基于过采样技术和随机森林的不平衡微阵列数据分类方法研究

于化龙[1] 高 尚[1] 赵 靖[2] 秦 斌[1]
(江苏科技大学计算机科学与工程学院 镇江 212003)[1]
(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)[2]

摘 要 近年来,应用DNA微阵列技术对疾病,尤其是癌症进行诊断,已逐渐成为生物信息学领域的研究热点之一。对比其它的数据载体,微阵列数据通常具有一些独有的特点。针对微阵列数据样本分布不平衡这一特点,提出了一种基于概率分布的过采样技术,通过该技术可以为少数类建立一些合理的伪样本,从而使各类的样本数达到均衡,然后使用随机森林分类器对其进行分类。该方法的有效性和可行性已经在两个标准的微阵列数据集上得到了验证。实验结果显示,与传统的方法相比,该方法可以获得更好的分类性能。

关键词 微阵列数据,样本分布不平衡,过采样技术,概率分布,随机森林
中图法分类号 TP391 **文献标识码** A

Classification for Imbalanced Microarray Data Based on Oversampling Technology and Random Forest

YU Hua-long[1] GAO Shang[1] ZHAO Jing[2] QIN Bin[1]
(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)[1]
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)[2]

Abstract In recent years, applying DNA microarray technology to diagnose for disease, especially for cancer, has been becoming one of hot topics in bioinformatics. In contrast with many other data carriers, microarray data generally holds some unique characteristics. A novel oversampling technology based on probability distribution was proposed to solve the problem brought by the characteristic of sample distribution imbalance of microarray data. By this technology, some reasonable pseudo samples would be created for the minority class to guarantee the balance between two classes. Then we used random forest to classify the samples belonging to different classes. Its effectiveness and feasibility were verified on two benchmark microarray datasets. Experimental results show that the proposed method can obtain better classification performance, compared with some traditional approaches.

Keywords Microarray data, Sample distribution imbalance, Oversampling technology, Probability distribution, Random forest

1 引言

近年来,DNA微阵列已经逐渐发展成为后基因组时代最重要的分子生物学技术之一,并且在多个领域得到了广泛的应用,尤其是在癌症分类领域,这项技术更是具有得天独厚的优势。与其它数据载体不同,DNA微阵列数据通常具有以下一些特点:高维数、小样本、高噪声、高冗余和样本分布不平衡等[1],这些特点为相应的分析方法和工具的开发带来了极大的挑战。目前,该领域的研究人员更多地关注其前4个特点,即如何有效地排除数据中的噪声和冗余信息,从而挑选出那些与疾病密切相关的特征基因[2-4],以及如何使用少量的样本在高维的空间上建立可靠的分类器[5,6],从而提高分类的精度,却忽略了其样本分布通常是不平衡的(即不同类别的样本数具有较大差别)这一特性。

事实上,在机器学习和数据挖掘领域,不平衡数据的分类问题早已成为研究的热点。在2000年的AAAI国际会议和2003年的ICML国际会议上,分别设立了相应的论坛,对该问题进行了专题讨论。在2005年的ICDM国际会议上,该问题更是被列为了数据挖掘领域的十大挑战性难题之一[7]。目前,研究人员围绕该问题已经提出了大量的方法,其主要可以归纳为以下两类:采样的方法和代价敏感学习方法[8]。前者又可以分为过采样和欠采样两类,它们分别通过为正类(少数类)添加样本和从负类(多数类)中删除部分样本来达到平衡的目的。代价敏感学习方法则赋予两类样本以不同的误分类代价,从而可有效地训练出无偏的分类器。

在微阵列数据分类领域,Yang等[9]最早注意到了这一问

到稿日期:2011-06-01 返修日期:2011-08-03 本文受国家自然科学基金(60873036),国家教育部博士点基金新教师项目(20070217051),江苏科技大学引进人才科研启动项目(35301002)资助。

于化龙(1982-),男,博士,讲师,CCF会员,主要研究领域为生物信息学、模式识别等,E-mail:yuhualong@just.edu.cn;高 尚(1972-),男,博士,教授,主要研究领域为模式识别;赵 靖(1972-),女,博士,教授,主要研究领域为机器学习、可信计算与移动计算;秦 斌(1981-),男,硕士,讲师,主要研究领域为模式识别与图像处理等。

题，并提出了两种专门针对不平衡数据集的特征基因评价方法。与传统分类方法相比，其具有更强的鲁棒性。然而，Yang 的方法仍然是以分类精度作为评价指标的，因而缺乏足够的可信性。Li 等[10]注意到了它的这一缺点，结合 Easy Ensemble 算法[8]和 SVM-RFE 算法[2]的思想提出了两种嵌入式特征基因选择算法，获得了更好的分类性能。文献[11]首先采用混合采样的策略来构建平衡样本集，然后使用 ReliefF 算法去筛选差异表达基因，最后通过 5 个不平衡微阵列数据集验证了其算法的有效性。以上几类算法虽然有效，但是它们有一个共同的特点，那就是都以选取特征基因为目标，而忽略了分类任务的本质。

有鉴于此，本文结合微阵列数据高维小样本的特点提出了一种新颖的过采样方法：基于概率分布的过采样方法。本方法根据少数类样本在每个基因上的实际分布情况为其随机地建立一些伪样本。新建立的伪样本既符合原始样本的概率分布，又具有一定的随机性，可有效地提升数据集的质量。此外，本文采用随机森林(RF，Random Forest)作为分类器，这主要是出于以下考虑：作为一种集成分类器，随机森林采用 Bootstrap 技术选择样本，构建基分类器，并通过各基分类器投票的方式来决定测试样本的类属，其本身就可以从一定程度上减轻不平衡数据所带来的影响；另外，随机森林在对样本进行分类的同时，会对特征的重要性做出相应的评估[12]。最后，采用 3 折交叉验证的方法，在两个标准的微阵列数据集上验证了本文方法的有效性和可行性。

2 基于概率分布的过采样方法

采样的方法是处理不平衡问题最常用的方法之一，主要分为过采样和欠采样两类。前者通过为少数类增加一些合理的伪样本来达到数据集的平衡，而后者则更关注多数类中的噪声和冗余数据，通过排除一些无关紧要的多数类样本来平衡数据集[8]。考虑到微阵列数据小样本的特性，本文将采用过采样的方法来获取平衡数据集。

增加少数类样本最简单与常用的一种方式是随机过采样，即通过复制正样本使两类数据趋于平衡。此方法具有简单宜用、计算复杂度低等优点，但是增加了数据过拟合的可能性，因而效果往往较差。为此，Chawla 等[13]提出了一种智能过采样方法：SMOTE (Synthetic Minority Oversampling TEchnique)。该方法不是简单地复制已有的样本，而是通过插值的策略从已有的样本中生成新的并不存在的样本。其主要步骤如下：a)以某个少数类样本为参考点，在其同类样本中找到 k 个近邻；b)在其近邻中随机选取一个；c)连接参考样本与随机选取出的近邻样本，新样本对应连接线上某个随机选取的点。SMOTE 方法有效地降低了数据过拟合的可能性，与随机过采样方法相比，其性能也有了明显的提高[13]。近年来，研究人员根据各应用领域的特点分别对 SMOTE 算法做出了一定的改进，其分类的性能也有了进一步的提升[14,15]。

考虑到微阵列数据具有高维小样本、高噪声、高冗余等一些特点，直接对其采用 SMOTE 算法可能会降低分类的性能，故本文提出了一种新颖的过采样方法：基于概率分布的过采样方法。该方法通过计算少数类样本在每一个基因上的概率分布情况来为其单独生成伪样本。这种做法既保证了概率分布的前后一致性，又增加了伪样本的随机性，使新生成的样本集更加合理。

我们最初的设想是通过计算少数类样本在每个基因上的均值和方差来评估其正态分布的位置。然而，由于样本数过少，这种评估往往是不准确的，甚至会引入大量的噪声信息，因此改用了非参数的概率密度估计方法。

使用本方法，首先需要将基因的取值范围进行离散化处理，即将其划分为多个等宽且互不相交的取值区间。如当基因的取值范围为[0,1]，离散化区间宽度为 0.1 时，则应将该基因划分为[0,0.1)，[0.1,0.2)，…，[0.9,1]，共 10 个离散化区间。然后，通过统计每个区间上出现的样本次数估计出实际的概率分布情况，如图 1 所示。

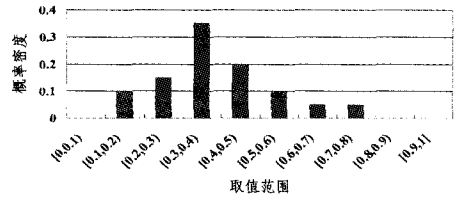


图 1 概率分布评估示意图

根据图 1，可以清晰地看出少数类样本的表达值在某个具体基因上的实际分布情况。事实上，我们希望新生成的伪样本也能服从这种分布。具体做法如下：a)由计算机自动生成一个(0,1)之间的随机数。b)根据各取值区间对应的概率密度分别计算出相应的随机数区间。以图 1 为例，由于[0.1,0.2)这一取值区间的概率密度为 0.1，因此其对应的随机数区间为(0,0.1]，相应的[0.2,0.3)取值区间所对应的随机数区间为(0.1,0.25]。依此类推，概率密度越大的取值区间，其所对应的随机数区间也越大，生成的伪样本也就有更大的概率取到该区间的值。c)找出随机数所对应的取值区间。d)取该区间中的某个随机数作为伪样本在此基因上的表达值。使用本方法，既可以保证新生成的伪样本符合原始的概率分布，又增加了一定的随机性，有效地提高了数据集的质量，同时与我们设计的初衷相吻合。

结合前面的语言描述，给出了基于概率分布的过采样方法相应的伪代码描述，如图 2 所示。

```
Input: Training sample set S, Number of genes N, Number of discretization intervals D.
1. Extract samples belonging to Majority class(save to S_Maj) and Minority class(save to S_Min), respectively.
2. For each iteration i=1,…,N
3. {
4. Discretize gene i into D regions with equal interval;
5. Evaluate probability distribution for each interval by the samples in S_Min;
6. Compute the corresponding random number region for each value interval;
7. }
8. For each iteration i=1,…,|S_Maj − S_Min|
9. {
10.    For each iteration j=1,…,N
11.    {
12.    Generate a random number R∈(0,1);
13.    Find the corresponding value interval to R;
14.    Generate a random number in the interval and save as S'_ij;
15.    }
16. }
Output: S+S'
```

图 2 基于概率分布的过采样方法的伪代码描述

在本算法中，离散化区间宽度的选取对最终分类性能的影响较大。为此，我们通过实验对该参数的性能进行了测试，发现当该参数较小或较大时，分类性能均不理想。究其原因，不难推断，当宽度过小时，新生成的伪样本与原始样本过于接近，因而缺乏足够的随机性；而当宽度过大时，尽管伪样本的随机性可以保证，但是其并不符合原始的概率分布，同样会降低分类的性能。当宽度为 0.1 时，二者之间基本上可以达到平衡，保证了最终的分类质量。

3 随机森林

随机森林是 Breiman 于 2001 年提出的一种集成分类器[12]，它与其它集成分类方法最大的不同在于只能使用决策树作为基分类器。其基本思想是通过 Bootstrap 技术生成大量训练样本子集，然后使用每个样本子集单独训练一个基分类器。在这一操作上它与 Bagging 的思想是一致的。为了进一步增加基分类器之间的差异度，随机森林在决策树的每个分支节点处，从全部 N 个特征中随机选择 n 个作为备选（$n \ll N$），通常设 $n=\sqrt{N}$，然后根据节点不纯度最小的原则，从这 n 个特征中选择一个进行节点生长。在树的生长过程中，通常不进行剪枝操作。当用随机森林对新样本进行判别与分类时，分类结果按基分类器的投票多少而定。

随机森林具有很多优点，如适合高维小样本的分类问题、不会出现过拟合、分类性能稳定、分类精度高、在分类的同时可以评估特征的重要性、分类速度快且容易实现并行化等。更为重要的是，由于采用了集成投票的方式，因此其对于样本不平衡分类问题具有先天的优势。

在微阵列数据分类领域，随机森林也已经得到了应用。Uriarte 等[16]采用 9 个微阵列数据集测试了随机森林的分类性能，并与很多传统的分类方法（包括 k 近邻分类器、支持向量机以及线性判别分析等）进行了比较，得出了随机森林分类性能更优的结论。故本文采用随机森林作为基准的分类器。

4 本文方法

结合上述基于概率分布的过采样方法和随机森林分类器，给出了一种不平衡微阵列数据的分类方法，其算法流程如图 3 所示。

```
Input: Training sample set S, Testing sample set T, Size of Random
       Forest M.
1. Construct balanced training sample set S+S' using Oversampling
   method based on probability distribution.
2. For each iteration i=1,…,M
3. {
4.   Select Si=|S+S'| samples by Bootstrap strategy;
5.   Use Si to train a decision tree classifier Ci;
6. }
7. For each iteration i=1,…,|T|
8. {
9.   For each iteration j=1,…,M
10.  {
11.    Use Cj to classify for testing sample Ti;
12.  }
13.  Classify for testing sample Ti by majority voting;
14. }
Output: Acc which is the classification accuracy for T.
```

图 3 本文方法的伪代码描述

5 实验结果及分析

5.1 数据集

本文采用两个标准的不平衡微阵列数据集对所提方法的有效性进行验证，其分别是 Colon 数据集[17]和 DLBCL 数据集[18]。这两个数据集可以从以下网址下载得到：http://datam.i2r.a-star.edu.sg/datasets/krbd/。关于这两个数据集的描述详见表 1。

表 1 本文所使用的微阵列数据集

数据集	基因数	样本数	类别数	不平衡度
Colon	2000	62(22:40)	2	1.82
DLBCL	7129	77(19:58)	2	3.05

5.2 性能评价指标

在不平衡数据分类任务中，仅仅使用整体分类精度作为其性能优劣的评价指标往往是不合理的。如某个包含 100 个样本的数据集，其中 99 个样本属于第一类，另一类只有 1 个样本，那么将所有样本都分为第一类，也可以获得 99%的分类精度，但不可否认的是，其对应的分类性能很差。因此，除整体分类精度（Acc，Accuracy）以外，本文也使用了其它一些不平衡分类问题中常用的评价指标，如 a)真正率（TPR，True Positive Rate），或称灵敏度（Sensitivity）、召回率（Recall）等；b)真负率（TNR，True Negative Rate），或称特指度（Specificity）；c)F-measure 以及 d)G-mean 等。这些评价指标的计算都需要用到表 2 所列的混淆矩阵。

表 2 二分类问题对应的混淆矩阵

	预测的正类	预测的负类
实际的正类	TP(正确预测的正样本数)	FN(错误预测的正样本数)
实际的负类	FP(错误预测的负样本数)	TN(正确预测的负样本数)

基于表 2 中所给出的混淆矩阵，各评价指标相应的计算公式如下：

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$TPR=Sensitivity=Recall=\frac{TP}{TP+FN} \tag{2}$$

$$TNR=Specificity=\frac{TN}{TN+FP} \tag{3}$$

$$Precision=\frac{TP}{TP+FP} \tag{4}$$

$$F\text{-}measure=\frac{2*Precision*Recall}{Precision+Recall} \tag{5}$$

$$G\text{-}mean=\sqrt{TPR\times TNR} \tag{6}$$

需要注意的是，本文中少数类样本被标记为正类，多数类样本则对应于负类。

5.3 结果与讨论

为了便于对样本进行分类，首先应对微阵列数据进行标准化处理，使各个基因处于同一量纲之下。本文的标准化方法如下：

$$g_{ij}^{*}=(g_{ij}-g_{\min})/(g_{\max}-g_{\min}) \tag{7}$$

式中，g_{ij} 与 g_{ij}^{*} 分别为第 j 个样本在第 i 个基因上的原始与变换后的表达值，$g_{\max}$ 与 $g_{\min}$ 分别表示第 i 个基因在所有样本上的最大与最小值。经过以上操作，每个基因的值将取在[0,1]范围内，且可以保证数据之间的原始比例关系不变。

本文的实验是在 Matlab7.0 环境下进行的，其中离散化区间宽度设为 0.1，同时参照文献[16]的参数设置，令随机森

林规模为 1000，决策树每个分支节点的备选基因数为√总基因数。同时，对本文方法与标准的随机森林、随机过采样和 SMOTE 过采样方法分别进行了性能比较，其中 SMOTE 方法的近邻数取为 3。分类性能评价方法为 3 折交叉验证，每个实验独立运行 10 次，各种方法的平均分类性能如表 3 和表 4 所列。

表 3　各类方法在 Colon 数据集上的平均分类性能(%)

分类方法	性能评价指标				
	Acc	TPR	TNR	F-measure	G-mean
标准随机森林	77.74	58.18	88.50	64.84	71.64
随机过采样+随机森林	79.03	63.64	87.50	68.29	74.62
SMOTE+随机森林	81.29	74.54	85.00	73.69	79.46
本文方法	82.90	76.36	86.50	76.02	81.26

表 4　各类方法在 DLBCL 数据集上的平均分类性能(%)

分类方法	性能评价指标				
	Acc	TPR	TNR	F-measure	G-mean
标准随机森林	86.25	54.74	96.55	66.12	72.48
随机过采样+随机森林	88.57	65.79	95.69	71.26	77.14
SMOTE+随机森林	90.39	71.17	96.72	76.22	82.54
本文方法	89.61	75.79	94.14	78.12	84.34

从表 3 和表 4 可以得出以下结论：

a)过采样技术有助于提高不平衡数据集的分类性能，这种提升不仅体现在 TPR，F-measure 以及 G-mean 等一些不平衡分类问题特有的评价指标上，而且包括总的分类精度。

b)相比其它指标，在大多数情况下真负率会随着样本集的平衡而略有下降，这种下降与真正率的提高是相对应的，表明训练所得到的分类面与真实的分类面更加接近。

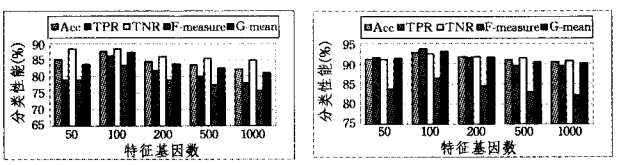
c)与随机过采样和 SMOTE 方法相比，本文所提出的基于概率分布的过采样方法在微阵列数据的分类问题上具有更好的性能，表明采用本方法能够生成更为合理的伪样本，同时验证了其有效性。

另外，我们测试了特征选择技术对本文方法性能的影响。作为最简单有效的特征选择方法之一，信噪比(SNR，Signal-Noise Ratio)方法[3]被用来选择那些与分类任务密切相关的差异表达基因。其计算公式如下：

$$SNR(g_i)=|\mu_1-\mu_2|/(\sigma_1+\sigma_2) \tag{8}$$

式中，μ_1 与 μ_2 分别表示基因 g_i 在两类样本上的平均值，而 σ_1 和 σ_2 则是它们的标准差。一个基因的信噪比越大，表明该基因与分类的关系越密切。

分别选用 50，100，200，500 和 1000 个特征基因测试了本文方法的性能，其它的参数与前文保持一致，其结果如图 4 所示。

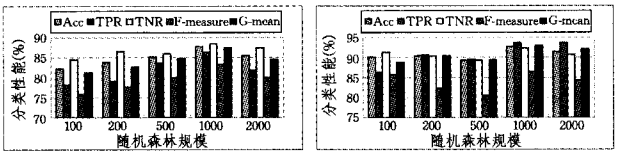


(a) Colon 数据集　　(b) DLBCL 数据集

图 4　本文方法基于不同特征基因数的分类性能

从图 4 中可以看出，特征选择有助于提升本文方法的分类性能。另外，从该图中可以观察到一个规律，即分类性能曲线随着特征基因的增加呈现出先上升后下降的一种走势。考虑其主要原因在于：当特征基因过少时，尽管森林中每棵决策树的质量可以得到保障，但是个体之间的差异度会有大幅的下降，导致分类性能降低；而当特征基因过多时，尽管决策树间的差异度有所增大，但其个体质量也会有较大程度的降低，同样会损害分类的性能。因此，需要在二者之间找到一个平衡。图 4 表明，当选取 100 个特征基因时，其分类性能基本可以达到最优。

接下来，我们测试了随机森林规模对本文方法性能的影响。决策树的个数分别取 100，200，500，1000 和 2000，特征基因数固定为 100，其它参数保持不变，其结果如图 5 所示。



(a) Colon 数据集　　(b) DLBCL 数据集

图 5　随机森林规模对本文方法分类性能的影响

从图 5 中可以看出，随着决策森林规模的增长，分类性能尽管可能会出现一定的波动，但在整体上还是呈现出增长的趋势。因此，在使用本文方法时，要根据实际情况力求在分类性能与时空复杂度之间寻找到一个平衡点。

结束语　针对不平衡微阵列数据的分类问题，本文结合过采样技术与随机森林分类器提出了一种新的分类方法。与前人的过采样技术不同，本文采用概率分布作为生成伪样本的准则，可以同时保证新生成样本的随机性与合理性。大量实验结果已经表明了本文方法的有效性和可行性。本文方法也为解决高维小样本不平衡数据分类问题提供了一个新的尝试。

参 考 文 献

[1] 于化龙，顾国昌，赵靖，等. 基于 DNA 微阵列数据的癌症分类问题研究进展[J]. 计算机科学，2010，37(10)：16-22

[2] Guyon I，Weston J，Barnhill S，et al. Gene Selection for Cancer Classification Using Support Vector Machines [J]. Machine Learning，2002，46(1-3)：389-422

[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(5439)：531-537

[4] Lee C P，Leu Y. A novel hybrid feature selection method for microarray data analysis [J]. Applied Soft Computing，2011，11(1)：208-213

[5] Paliwal K K，Sharma A. Improved direct LDA and its application to DNA microarray gene expression data[J]. Pattern Recognition letters，2010，31(16)：2489-2492

[6] 于化龙，顾国昌，刘海波，等. 基于相关性分析的微阵列数据集成分类研究[J]. 计算机研究与发展，2010，47(2)：328-335

[7] 翟云，杨炳儒，曲武. 不平衡类数据挖掘研究综述[J]. 计算机科学，2010，37(10)：27-32

[8] Liu X Y，Wu J X，Zhou Z H. Exploratory Under-sampling for Class-Imbalance Learning[C]// Proceedings of the Sixth International Conference on Data Mining. Hongkong：IEEE Press，2006：965-969

[9] Yang K，Cai Z，Li J，et al. A stable gene selection in microarray data analysis[J]. BMC Bioinformatics，2006，7：228

[10] Li G Z，Meng H H，Ni J. Embedded Gene Selection for Imbalanced Microarray Data Analysis[C]// Proceedings of Third International Multi-symposiums on Computer and Computational Sciences. Shanghai：IEEE Press，2008：17-24

[11] Kamal A H M, Zhu X Q, Narayanan R. Gene Selection for Microarray Expression Data with Imbalanced Sample Distributions [C] // Proceedings of 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing. Shanghai: IEEE Press, 2009: 3-9

[12] Breiman L. Random Forest[J]. Machine Learning, 2001, 45: 5-32

[13] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357

[14] Han H, Wang W Y, Mao B H. Borderline-smote: a new over-sampling method in imbalanced data sets learning[J]. Lecture Notes in Computer Science, 2005, 3644: 878-887

[15] Liu A, Ghosh J, Martin C. Generative oversampling for mining imbalanced datasets[C] // Proceedings of the Seventh International Conference on Data Mining. Las Vegas, Nevada, USA: CSREA Press, 2007: 66-72

[16] Uriarte R D, Andres S A D. Gene selection and classification of microarray data using random forest[J]. BMC Bioinformatics, 2006, 7: 3

[17] Alon U, Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by Oligonucleotide array[J]. Proceedings of the National Academy of Sciences, 1999, 96(12): 6745-6750

[18] Shipp M A, Ross K N, Tamayo P, et al. Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene Expression Profiling and Supervised Machine Learning[J]. Nature Medicine, 2002, 8(1): 68-74

(上接第 189 页)

$$\begin{cases} f_1(x)=x_1^2-0.25428722-0.18324757x_3^2x_4^2x_9^2 \\ f_2(x)=x_2^2-0.37842197-0.16275499x_1^2x_6^2x_{10}^2 \\ f_3(x)=x_3^2-0.27162577-0.16955071x_1^2x_2^2x_{10}^2 \\ f_4(x)=x_4^2-0.19807914-0.15585316x_1^2x_6^2x_7^2 \\ f_5(x)=x_5^2-0.44166728-0.19950920x_3^2x_6^2x_7^2 \\ f_6(x)=x_6^2-0.14654113-0.18922793x_5^2x_8^2x_{10}^2 \\ f_7(x)=x_7^2-0.42937161-0.21180486x_2^2x_5^2x_8^2 \\ f_8(x)=x_8^2-0.07056438-0.17081208x_1^2x_6^2x_7^2 \\ f_9(x)=x_9^2-0.34504906-0.19612740x_6^2x_8^2x_{10}^2 \\ f_{10}(x)=x_{10}^2-0.42651102-0.21466544x_1^2x_4^2x_8^2 \end{cases} \tag{18}$$

由表 2 可以看出，本算法能 100%求出方程(3)、(4)的解，50 次解对应的最大方程值的平均值精度均很高。由此可以看出本文所提的自适应正交差分进化算法是一种求解非线性方程组的有效算法。

表 2　问题 3、4 的收敛情况

问题	仿真次数	本算法成功率	本算法最大绝对值	平均求解时间
问题 3	50	1	1.3197e-008	11.7
问题 4	50	1	2.3193e-008	17.6

结束语　本文分析了极大熵收敛性对函数形式具有一定的依赖性，而修正极大熵(凝聚态函数)收敛性不依赖于函数形式的特点，通过代理约束将方程组转化为不可微的无约束优化问题，然后利用修正极大熵将其转化为可微的无约束优化问题。同时，针对文献[6]自适应正交交叉算子的不足之处，设计了新的自适应正交交叉算子，该算子不受问题收敛精度的影响，自适应调节相似度的下界，以避免个别相似度极大的基因作为因素的分割位置而影响算法的收敛速度，从而提出了自适应正交交叉差分进化算法，用于求解方程组对应的修正极大熵函数。通过求解 4 个不同特点的方程组，并比较其他算法求解结果，表明了该算法求解非线性方程组的有效性。

参 考 文 献

[1] 李兴斯. 解非线性规划的凝聚函数法[J]. 中国科学：A 辑，1991，12：1283-1288

[2] 颜世建. 关于解非线性规划的一个修正凝聚函数法的注记[J]. 南京师大学报，2002，25(2)：93-96

[3] Yang Yan, Zhou Yong-quan, Gong Qiao-qiao. Hybrid Artificial Glowworm Swarm Optimization Algorithm for Solving System of Nonlinear Equations [J]. Journal of Computational Information Systems, 2010, 6(10): 3431-3438

[4] 雍龙泉. 基于极大熵微粒群混合算法的非线性方程组求解[J]. 海军工程大学学报，2009，21(3)：28-32

[5] 陈海霞，杨铁贵. 基于极大熵差分进化混合算法求解非线性方程组[J]. 计算机应用研究，2010，27(6)：2028-2030

[6] 江中央，蔡自兴，王勇. 求解全局优化问题的混合自适应正交遗传算法[J]. 软件学报，2010，21(6)：1296-1307

[7] 田巧玉，古钟璧，周新志. 基于混合遗传算法求解非线性方程组[J]. 计算机技术与发展，2007，17(3)：10-12

[8] Hentenryck P V, Mcallester D, Kapur D. Solving polynomial systems using a branch and prune approach [J]. SIAM J. NUMER. ANAL, 1997, 34(2): 797-827

[9] Sonia K. Extension of the Lanczos and CGS methods to systems of nonlinear equations [J]. Journal of Computational and Applied Mathematics, 1996, 69: 181-190

[10] Leung Y-W, Wang Yu-ping. An Orthogonal Genetic Algorithm with Quantization for Global Numerical Optimization [J]. IEEE Transactions on Evolutionary Computation, 2001, 5(1): 41-53

[11] Wang Yong, Cai Zi-xing, Zhang Qing-fu. Enhancing the Exploration Ability of Differential Evolution through Orthogonal Crossover [R]. working report. 2011

[12] 罗亚中，袁端才，唐国金. 求解非线性方程组的混合遗传算法[J]. 计算力学学报，2005，22(1)：109-114

[13] Storn R, Price K. Differential Evolution-A Simple and Efficient Heuristic for global Optimization over Continuous Spaces [J]. Journal of Global Optimization, 1997, 11: 341-359

[14] 李耀辉，薛继伟，冯勇. 基于预处理和区间计算的非线性方程组实根求解[J]. 四川大学学报，2005，36(5)：86-93

[15] 段治健，杨永，马欣荣，等. 求解带状线性方程组的一种并行算法[J]. 计算机科学，2010，37(3)：242