·基础理论与方法·

应用广义多因子降维法分析数量性状的交互作用

陈卿　唐迅　胡永华

【导读】 介绍广义多因子降维法(GMDR)在交互作用分析,尤其是数量性状的基因-基因交互作用分析中的应用。文中简述GMDR的原理、基本步骤及其特点,并结合实例说明如何在研究中对GMDR进行应用。GMDR是无模型的交互作用分析方法,能够处理连续型结局变量,还可纳入协变量改善预测准确率,目前已成功应用于尼古丁依赖等疾病的研究。GMDR能够处理多种样本类型和结局变量类型,与其他连续变量交互作用分析方法相比具有一定优势。

【关键词】 广义多因子降维法;数量性状;交互作用

Detecting interaction for quantitative trait by generalized multifactor dimensionality reduction

CHEN Qing, TANG Xun, HU Yong-hua. Department of Epidemiology and Biostatistics, School of Public Health, Peking University, the Key Laboratory of Epidemiology Ministry of Education, Beijing 100191, China

Corresponding author: HU Yong-hua, Email: yhhu@bjmu.edu.cn

This work was supported by grants from the National Natural Science Foundation of China (No. 30671807, 30872173) *and the Specialized Research Fund for the Doctoral Program of Higher Education* (No. 20060001111).

【Introduction】 To introduce the application of generalized multifactor dimensionality reduction (GMDR) method for detecting interactions, especially gene-gene interactions for quantitative traits. Principles, basic steps as well as features of GMDR were discussed, illustrated with a practical research case. As an interaction analysis method, GMDR was model-free, available for studies on different outcome variables including continuous ones, and permitted adjustment for covariates to improve prediction accuracy. Evidences of its capacity had been supposed by research on different diseases, e.g. nicotine dependence. GMDR method was applicable to different types of samples and outcome variables, which was superior to other statistical approaches for continuous variables in some aspects.

【Key words】 Generalized multifactor dimensionality reduction; Quantitative trait; Interaction

在肿瘤、心血管病等复杂疾病中广泛存在着基因-基因交互作用和基因-环境交互作用,怎样发现这些交互作用是目前面临的一个问题。多因子降维法(multifactor dimensionality reduction, MDR)是近年来发展出的一种非参数、无需遗传模式的高阶交互作用分析方法[1],被成功应用在多发性硬化、2型糖尿病等疾病的病因研究中[2]。但由于该法不能用于数量性状的研究,为此Lou等[3]在2007年提出了一种基于MDR基本原理的扩展方法——广义多因子降维法(generalized multifactor dimensionality reduction, GMDR),又称基于计分的多因子降维法(score-based MDR)。该法可以通过将广义线性模型的概念引入到MDR中,使其不但能够分析连续变量,且能够纳入协变量,从而控制协变量引起的干扰,提高预测的准确度。其主要特点:①分析的基因表型和校正因素不限于离散型变量,也可以是连续型变量;②可应用于多种数据结构(病例对照资料、人群随机抽样样本或其他类型样本);③结合GMDR software软件,可识别多个位点或环境因素之间的交互作用。

基本原理

GMDR是对原始MDR的扩展,其基本原理包括计分统计量(score statistic)和交叉验证(cross validation)。交叉验证是用于评估GMDR得出的模

DOI: 10.3760/cma.j.issn.0254-6450.2010.08.024
基金项目:国家自然科学基金(30671807, 30872173);高等学校博士学科点专项科研基金(20060001111)
作者单位:100191 北京大学公共卫生学院流行病与卫生统计学系 教育部流行病学重点实验室
通信作者:胡永华, Email: yhhu@bjmu.edu.cn

型统计学意义的手段，其基本思想是：先利用部分数据（从全部数据中随机抽取）得出模型，再在剩余的数据中加以检验；并且多次重复这一过程以避免数据的机会性划分对结果造成的影响[4]。计分统计量是GMDR新引入的概念，也是与MDR的区别所在。

1. 计分统计量：在构建计分统计量之前，首先需要通过某种恰当的函数将暴露变量，包括研究者关注的暴露变量（研究变量）和其他协变量，与结局变量联结起来。运用广义线性模型的方法，对于指数族分布（如正态分布、Poisson分布、伯努利分布等）的结局变量可通过如下形式表示

$$l(\mu_i) = \alpha + X_i^T\beta + Z_i^T\gamma \quad (1)$$

式中i表示第i个研究对象，μ_i是结局变量的期望值，$l(\mu_i)$是对μ_i进行的某种转换，即关联函数。α为截距，X_i为研究变量间交互作用的向量，Z_i为协变量的向量，β和γ为两者相应的参数向量。

对于不同分布类型的结局变量，$l(\mu_i)$有所不同。例如对于符合伯努利分布的二分类结局变量，可以建立logit关联函数

$$l(\mu_i) = \log\left[\frac{\mu_i}{(1-\mu_i)}\right] \quad (2)$$

而对符合正态分布的连续变量，可不进行关联函数的转换。

对于个体i的结局变量观察值y_i[即是说μ_i是y_i的期望值，$E(y_i)=\mu_i$]，其对数前瞻似然（log-prospective likelihood）函数可表示为

$$\log L(Y \mid \Omega) = \sum_{i=1}^{n} \{ y_i l(\mu_i) - f \mid l(\mu_i) \mid \} \quad (3)$$

其中$\Omega=(\alpha,\beta,\delta)$，$f=\mid l(\mu_i) \mid$是$l(\mu_i)$的函数，具有如下性质：$\partial f \mid l(\mu_i) \mid / \partial l(\mu_i) = \mu_i$。求式(3)的一阶偏导数，并设定$\beta=0$，即得到残差计分向量（residual score vector）：

$$S_\beta(\hat{\alpha}_0, \beta=0, \hat{\gamma}_0) = [S_{\beta_j}(\hat{\alpha}_0, \beta=0, \hat{\gamma}_0)] \quad (4)$$

其中

$$S_{\beta_j}(\hat{\alpha}_0, \beta=0, \hat{\gamma}_0) = \sum_{i=1}^{n} x_{ij}(y_i - \hat{\mu}_i) \quad (5)$$

式中j表示第j个交互作用变量，$\hat{\mu}_i$为$l^{-1}(\hat{\alpha}_0 + Z_i^T\hat{\gamma}_0)$，而$\hat{\alpha}_0$和$\hat{\gamma}_0$为$\beta=0$（即不存在研究变量的交互作用）时的最大似然估计值。对于每个个体，将其统计量正态化，即得到

$$S_i^T = \sum_j \frac{x_{ij}(y_i - \hat{\mu}_i)}{\sqrt{\mathrm{Var}(y_i)}} \quad (6)$$

此即GMDR方法中通常使用的每个个体的计分统计量。

2. 基本步骤：GMDR的分析过程也是采用降维策略来发现交互作用。以基因-基因交互作用分析为例，基本步骤如图1所示。

第1步，将数据随机分为若干等份（图1中为10等份），取其中一份为检验样本，其余为训练样本，以便进行交叉验证。第2步，确定探索几因素交互作用模型，即需要从研究变量中选出多少个（设为n）进行组合。第3步，根据选出的研究变量划分出单元格（图1中以$n=2$位点的基因型交互作用分析举例，将形成$3^2=9$个单元格），而训练样本中的研究对象也将据此被分配到各个单元格中。计算每个研究对象的计分（图1中各单元格内显示的分别的是正计分之和及负计分之和）。第4步，每个单元格将根据其内个体计分的均值是否超过某个设定标准（例如0），分别标记为“高危”或“低危”，单元格因此被分为2类，形成一维两水平的模型。第5步，GMDR将对所有可能的n位点组合都进行一次分析，例如研究共有4个位点的信息，那么2位点组合就有$C_4^2=6$个（图1中以位点1和位点4的组合举例）。在这些组合中选出对训练样本中个体状态（高危/低危）判断正确率最高的模型（图1中此模型为“1，3”）。第6步，利用检验样本评价上一步选出模型的预测准确率，即该模型将检验样本中的个体正确判断为高危或低危的比例。在第一步随机划分的样本中，每一份都将被选作检验样本一次，相应的其余样本作为训练样本，重复上述步骤。在此过程中某个模型被选中（第5步）的次数称为其交叉验证一致性（cross- validation consistency）。预测准确率由多次结果的平均值表示。模型的统计学检验采用符号检验（sign test），以交互验证过程中检验样本预测准确率高于0.5的次数为指标进行统计检验[5]。此外，也可通过置换检验（permutation test）对模型的预测准确率进行统计检验[6]。对于不同的n（单位点、2位点、3位点、……），可以重复上述过程，分别得出模型，在这n个模型中取预测准确率最高和/或交互验证一致性最大者作为最优模型。

3. GMDR软件分析：目前最新的GMDR软件（版本beta 0.7）是基于Java程序编写的源代码开放的免费软件（可在http://www.healthsystem.virginia.edu/internet/addiction-genomics/Software/免费下载）。该软件支持多种操作系统，可在图形用户界面（GUI）下进行操作，令使用过程非常简便。GMDR软件是在MDR软件的基础上开发的，其运行环境、操作界面以及对数据文件的格式要求与后者非常相

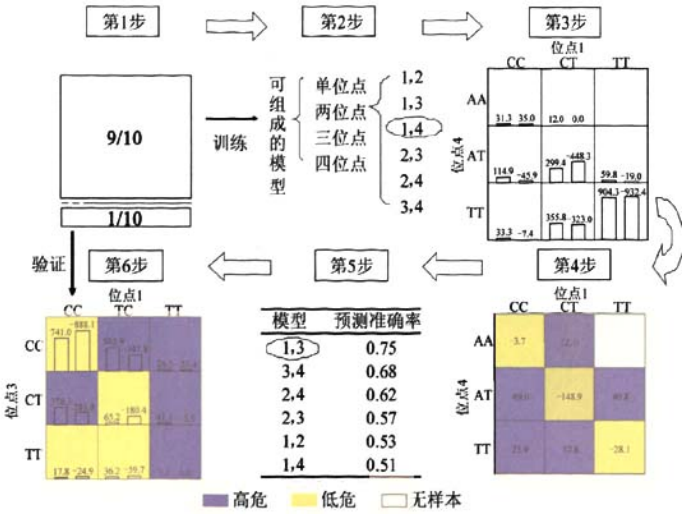

图1　GMDR基本步骤

近[4]。主要区别在于GMDR软件的主菜单中增加了“计分计算”(score calculation)标签，用户在需要时可以载入协变量和/或连续型结局变量，选定计分的计算方法(线性回归/logistic回归，前者适用于连续型结局变量，后者适用于分类结局变量)，从而生成计分。如果用户已经自行计算了计分统计量并按要求格式生成了数据文件，也可以通过“分析”(analysis)标签中新增加的“load scorefile”和“view scorefile”选项来载入和阅读此文件。

(1)文件类型和格式要求：与其选项相对应，GMDR能够识别的文件包括三种，分别为标记文件(marker file)，协变量与表型文件(covariate & phenotype file)以及计分文件(score file)。三者均为文本文件(.txt)，其中标记文件的内容是研究变量和分类结局变量，其格式要求与MDR相同。协变量与表型文件的内容是协变量和连续型结局变量，文件的首行是变量名，各变量名之间用空格隔开，结局变量的变量名前应加“#”。计分文件的内容为用户自己定义的计分数据，首行也是空格分隔的变量名。需注意的是，如果研究的结局是连续变量，那么标记文件中的最后一列(二分类结局变量)将不被软件纳入计算过程，但是用户仍然需要设定这一列(可使用0和1随意赋值)，否则软件无法识别文件。

(2)结果输出与保存：GMDR输出的结果也与MDR相似，但在生成的图表中，每个单元格的颜色由单元格的计分决定，蓝色为计分超过设定值，黄色为未超过设定值，白色表示单元格内没有数据。单元格中的左侧条带表示正计分之和，右侧条带表示负计分之和。数据的分析结果和图形文件可以分别通过分析标签中的“save analysis”和“save”进行保存。前者保存为“.txt”格式后，能够直接打开或通过“load analysis”在GMDR中读取；图形文件可保存为“.bmp”等各种格式，用相应程序打开。

实例分析

Chan等[7]在2008年为研究哮喘的遗传学基础开展了一项病例对照研究，并在分析哮喘相关的数量性状——外周血嗜酸性粒细胞计数的基因-基因交互作用时应用了GMDR。研究者在香港某医院选取298例确诊慢性稳定性哮喘的中国儿童和年龄、性别与之匹配的175名对照，测定了8个基因的18个多态性位点(表1)以及外周血嗜酸性粒细胞计数等指标。

表1　研究中选择的多态性位点

染色体位置	候选基因	多态性位点
1q25.2-25.3	*PTGS2*	C-4231A、A-1195G、PTGS2.5209T→G、PTGS2.8473T→C
2q33	*CTLA4*	-1147CT、+49A/G、CT60、JO31、JO30、JO27_1
5q31-33	*IL13*	R130Q
5q31	*CD14*	C-159T
5q32-34	*ADRB2*	R16G、E27Q
11q13	*FCERIB*	Rsal_in2、Rsal_ex7
16p12	*IL4RA*	150V
16q13	*TARC*	C-431T

将所有473名儿童作为整体纳入GMDR进行分析，探索1位点至5位点的交互作用模型，并进行5000次置换检验。结果发现最优模型(表2中的两位点模型)分别包含了*TARC*基因和*FCERIB*基因的一个位点，预测误差为40.22%，交互验证一致性为9。同时研究者还使用广义线性模型对这一交互作用进行验证，结果同样有统计学意义(P=0.029)；而单独考察每个位点时，所有位点均不能进入方程，提示影响外周血嗜酸性粒细胞计数的是位点间的交互作用而非这些位点的主效应。

表2　使用GMDR方法分析外周血嗜酸性粒细胞计数得出的交互作用模型

位点数	最佳模型	交互验证一致性	预测误差*(%)	5000次置换检验的P值
1	R130Q	8	44.62	0.094
2	C-431T, RsaI_in2	9	40.22	0.009
3	C-431T, RsaI_in2, R16G	7	45.29	0.145
4	C-431T, RsaI_in2, R16G, -1195A/G	8	40.44	0.022
5	C-431T, RsaI_in2, R16G, -1195A/G, JO27_1	7	42.09	0.113

注：*预测误差=(1-预测准确率)×100%

讨 论

GMDR是对原始MDR方法的扩展，由于引入了计分统计量，GMDR能够处理连续型结局变量，并且纳入协变量，使其应用范围和预测准确率得到改善。Lou等[3]曾利用模拟数据检验GMDR应用于连续变量时发现交互作用的能力，结果提示GMDR不但适用于连续变量，而且能够处理数据存在2个以上潜在分组的情况（在模拟数据为三峰分布的情况下，也成功发现了交互作用）。通过比较还发现，同样使用GMDR，纳入协变量以后模型的预测准确率将优于或不逊于未纳入协变量的分析。当然，GMDR仍然可以处理二分类结局变量，事实上，在不纳入协变量的情况下使用GMDR分析二分类结局变量等同于MDR。GMDR还拓宽了数据结构的适用范围，不要求研究是病例对照设计，随机抽样或其他抽样方式选取的样本都可进行处理。另外，虽然GMDR主要关注研究变量之间的交互作用，但在需要时还可对关联函数和计分统计量进行扩展，将分析研究变量与协变量之间的交互作用也纳入分析的范畴[3]。正因为GMDR具有这些优点，目前已有多项研究应用该方法，并在尼古丁依赖[8]、哮喘[7,9]、2型糖尿病[10]、糖尿病肾病[11]、脑卒中[12]等研究领域中成功发现交互作用。

需要注意的是，虽然GMDR中引入了计分统计量，但并不强调这种方法必须以似然函数为基础，必要时也可使用其他统计量进行替代。因此GMDR与MDR一样，也可视为是无模型（model free）。而由于GMDR与MDR一样，只关注变异中的主要信号（得出的模型只将数据划分为一维两水平），所以相对于组合划分法（combinatorial partitioning method）和限制划分法（restricted partition method）等其他连续变量分析方法极大地降低了计算负担。但这一优点也有限度，当研究变量的个数N超过10个且探索所有1位点至N位点模型时计算负担仍会较大。此时如果对模型的位点数进行限制，例如只探索1～5位点模型，可以有效降低计算负担。

综上所述，GMDR对MDR进行了扩展，可用于连续型结局变量分析，能够纳入协变量以提高预测准确率，适用的数据结构也更加宽泛。而且GMDR是一种无模型的分析方法，其计分统计量可由其他适当的统计量进行替代，因此具有进一步改进的潜力。该方法在基因-基因交互作用和基因-环境交互作用的探索中具有广阔的应用前景。

参 考 文 献

[1] Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet, 2001, 69(1): 138-147.

[2] Tang X, Li N, Chen DF, et al. Recent advances in applications of multifactor dimensionality reduction for detecting gene-gene interactions. Chin J Epidemiol, 2007, 28(9): 918-921. (in Chinese)
唐迅，李娜，陈大方，等. 多因子降维法分析基因-基因交互作用的应用进展. 中华流行病学杂志，2007，28(9)：918-921.

[3] Lou XY, Chen GB, Yan L, et al. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet, 2007, 80(6): 1125-1137.

[4] Tang X, Li N, Hu YH. The application of multifactor dimensionality reduction for detecting gene-gene interactions. Chin J Epidemiol, 2006, 27(5): 437-441. (in Chinese)
唐迅，李娜，胡永华. 应用多因子降维法分析基因-基因交互作用. 中华流行病学杂志，2006，27(5)：437-441.

[5] Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics, 2003, 19(3): 376-382.

[6] Li N, Tang X, Chen DF, et al. Identification of gene-gene interactions related to the etiology of complex disease: a multifactor dimensionality reduction-genotype pedigree disequilibrium test. Chin J Epidemiol, 2007, 28(10): 1036-1040. (in Chinese)
李娜，唐迅，陈大方，等. 复杂疾病病因研究中基因间交互作用分析：基于基因型传递不平衡的多因子降维法. 中华流行病学杂志，2007，28(10)：1036-1040.

[7] Chan IHS, Tang NLS, Leung TF, et al. Study of gene-gene interactions for endophenotypic quantitative traits in Chinese asthmatic children. Allergy, 2008, 63(8): 1031-1039.

[8] Motsinger-Reif AA, Reif DM, Fanelli TJ, et al. A comparison of analytical methods for genetic association studies. Genet Epidemiol, 2008, 32(8): 767-778.

[9] Lee JH, Moore JH, Park SW, et al. Genetic interactions model among Eotaxin gene polymorphisms in asthma. J Hum Genet, 2008, 53(10): 867-875.

[10] Lin E, Pei D, Huang YJ, et al. Gene-gene interactions among genetic variants from obesity candidate genes for nonobese and obese populations in type 2 diabetes. Genet Test Mol Biomark, 2009, 13(4): 485-493.

[11] Wu LSH, Hsieh CH, Pei D, et al. Association and interaction analyses of genetic variants in ADIPOQ, ENPP1, GHSR, PPAR gamma and TCF7L2 genes for diabetic nephropathy in a Taiwanese population with type 2 diabetes. Nephrol Dial Transplant, 2009, 24(11): 3360-3366.

[12] Liu JH, Sun K, Bai YY, et al. Association of three-gene interaction among MTHFR, ALOX5AP and NOTCH3 with thrombotic stroke: a multicenter case-control study. Hum Genet, 2009, 125(5-6): 649-656.

（收稿日期：2009-12-23）
（本文编辑：张林东）