基于神经网络和粒子群算法的遗传位点与患病信息的关联性分析

李 杰[1] 李志强[2*] 刘 晓[1] 闫白鹭[2]
(北京化工大学 1. 经济管理学院; 2. 理学院, 北京 100029)

摘 要: 基于遗传疾病与某些遗传基因位点存在的较强关联性,并考虑到位点间存在交互作用的情形,提出了关联性最强的位点组合的筛选方法。将每个候选位点组合对应的基于神经网络的预报准确率作为评价标准,用粒子群算法(PSO)通过迭代逼近找出最优的位点组合,并与神经网络权重分析法进行比较。结果表明,由本文方法得到的位点组合预报精度更高,对患病情况有着较好的识别效果,可为遗传疾病诊断等提供参考方法。

关键词: 遗传位点; 交互作用; 粒子群算法(PSO); 神经网络
中图分类号: F064.1 **DOI**: 10.13543/j.bhxbzr.2018.01.016

引 言

大量研究表明,人体的许多表型性状差异以及对药物和疾病的易感性都可能与某些基因位点相关联,故选取恰当的方法来找寻出相关致病位点,对于疾病的治疗和预防具有重要意义。

目前关于遗传位点与患病信息关联性的研究成果较多,其中全基因组关联分析是研究患病信息和患病位点之间关联性的重要方法,这种方法通过对试验个体的健康状况和位点编码的统计关联分析来确定致病位点,从而发现遗传病或性状的遗传机理[1]。皮尔逊卡方检验[2]和 Logistic 回归模型[3-4]也是致病位点关联性分析中常用的方法,目的是剔除与患病信息无关或影响作用非常小的基因位点。为了避免模型设定的错误,部分学者提出利用神经网络模型等非参数的方法拟合数据并进行位点筛查[5-6],但神经网络筛选方法是根据对每个位点在神经网络中的权重贡献率大小来挑选的,并没有考虑到位点之间是否存在交互作用,因此有时不能准确反映位点与遗传疾病的关联强度。

实际上,如癌症、糖尿病等复杂疾病并不是由单个位点独立引起的,而是与多个遗传位点的联合作用有关,即位点之间存在交互作用。交互作用在疾病的认识和发展中扮演着极其重要的角色,如果没有考虑位点间的交互作用,就无法真实准确地描述位点与患病的效应[7-8]。探讨位点间交互作用对于提高复杂疾病的遗传解释度、构建复杂疾病的遗传风险评估模型、开发疾病诊疗个性化药物靶点并最终降低复杂疾病负担等方面均具有重要的理论和现实意义[9-10]。文献[11-12]通过 Logistic 回归模型研究了位点及位点交互作用对遗传疾病的影响,得到对遗传疾病影响较大的位点组合。但是 Logistic 回归模型对位点及位点交互作用作了较强的参数形式的假定,这在处理实际问题时可能会产生模型设定误差,从而导致错误的推断结果。

因此,针对位点之间可能存在交互作用的情形,本文提出新的方法筛选与遗传疾病关联性最强的位点组合。针对每个可能的候选位点组合,利用神经网络方法训练数据,并以预报准确率作为评价标准,利用粒子群算法(PSO)通过迭代计算逐步逼近与遗传疾病相关性最强的位点组合,该方法缩短了计算时间,避免了 Logistic 回归模型等方法对位点作用较强的参数形式的假定。

1 基于粒子群算法的位点组合选取

1.1 基于粒子群算法的迭代计算

本文将预报准确率作为评价标准,将全部给定的候选位点集合看作样本空间,把样本空间的所有非空子集作为候选的位点组合,根据给定的评价标

收稿日期: 2016-12-12
基金项目: 北京化工大学研究生教改项目(11120024018)
第一作者: 男,1993 年生,硕士生
*通讯联系人
E-mail: li-zhiqiang2000@163.com

准找出关联性最强的位点组合。其中预报准确率根据以下方法得到:首先将所有样本数据分为训练样本和预测样本,然后以相应位点的取值为输入、遗传疾病信息为输出,分别对某个指定的候选位点组合利用神经网络方法训练数据,最后基于预测样本进行预报并计算出相应的预测精度。

但实际应用中由于候选的全部基因位点数目巨大,导致寻找最优位点组合的计算强度大时间过长,因此本文利用粒子群算法,通过迭代逼近最终筛查出关联性最强的位点组合。粒子群中的每一个粒子均对应一个候选的位点组合,其包含的位点在全部候选位点中的位置与粒子中分量为 1 的位置相对应。以每个粒子对应的预测精度为适应度函数值,根据粒子群算法对粒子进行更新迭代直至收敛,从而得到给定评价标准下的最优粒子,即最优位点组合。

粒子群算法的原理为:在迭代搜寻的过程中,粒子通过向个体最优位置和群体最优位置学习的经验来搜索最优解,在搜寻过程中通过不断修正粒子的适应度、速度和位置来学习。正是由于每个粒子都在不断调整和其他较好位置粒子的差异,不断向前期找到的较好位置靠拢,才使得它们的收敛速度很快[13-14]。吕思晨[15]提出了将粒子群算法和遗传算法结合起来研究位点与疾病关联分析的算法,并用来寻找与疾病关联性较强的单个位点。

假定粒子群中有 n 个粒子,每个粒子均是分量为 0 或 1 的 D 维向量(D 为可选的位点个数),即 $S=(S_1,S_2,\cdots,S_n)$ 为粒子种群,其中 S_i 为第 i 个粒子。为了计算每个粒子的适应度,分别以粒子中所有取值为 1 的分量所对应的位点取值作为 BP 神经网络的输入值,则对预测样本的预报正确率可作为对应粒子的适应度函数值。在下一次迭代时,基于每个粒子的适应度值以及种群中最大的适应度值来调整其分量的飞行速度,然后基于速度修正粒子的选取概率,并重新生成一个新的具有 n 个粒子的种群。重复以上迭代步骤进行计算,直至收敛。

具体来说,粒子的速度决定粒子移动的方向和距离,在每次迭代过程中,第 i 个粒子的第 j 个分量根据个体适应度极值和群体适应度极值更新自身的速度,即:

$$v_{ij}^{k+1}=\omega v_{ij}^{k}+c_1r_1(P_{ij}^{k}-s_{ij}^{k})+c_2r_2(P_{gj}^{k}-s_{ij}^{k}) \quad (1)$$

其中,ω 为惯性权重;k 为当前迭代次数;c_1 和 c_2 是非负常数,称为加速度因子;r_1 和 r_2 是[0,1]上的随机数;$\boldsymbol{P}_i=(P_{i1},P_{i2},\cdots,P_{iD})^{\mathrm{T}}$ 为第 i 个个体适应度最优时对应的取值,表示迭代到第 k 次时第 i 个粒子的适应度最大取值(比较的是同一个粒子在整个迭代中的适应度值);$\boldsymbol{P}_g=(P_{g1},P_{g2},\cdots,P_{gD})^{\mathrm{T}}$ 为全局适应度最大取值,表示迭代到第 k 次时所有粒子的适应度最大值(比较的是所有粒子在整个迭代中的适应度值)。

根据第 i 个粒子中第 j 个分量的速度 v_{ij} 更新对应位置的选取概率 $f(v_{ij})$,基于选取概率来判定第 i 个粒子对应的第 j 个分量的取值 S_{ij}

$$S_{ij}=\begin{cases}1 & R<f(v_{ij})\\ 0 & \text{others}\end{cases} \quad (2)$$

其中 R 是服从[0,1]上均匀分布的随机数,函数 $f(v_{ij})=\dfrac{\exp(v_{ij})}{1+\exp(v_{ij})}$,$v_{ij}$ 不同,粒子的更新位置也不同,v_{ij} 越大,粒子在该位置取 1 的概率越大。为了能够迭代计算,在初次迭代时,每个粒子的每个分量的取值可利用等概率来选择。模拟结果显示,算法的结果不受初值的影响。

1.2 基于神经网络的适应度计算

假定输入某 l 个位点构成的位点组合为

$$\boldsymbol{X}_p=(x_1^p,x_2^p,\cdots,x_l^p)^{\mathrm{T}}$$

其中 $p=1,2,\cdots,W$(W 为训练样本总数)。神经网络第 p 个样本隐层第 j 个结点的输出值为

$$h_j^p=f\left(a_{0j}+\sum_{i=1}^{l}\omega_{ij}x_i^p\right),\ j=1,2,\cdots,L$$

式中,$f(\cdot)$ 为激活函数,其形式同式(2)中 $f(v_{ij})$,ω_{ij} 为输入层和隐层之间的连接权值,a_{0j} 为常数项,j 为隐层结点数。

BP 神经网络第 p 个样本输出层第 k 个结点值为

$$y_k^p=\sum_{j=1}^{L}\omega_{jk}h_j^p,k=1,2,\cdots,N \quad (3)$$

其中,ω_{jk} 为隐层到输出层的权值,h_j^p 为第 j 个隐层结点的输出值,N 为输出层结点数。

得到训练的网络结构后,输入预测样本并比较预测结果和真实结果,以位点组合对预测样本的预测准确率作为对应的粒子在粒子群算法中的适应度函数值。

1.3 算法构建

位点组合选取迭代计算步骤如下。

(1) 利用两点分布 $B(1,0.5)$ 生成 n 个粒子的每个分量,建立 BP 神经网络;对每个粒子,以粒子

中所有位置为 1 的位点作为输入值,以预测样本的正确率作为适应度函数评估各粒子,记录第 i 个粒子的个体最优值并作为当前粒子适应度,全局最优值为适应度值最大的粒子。

(2) 根据 PSO 算法的公式(1)和(2)更新 n 个粒子的速度和位置,产生新的一组粒子,建立 BP 神经网络;将每个粒子以粒子中所有位置为 1 的位点作为输入值,以预测样本的正确率为适应度函数来评估各粒子。比较当前第 i 个粒子和个体最优的适应度函数值,将其中具有较大适应度函数值的粒子作为第 i 个粒子的个体最优;比较所有粒子和全局最优的适应度函数值,将具有最大适应度函数值的粒子作为全局最优。

(3) 判断是否满足停止准则,若满足,则将全局最优输出,结束;若不满足,则返回步骤(2)。

(4) 重复步骤(1) ~ (3),可以得到多个位点组合。将出现在位点组合中次数最多的部分位点作为具有较好预测效果的位点组合。

2　实验及结果分析

2.1　数据来源

选取 2016 年研究生数学建模竞赛 B 题数据(http://gmcm.seu.edu.cn/01/1d/c12a285/page.htm),样本由 1000 位试验者的患病信息和 9445 个位点信息遗传信息构成。样本分为患病者和健康者,两者各占 50%,用 1 表示患病者、0 表示健康者。每位试验者对应 9445 个位点,位点信息由碱基 A, T, C, G 的不同组合来表示,用两个碱基的组合表示一个位点的信息,一个位点有 3 种不同编码。

因为样本所给的位点有 9445 个,直接实施本文算法会导致计算量非常大,计算时间过长,所以需要首先对与患病信息无关的基因位点初步筛除。本文选择皮尔逊卡方检验的 p 值和 Logistic 回归中单变量的 t 检验的 p 值作为统计相关性度量指标。筛选掉与遗传疾病信息完全独立的基因位点和本身作用很小的基因位点,然后从剩余的相关性较强的位点集合中分别利用本文方法和神经网络权重分析法筛查关联性最强的部分位点,并根据预报准确率对两种方法进行对比。

2.2　初步筛选步骤

2.2.1　皮尔逊卡方检验

皮尔逊卡方统计量[2]是用于检验实际分布与理论分布拟合优度指标,可以用于两个指标的独立性检验。以位点 rs2273298 位置为例,该位点取值为 AA、AG 和 GG,并分别赋值为 1,2,3;疾病信息取值为 0、1。令 t_{ij} 为位点取值 j 而疾病取值为 i 的个体的数量,则位点 rs2273298 与疾病信息的独立性检验的卡方值为

$$\chi^2 = \sum_{i,j} \frac{(t_{ij} - f_i q_j/1000)^2}{f_i q_j/1000} \tag{4}$$

其中, $f_i = \sum_{j=1}^{3} t_{ij}, i = 0,1$; $q_j = \sum_{i=1}^{2} t_{ij}, j = 1,2,3$。

当位点与疾病独立时,式(4)中统计量服从自由度为 2 的卡方分布,因此可计算出样本卡方值对应的 p 值。若 p 值很小,说明位点与疾病之间存在相关性;若 p 值大于给定的临界值,说明两者之间相互独立。

利用 MATLAB 计算 9445 个位点与疾病之间的卡方统计量值和相应的 p 值,以卡方检验的 p 值小于 0.01 为标准,从中筛选出与疾病存在相关性的 73 个位点,结果如表 1 所示。

表 1　卡方检验的 73 个筛选结果

Table 1　73 screening results of the Hi-square test

位置	位点	p 值
2938	rs2273298	0.0000005
292	rs2250358	0.0000713
8380	rs7533305	0.0003287
7737	rs93272	0.0002207
80	rs1263	0.0003229
⋮	⋮	⋮
8589	rs933306	0.0003303
932	rs1225350	0.0005353
1531	rs7368252	0.0006903
962	rs33926	0.0007627
3588	rs5736051	0.0020900
⋮	⋮	⋮

2.2.2　Logistic 回归

本文中位点的取值为碱基对,每一个位点编码方式均为 3 种。考虑到每个位点有 3 个取值,将位点拆分为两个 0 - 1 型变量。以第 i 个位点 rs2273298 为例,样本中位点 rs2273298 有 AA、AG 和 GG 3 种不同编码方式,将此位点拆分为两个示性变量

$$x_{2i-1} = \begin{cases} 1 & AA \\ 0 & \text{other} \end{cases}$$

$$x_{2i}=\begin{cases}1 & GG\\ 0 & \text{other}\end{cases}$$

根据此方法将通过卡方检验得到的 73 个位点拆分成146 个示性变量。

观测指标变量为$(\boldsymbol{X},y)$,其中 $\boldsymbol{X}=(x_1,x_2,\cdots,x_{2m})$是 m 个$(m=73)$位点的信息,$y\in\{0,1\}$是一个二分类的属性变量($y=1$ 表示第 i 个样本是患病者,$y=0$ 表示第 i 个样本是健康者)。用 Logistic 回归[3]建立患病识别准则模型:

$$p=\frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_{2m}x_{2m})}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_{2m}x_{2m})} \quad (5)$$

其中,$\beta_i(i=1,2,\cdots,2m)$为变量系数,Logistic 回归值 p 为个体样本是否患病的概率。

根据回归系数的 t 检验来判断位点与疾病之间的相关性是否显著。由于样本量较大,t 检验统计量近似服从标准正态分布

$$T=\frac{\hat{\beta}_i-\beta_i}{\sigma_{\hat{\beta}_i}}\sim N(0,1) \quad (6)$$

由式(6)结合标准正态分布表可计算出 t 检验统计量的 p 值

$$p=P\{|T|>t\}$$

以 146 个示性变量为自变量进行 Logistic 回归,以 t 检验的 p 值来衡量位点是否对患病概率存在影响。以 p 值小于 0.05 作为筛选标准进一步筛选出变量所对应的位点,得到显著相关水平较高的 55 个位点如表 2 所示。

表 2 Logistic 回归的 55 个筛选结果

Table 2 55 screening results of Logistic regression

位置	位点	位置	位点
2938	rs2273298	962	rs33926
292	rs2250358	3588	rs5736051
8380	rs7533305	1593	rs752233
7737	rs93272	3753	rs2999878
80	rs1263	353	rs3636092
8589	rs933306	757	rs880801
932	rs1225350	5937	rs283567
1531	rs7368252		

2.3 最优位点组合求解

根据初步筛选得到 55 个与某遗传疾病存在关联性的位点及位点组合的预报准确率,可以找出对遗传疾病发生可能性影响最大的位点组合。但是可能的位点组合数目有 $2^{55}-1$ 个,要逐一计算对比寻找最优位点组合将产生非常大的计算量。为此本文利用粒子群算法通过迭代逼近来求出最优位点组合,在采用神经网络方法针对候选位点训练拟合数据时将样本分为训练样本和预测样本,其中训练样本 900 个,预测样本 100 个。将 BP 神经网络对预测样本的预测正确率作为粒子群算法中的适应度函数值,然后进行迭代求解。

2.4 结果对比及分析

根据在多个最优解中出现的次数选出重要程度最高,即与疾病关联性最强的 11 个位点如表 3 所示。

表 3 与疾病关联性最强的 11 个位点

Table 3 The strongest 11 genetic loci associated with disease

位置	位点	位置	位点
92	rs28337	3307	rs273530
292	rs2250358	3927	rs28095
392	rs382033	6077	rs1573253
962	rs33926	7737	rs93272
1531	rs7368252	8589	rs933306
2938	rs2273298		

将上述 11 个位点作为输入序列,患病信息为输出量,从真实数据和预测数据的比较(图 1)可以看出模型具有较好的预测效果。

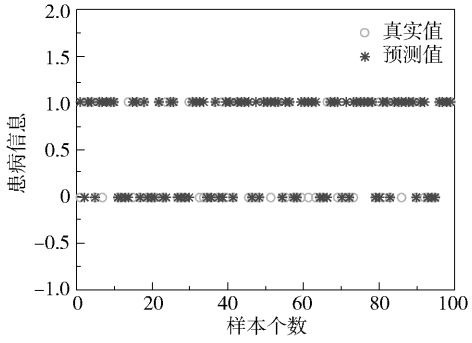

图 1 预测样本中真实值与预测值对比图

Fig. 1 A comparison of the true and predicted values for the predictiv sample

为了比较本文所提出的基于神经网络和粒子群算法的位点组合的筛选方法(方法 1)与单一神经网络权重分析法(方法 2)的准确度差异,以 55 个初选得到的位点取值为输入值直接训练数据,以每个位点在神经网络中的权重贡献率大小排序,再从中选取权重贡献率最大的 11 个位点作为与疾病关联性最强的位点。

利用混淆矩阵和受试者工作特征曲线 ROC 下方面积(AUC)检验这两个方法预测效果，具体结果如表 4 所示。

表 4　两种方法的预测效果比较

Table 4　Comparison of the predictive effects of the two methods

实际状态	分类准确率/%		AUC	
	方法 1	方法 2	方法 1	方法 2
不患病	62	56	—	—
患病	86	74	—	—
总计	74	65	0.77	0.72

由表 4 结果可知，基于粒子群算法所筛选的位点组合预测效果比单一神经网络权重法更好，尤其是针对患病者的预测精度高达 86%，说明本文方法分类正确率更高。

3　结束语

针对寻找与疾病存在关联性的位点问题，本文提出以神经网络的预报准确率为评价标准，基于粒子群算法筛选与某遗传疾病相关的重要位点组合。建立的位点筛选方法具有较强的适应性，包含了位点之间存在交互作用的情形，与只考虑选取单个重要位点的单一神经网络权重法相比，由本文方法得到的位点组合的预测效果更好，适用范围也更广泛，为疾病诊断和遗传疾病分析提供了一种新方法。

参考文献：

[1] Taylor K C, Evans D S, Edwards D R V, et al. A genome-wide association study meta-analysis of clinical fracture in 10,012 African American women[J]. Bone Reports, 2016,5:233-242.

[2] 赵冀,周超,邓小凡,等. microRNA-137 基因与原发性肝细胞癌患病风险和手术治疗预后分析[J]. 中国临床研究, 2016, 29(7):880 - 883.
Zhao J, Zhou C,Deng X F, et al. Association of the micro RNA-137 gene with morbid risk and surgical treatment prognosis for primary hepatocellular carcinoma [J]. Chinese Journal of Clinical Research, 2016,29 (7):880-883. (in Chinese)

[3] Nikolic Z, Savic P D, Vucic N, et al. Assessment of association between genetic variants in microRNA genes hsa-miR-499, hsa-miR-196a2 and hsa-miR-27a and prostate cancer risk in Serbian population[J]. Experimental & Molecular Pathology, 2015, 99(1):145-150.

[4] 杨亮, 李涌涛, 齐新, 等. CYP1B1 基因 rs1056836 位点多态性与新疆维吾尔族乳腺癌易感性的研究[J]. 临床肿瘤学杂志, 2014(8): 728-733.
Yang L, Li Y T, Qi X, et al. The relationship between the polymorphism in CYP1B1 gene rs1056836 and the susceptibility to breast cancer in Xinjiang Uygur women [J]. Chinese Clinical Oncology, 2014, 19(8): 728-733. (in Chinese)

[5] Falk C T, Gilchrist J M, Pericak-Vance M A, et al. Using neural networks as an aid in the determination of disease status: comparison of clinical diagnosis to neural-network predictions in a pedigree with autosomal dominant limb-girdle muscular dystrophy[J]. American Journal of Human Genetics, 1998, 62 (4): 941-949.

[6] 杜文聪, 陆莹, 叶新华, 等. 应用 BP 人工神经网络探讨脂联素基因多态性位点间交互作用与汉族人群 2 型糖尿病遗传易感性的关系[J]. 中国糖尿病杂志, 2012, 20(1): 20-23.
Du W C, Lu Y, Ye X H, et al. Association between adiponectin (APN) gene polymorphism locus interacts and type 2 diabetes risk in a Chinese Han population studied by BPANN[J]. Chinese Journal of Diabetes, 2012, 20 (1): 20-23. (in Chinese)

[7] Wu X S, Jin L,Xiong M M. Composite measure of linkage disequilibrium for testing interaction between unlinked loci[J]. Eur J Hum Genet, 2008, 16(5): 644-651.

[8] 徐静. 基于得分检验的整体基因间共关联作用统计方法研究[D]. 济南:山东大学,2016.
Xu J. Statistical method study for detecting co-association of whole genes based on score test [D]. Jinan: Shandong University,2016. (in Chinese)

[9] Eichler E E, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease[J]. Nature Reviews Genetics, 2010, 11 (6): 446-450.

[10] 李芳玉. 多数量性状的整体基因间交互作用统计推断方法研究[D]. 济南: 山东大学, 2014.
Li F Y. Statistical methods for detecting gene-based gene-gene interaction on multiple quantitative traits [D]. Jinan: Shandong University, 2014. (in Chinese)

[11] 彭倩倩. 群体病例对照研究设计的整体基因关联分析统计推断方法研究[D]. 济南:山东大学,2009.
Peng Q Q. Whole-gene-statistical-method research for population-based case-control study [D]. Jinan: Shandong University, 2009. (in Chinese)

[12] Schaid D J, McDonnell S K, Hebbring S J, et al. Nonpara-

metric tests of association of multiple genes with human disease[J]. Am J Hum Genet, 2005, 76(5): 780-793.

[13] Xu S H, Mu X D, Chai D, et al. Multi-objective quantum-behaved particle swarm optimization algorithm with double-potential well and share-learning[J]. Optik-International Journal for Light and Electron Optics, 2016, 127(12): 4921-4927.

[14] Karami A, Guerrero-Zapata M. A hybrid multiobjective RBF-PSO method for mitigating DoS attacks in named data networking[J]. Neuro-computing, 2015, 151: 1262-1282.

[15] 吕思晨. 基于遗传和粒子群搜索的SNP关联分析算法[D]. 西安: 西安电子科技大学, 2014.
Lv S C. SNP association study by genetic particle swarm optimization [D]. Xi'an: Xidian University, 2014. (in Chinese)

Correlation analysis of genetic site and disease information based on neural networks and particle swarm optimization

LI Jie[1] LI ZhiQiang[2*] LIU Xiao[1] YAN BaiLu[2]

(1. School of Economics and Management; 2. Faculty of Science, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: The method of screening the most powerful loci combinations has been studied under consideration of the interactions between loci when genetic diseases are associated with these genetic loci. In this paper, the prediction accuracy based on neural networks is taken as the evaluation criterion to find the optimal combination of loci by the particle swarm algorithm through iterative approximation. Compared with the weight analysis method, this method has higher accuracy, and has a good recognition effect for a disease, and can thus provide a reference for disease diagnosis.

Key words: genetic locus; interaction; particle swarm optimization (PSO); neural network

(责任编辑:汪 琴)