

[3] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286: 531-537

[4] Hastie T, Tibshirani R, Eisen M, et al. Gene shaving as a method for identifying distinct sets of genes with similar expression patterns[J]. Genome Biology, 2000, 1(2): 1-21

[5] Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data[J]. Journal of the American Statistical Association, 2002, 97(457): 77-88

[6] Xing E, Jordan M, Karp R. Feature selection for high-dimensional genomic microarray data[C]// Proceedings of the Eighteenth International Conference on Machine Learning. 2001: 601-608

[7] Jafari P, Azuaje F. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors[J]. BMC Medical Informatics and Decision Making, 2006, 6(1): 27

[8] Miller L D, Long P M, Wong L, et al. Optimal gene expression analysis by Microarrays[J]. Cancer Cell, 2002, 12: 353-361

[9] Xiong M, Fang Z, Zhao J. Biomarker identification by feature wrappers[J]. Genome Research, 2001, 11: 1878-1887

[10] Blum A L, Langley P. Selection of relevant features and examples in machine learning[J]. Artificial Intelligence, 1997, 97: 245-271

[11] Liu H, Motoda H. Feature selection for knowledge discovery and data mining[M]. Boston: Kluwer Academic Publishers, 1998

[12] Somol P, Pudil P, Kittler J. Fast branch & bound algorithms for optimal feature selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(7): 900-912

[13] Li S, Tan M. Gene selection and tissue classification based on support vector machine and genetic algorithm[C]// The 1st International Conference on Bioinformatics and Biomedical Engineering(ICBBE 2007). vol. 6, 2007: 192-195

(下转第 196 页)

差结果(平均差异可达 6.4%)。故表 2(及表 3)所列数据虽为随机一次实验的结果,但足以代表相关算法的普遍性能。

3.2 异构调度问题

第二组实验测试 PSACS 求解异构调度问题的性能。就笔者所知,目前没有异构 DAG 调度问题的标准测试集,因此没有最优解可供参照。因 PSACS 嵌入了 HEFT 使用的启发式知识,本文使用 HEFT 作为基准比较其它算法的性能。异构 DAG 图由本文 3.1 节所使用的测试用例转化而来,方法如下:保持 DAG 图结构不变,XXX.stg 文件中登记的任务执行时间 $weight_i$ 为任务 t_i 在异构处理器上运行时间的均值,$ETC_{i,j}$ 为[0,2 * $weight_i$]区间内均匀分布的随机变量。

实验结果(表 3)为:随机对所有 300 个被测试的异构 DAG 调度问题进行一次实验,21%的 PSACS 调度结果优于 HEFT 10%以上,最高改善比可达 27.59%。纯遗传算法的性能远远差于 PSACS:300 个问题中仅有 2 个问题的纯遗传算法的调度结果优于 HEFT 10%(最高改善比 11%)。此结论表明,PSACS 性能不仅优于纯遗传算法,而且优于最近的研究成果(DAG 调度问题的粒子群优化算法[12]),后者取 10 次实验中的最优值,对 HEFT 的改善比最高仅为 18.3%。

表 3 PSACS 求解异构调度问题的能力
(随机一次试验的统计结果)

	优于 HEFT(10%)的次数		对 HEFT 的最优改善比例	
	PSACS	纯 GA	PSACS	纯 GA
2pe random	3	0	12.8%	0.7%
4pe random	5	0	16.2%	−4%
8pe random	14	0	20.3%	1%
16pe random	17	0	24.4%	−29.3%
32pe random	9	0	27.6%	−29.2%
8pe high	14	2	27.4%	10.6%

结束语 本文给出蚁群系统 PSACS,用以解决异构 DAG 调度问题。因恰当地使用了启发式知识,算法性能明显优于遗传算法和粒子群算法。值得注意的是,本文并未使用任何复杂的演化技术,仅使用 ACS 默认参数和解空间搜索技术,PSACS 的性能便已相当可观。此结论表明蚁群优化方法适合求解异构 DAG 调度问题。

参 考 文 献

[1] UllMan J D. NP-complete scheduling problems[J]. J. of Computer and Systems Sciences, 1975, 10: 384-393

[2] Topcuoglu H, Hariri S, Wu Min-you. Performance effective and low-complexity task scheduling for heterogeneous computing [J]. IEEE Transactions on Parallel and Distributed Systems, 2002, 13(3): 260-274

[3] Bajaj R, Agrawal D P. Improving Scheduling of Tasks in A Heterogeneous Environment[J]. IEEE Transactions on Parallel and Distributed Systems, 2004, 15(2): 107-118

[4] Cirou B, Jeannot E. Triplet: A clustering scheduling algorithm for heterogeneous systems[C]// International Conference on Parallel Processing Workshop. 2001: 231-236

[5] Hou E S, Ansari N, Ren H. A Genetic Algorithm for Multiprocessor Scheduling[J]. IEEE Transactions on Parallel and Distributed Systems, 1994, 5(2): 113-120

[6] Correa R C, Ferreira A, Rebreyend P. Scheduling Multiprocessor Tasks with Genetic Algorithms[J]. IEEE Transactions on Parallel and Distributed Systems, 1999, 10(8): 825-837

[7] Dhodhi M K, Ahmad I, Yatama A, et al. An Integrated Technique for Task Matching and Scheduling onto Distributed Heterogeneous Computing Systems[J]. Parallel and Distributed Computing, 2002, 62(9): 1338-1361

[8] Dorigo M. Optimization, Learning and Natural Algorithms[D]. DEI, Polytecnico di Milano, Milan, Italy(in Italian), 1992

[9] Marco D, Mauro B, Thomas S. Ant Colony Optimization: Artificial Ants as a Computational Intelligence Technique[J]. IEEE Computational Intelligence Magazine, 2006(11): 28-39

[10] Chen Wei-neng, Zhang Jun. An Ant Colony Optimization Approach to a Grid Workflow Scheduling Problem with Various QoS Requirements[J]. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, 2009, 39(1): 29-43

[11] Udo H, Wolfram S. A Comprehensive Test Bench for the Evaluation of Scheduling Heuristics[C]// Proc. of the 16th International Conference on Parallel and Distributed Computing and Systems (PDCS'04). Cambridge, USA: ACTA Press, 2004: 840-845

[12] Xiao H K, Jun S, Wen B X. Permutation-based particle swarm algorithm for tasks scheduling in heterogeneous systems with communication delays[J]. International Journal of Computational Intelligence Research, 2008, 4(1): 61-70

(上接第 174 页)

[14] Gupta P, Doermann D, DeMenthon D. Beam search for feature selection in automatic SVM defect classification[C]// the 16th International Conference on Pattern Recognition. 2002: 212-215

[15] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46: 389-422

[16] Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer[J]. N Engl J Med, 2001, 344: 539-548

[17] Alon U, Barkai N, Notterman D A. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. Proc Natl Acad Sci USA, 1999, 96: 6745-6750

[18] Alizadeh A, Eisen M B, Davis R E, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling [J]. Nature, 2000, 403: 503-511

[19] van't Veer L J, Dai H, van de Vijver M J, et al. Gene expression profiling predicts clinical outcome of breast cancer[J]. Nature, 2002, 415(6871): 530-536

[20] Tourassi G D, Frederick E D, Markey M K, et al. Application of the mutual information criterion for feature selection in computer-aided diagnosis[J]. Med Phys, 2001, 28: 2394-2402

[21] Weston J, Mukherjee S, Chapelle O, et al. Feature selection for SVMs[M]. Advances in Neural Information Processing Systems. vol. 13. Cambridge, Massachusetts: MIT Press, 2000