基于 Copula 熵和随机方法的基因数据分析

摘要

基因数据分析是生命科学领域的重要研究课题。本文以针对红眼病基因，围绕来自 120 个独立实验体的 18976 个基因点位数据，建立数学模型，分析了其他基因与目标基因 (**1389163_at**) 的非线性关系，找出了对目标基因有影响的基因并对影响程度进行了量化排序，最后给出了一种策略，能够以最少的自变量数目对目标基因建立一个合适的模型。

在评价其他基因点位是否与目标基因具有相关性时，本文以 Copula 熵 (CE) 作为评判基因之间相关性的核心指标，其中 CE 值越高则代表数据的统计相关性越强。为了得到 CE 值的阈值以判别是否相关，本文采取蒙特卡洛模拟的方法得到了无关随机变量之间的 CE 值分布，再以 99.9% 单侧置信限为阈值，筛选出了置信程度为 99.9% 条件下，可以认为与目标基因有关的基因，共计 1437 个。

进一步对变量重要性进行排序时，我们首先选用多重感知机模型 (MLP) 对相关基因和目标基因进行建模。为了得到重要性指标，参考了嵌入法的变量重要性排序方式，利用训练后的 MLP，将自变量置为随机 (视为一种随机扰动)，观察扰动后预测误差的变化情况，并以此作为重要性指标进行排序。

最后，为了以最少自变量对目标基因建立一个合适的模型，我们以 MLP 在测试集上的均方误差 (MSE) 作为指标，认为 MSE 低于 0.15 的模型即为可用的模型，并在此标准下尽可能降低自变量个数。在实际操作过程中，在 1437 个有关变量中进行自变量的选取是一个难题。本文给出了一种基于随机采样和贪婪算法的选取方式：首先按照变量重要性从高到低依次将相关变量加入模型中，直至模型在测试集上的 MSE 达到要求；其次利用贪婪思想，在保持 MSE 达到要求的前提下，尽可能删去模型中变量，达到一个局部最优解；而后保留当前局部最优的变量个数，对相关变量进行随机采样，适当保留随机值使得 MSE 逐步下降，最后，对随机采样出的自变量利用贪婪思想进行第二次的删减，便得到一组最优解，最后得到了含有 7 个变量的 MLP 模型，具有 0.146MSE 的同时 R^2 达到了 0.996。

本文中用到了较多的随机方法，如蒙特卡洛模拟，MLP 模型的随机优化，重要性排序对自变量进行随机扰动，以及在筛选变量时的随机采样方法，从而数值结果可能不具有可复现性。但值得一提的是，本文在最后对最少自变量数模型进行了稳定性分析，在八次变量筛选中，最少变量数均保持在 6 至 12 之间，证明了本文的结果相对较好，且采用的随机方法也具有一定的稳定性。

关键词：基因分析；数据分析；相关变量筛选；变量重要性排序；模型最少自变量数；Copula 熵；蒙特卡洛模拟；多重感知机；随机采样；贪婪算法

1 问题重述

对于基因数据的分析是近年来生命科学领域的一个重要课题，高效的基因分析技术可帮助医学工作者进行对各种疾病的诊断，并判断不同个体的潜在患病风险，从而及时有效地采取预防措施。

本文以红眼病基因为主要研究对象，尝试分析 120 个独立实验体获取的 18976 个基因位点数据 [1]，其中，位于 **1389163_at** 上的基因被发现与实验体患有红眼病的情况具有密切关系 [2]。

基于上述背景,本文研究的目标主要为:建立数学模型,分析其他基因与目标基因 (**1389163_at**) 表达的**非线性关系**，尝试找出有影响的基因并对其影响力进行排序，在此之后，考虑作为自变量基因之间的相关性，给出一种策略能够以最少的变量数对目标基因建立一个合适的模型。

2 问题分析

2.1 有关基因筛选问题的分析

为了筛选出与目标基因 (**1389163_at**) 相关的基因，我们需要量化获取数据中其他基因与目标的相关性，并以此筛选影响因子大的相关位点。考虑基因位点之间未知的相关性与独立性，我们选用 **Copula 熵模型**作为指标对自变量基因进行初步筛选，从而解决此问题。

而考虑到实际过程中，无法先验地确定 Copula 熵阈值以划分有关基因和无关基因。因此，我们尝试利用**蒙特卡洛方法**对随机无关变量间的 Copula 熵的分布进行模拟，基于估计出的分布函数，将落在单侧置信区间内的基因视作无关基因，置信区间外的基因视作有关基因。也就是说，**利用 Copula 熵进行变量筛选的阈值是由无关变量 Copula 熵近似分布的单侧置信限决定的**。

2.2 基因重要性排序问题的分析

在初步筛选出有关基因后，为了更加细致地将各个有影响基因对于目标基因的重要程度 (或影响力) 进行排序，我们参考了机器学习中嵌入法模型中的变量重要程度排序算法，即将某一个变量置为随机，代入模型中检验模型准确率因此下降的幅度，以下降幅度作为衡量模型重要性的指标。

具体到本文面对的问题中，对于每一个有影响基因，我们将该基因的数值**置为随机** (为了简便，在实际操作过程中取为正态随机量)，再代入模型中比较预测目标基因值与实际目标基

因值的差距，观察并量化**模型的准确率下降程度**，以衡量每个有影响基因对目标基因的影响程度。

基于上述讨论，首先对有影响基因和目标基因之间的关系进行建模求解是必要的，这里我们**选取了多重感知机模型以衡量基因之间复杂的非线性关系**，在完成训练后对测试集中有影响基因值进行随机化并计算模型准确率的下降程度，从而以此为标准对各有影响基因的重要性进行量化并排序。

2.3 最少自变量个数问题的分析

由于只有一个表达值表征每只动物的每个转录位点, 而这个值本质上体现了对该动物中该基因表达的所有遗传影响的总和。因此，变量之间的相互影响对于最少自变量个数的分析是至关重要的。为了更加具体的认识问题，我们将**“以最少的自变量建立适当的模型”**等价为**“在保证模型准确率的前提下尽可能减少自变量个数”**，以便于后续的分析求解。

在这种情况下我们首先选择**对目标基因影响力高的基因**建立准确率较高的模型，以迅速地得出初始解。再以归一化后的 CE 值为权重，利用**随机采样方法**选择变量，以较低的代价求出更优的变量组合作为新解，保证了解的优度的同时不会带来过大的运算量和代码运行时间。而为了进一步加快求出最优解的速度，我们在得出初始解和随机采样后得到最优解后，在保证模型的准确率依旧较高的前提下，**利用贪婪算法尽可能筛去模型中的变量**。

同时, 由于选用了随机模型, 我们将重复随机多次检验两点：一是每次随机求出的**最小变量个数的波动情况**，二是最小变量模型里选取的变量是否呈现某些**集中性趋势**, 进而验证我们提出的筛选机制是否具有稳定的表现。

3 模型假设

为了便于处理分析，我们在对问题进行建模求解时给出了以下假设：

1. 每一只作为实验体的小鼠的基因数据互为相互独立的随机变量，小鼠之间互不影响且每个小鼠实验时的所有外部条件均相同；

2. 目标基因的表达值仅与所给数据中其他基因的表达值相关；

3. 若其他基因被验证与目标基因分布无关，则认为其他基因的表达值与目标基因的值没有任何关系，在分析中可以直接删去。

4 符号说明

表 1: 符号说明

符号	含义	说明
X	随机变量	
$\{x_i\}$	X 的一组样本	
$F_i(x_i)$	经验 Copula 密度函数	
CE_i	基因 i 的 Copula 熵值	由经验 Copula 密度函数估计得到的 Copula 熵值
α	置信程度	
R^2	决定系数	决定系数越大，模型拟合程度越好
ΔM	均方误差 MSE 上升均值	同时作为变量重要性指标
P_i	随机采样基因 i 的概率	
CE_i	归一化后的 Copula 熵	

5 模型建立与求解

5.1 基于 Copula 熵的变量筛选模型的建立与求解

5.1.1 Copula 熵 (CE) 与 Copula 熵估计

CE 可以看作由 Copula 理论得到的一种特殊香农熵 [3]，具有连续性，对称性，可加性等基本性质，而 CE 与变量相关程度呈正相关关系，即 CE 越大，相关性越强。

应当指出，当 CE 应用与变量选择问题时，相比于传统的变量选择方法，其具有**模型无关**、**数学理论坚实**、**物理上可解释**以及**具有非参数估计算法**等显著优点。在给定随机变量 **X** 的一组独立同分布样本 $\{x_1, x_2, \cdots, x_N\}$ 时，由马健 [3] 给出的 CE 估计算法如下：

1. 由公式 (1) 估计经验 Copula 密度函数：

$$F_i(x_i) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}(x_n^i < x_i) \tag{1}$$

其中 **1** 为示性函数，x_n^i 为给定样本的次序统计量

2. 由经验 Copula 密度函数估计 CE，由于这本质上是一个熵估计问题，这里我们选用非参数方法 k-邻近法来估计 CE。

综上我们得到了 Copula 熵 (CE) 的一个非参数估计模型，我们首先利用这个模型对数据中的变量进行初筛，以大幅降低维数，便于后续进一步的模型建立与求解。

5.1.2 基于 Copula 熵估计的变量筛选

本文将目标基因数据与其他自变量基因数据两两配对，按照 CE 估计算法进行估计，得到了每个自变量基因与目标基因的 Copula 熵估计值，其中 CE 分布图与排序后的散点图分别如图 1,2 所示：

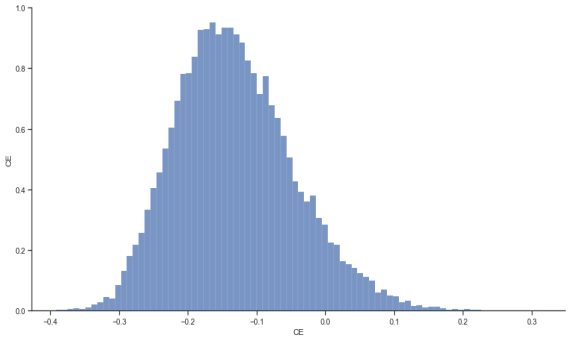

图 1: CE 分布图

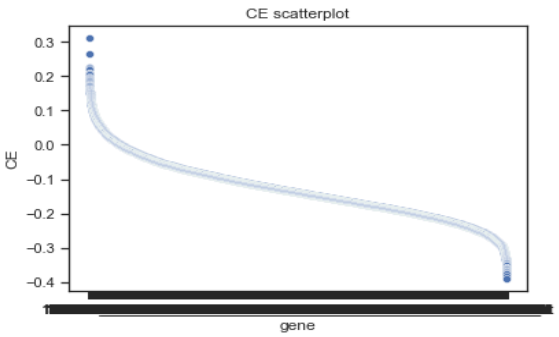

图 2: 排序后的 CE 散点图

从两个图中均可由看出，CE 值较大与较小的值均不多，主要集中在中间的部分，近似于正态分布。而为了利用 CE 初步筛选出可能与目标基因有关的变量，我们首先需要得知，在无关的情形下，CE 的分布是怎么样的。

基于这样的想法，我们利用蒙特卡洛方法进行随机数值模拟，又考虑到 CE 的估计具有单调变换不变性 [3]，从而只需随机取 150 维的标准正态向量，均值和方差简单取 0,1 即可。计算独立随机样本和目标基因的 CE 值，画出 30000 次随机模拟得到的 CE 值的分布图及正态性检验的结果如图 3 所示：

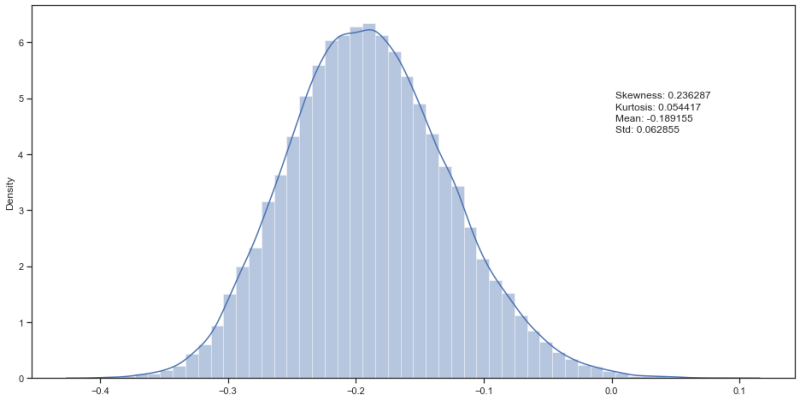

图 3: 随机模拟的 CE 值分布图及正态性检验，$\mu = -0.189155, \sigma = 0.062855$

利用蒙特卡洛方法采样后的样本均值与方差，我们对无关变量之间的 CE 值分布给出了估计：

$$CE \sim N(-0.189155, 0.062855^2) \tag{2}$$

基于此估计，我们利用考虑置信程度 α 单侧置信区间对变量进行了筛选，即认为满足 CE 值位于置信区间外的自变量基因属于对目标基因有影响的基因，否则则认为是无关基因。

我们分别取置信程度 $\alpha = 99\%, 99.5\%, 99.9\%$，得到的有关变量个数如表 2 所示：

表 2: 不同置信程度与对应选出有影响基因数

置信程度 α	99%	99.5%	99.9%
单侧置信区间限	$\mu + 2.33\sigma$	$\mu + 2.575\sigma$	$\mu + 3\sigma$
有影响基因数	2873	2281	1437

参考上述表格，我们选取 $\alpha = 99.9\%$ 作为置信程度，对应地选取了 1437 个自变量基因作为对目标基因有影响的基因，其余则作为无关基因洗去相关数据。

总之，根据上述方法，最后筛选出了对目标基因有影响的基因，其中部分基因及其对应 CE 值如表 3 所示：

表 3: 筛选出的部分相关基因及其对应 CE 值

基因名	1370205_at	1372736_at	$\cdots$	1379971_at
CE 值	0.312824	0.265163	$\cdots$	-0.000367

5.2 变量重要性模型的建立与求解

5.2.1 多重感知机模型

考虑到基因表达之间具有复杂的非线性关系，我们采用最简易的神经网络，即多重感知机 (MLP) 模型 [4] 以探索求解目标基因的表达与选出的有关自变量基因之间的关系。

在本文中，我们首先利用 150 组样本数据，以选出与目标基因有关的基因数据作为自变量，目标基因数据作为标签，训练出多重感知机模型。本节后续的工作将基于这样一个训练完成的多重感知机展开。

搭建的多重感知机网络各种参数及训练后评分如表 4 所示：

表 4: 多重感知机参数

Layer	Param
Input Layer	1437
Hidden Layer1	300
Hidden Layer2	40
Hidden Layer3	20
Output Layer	1

在此条件下，实际训练的结果在训练集上达到 $\mathbf{R^2 = 0.9997}$，其中 $R^2 \in [0, 1]$，越接近于 1，模型效果越好，同时在测试集上达到了 0.1651 的均方误差。可以看出，本文中给出的多重感知机模型在利用筛选出的相关基因对目标基因表达程度的估计问题中表现良好，可以作为后续检验变量重要性的基准。

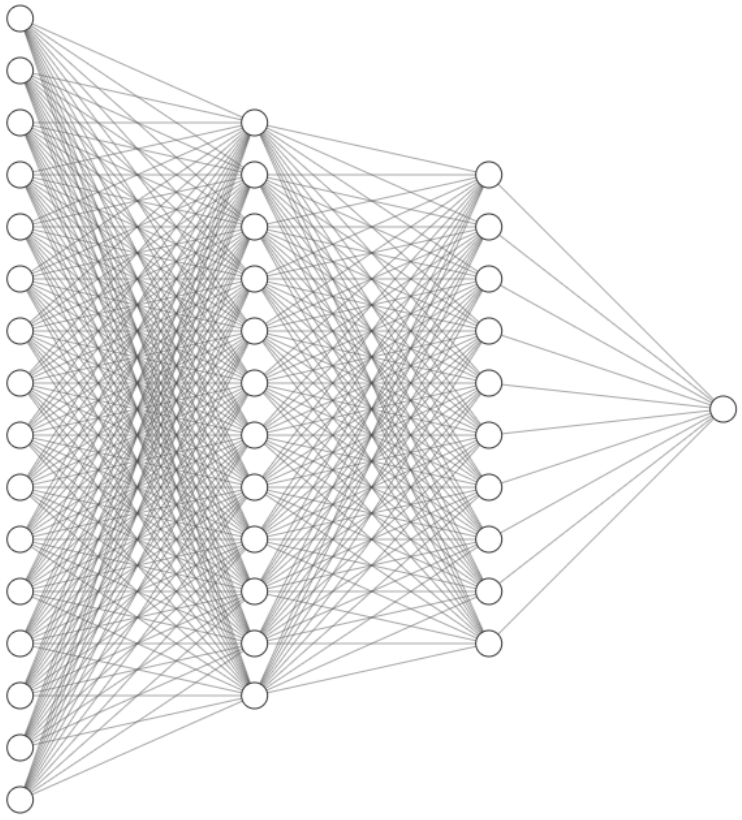

图 4: 小规模 MLP 结构示意图

5.2.2 变量重要性指标模型以及基因重要性排序

为了衡量各个基因对于目标基因的重要性，我们参考嵌入法 [4] 计算变量重要性的思想，即将某一个变量置为随机值，代入模型，观察模型准确度是否受到影响，以准确率波动的大小来衡量某个变量对于该模型的重要程度。

将每一个变量多次置为随机值得到 MLP 模型准确率下降值的数据，并绘图如图 5 所示：

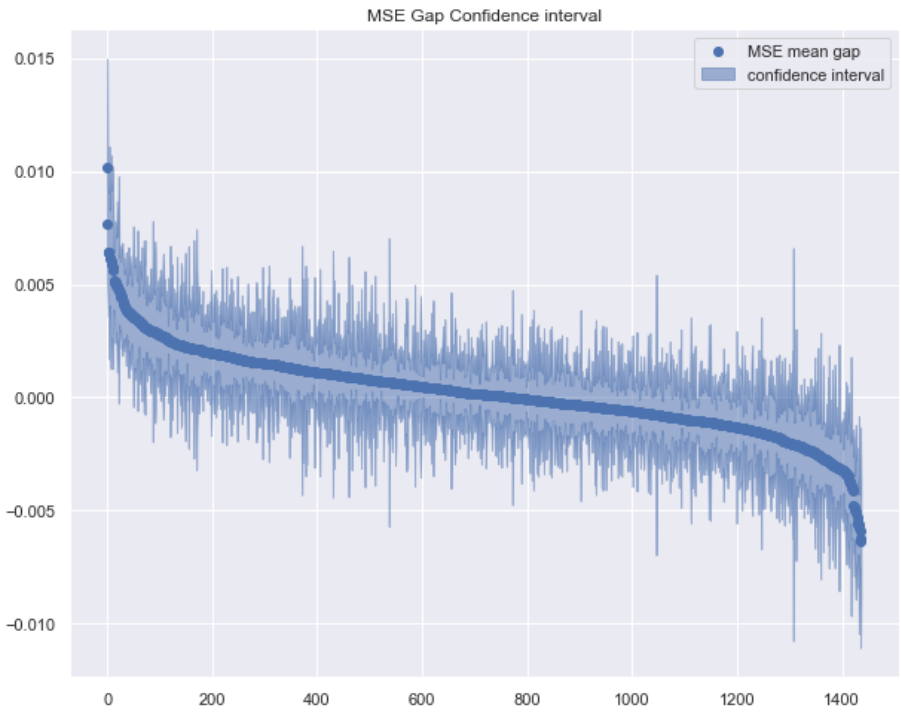

图 5: 排序后的 ΔM 及其置信区间图

可以看到，将不同基因置为随机后，预测模型受到的影响幅度有着明显差异，于是将均方误差 MSE 上升的均值 ΔM 作为变量重要性指标，即可得到基因重要性排序，部分基因的次序及其重要性指标如表 5 所示：

表 5: 重要性前四的基因及其重要性指标

次序	1	2	3	4
基因名	1369662_at	1382362_at	1390669_at	1371678_at
ΔM	0.0102±0.0048	0.0077±0.0033	0.0064±0.0029	0.0064±0.0026

可以看出，每一个变量置为随机后对模型准确率的影响都不大，这也说明了选取了所有有关变量后得到的 MLP 模型中，自变量信息重合的现象严重。因此，尽可能的减少参数以获得自变量尽可能相互独立的预测模型是很有必要的。

5.3 以低自变量数目为目标的变量筛选机制模型的建立与求解

5.3.1 变量筛选机制简介

基于多重感知机 (MLP) 模型，我们提出了一种基于随机采样方法和贪婪算法的变量迭代筛选机制以获取含有尽可能少自变量的预测模型。具体步骤为：

- 初始化可行解：在 MLP 模型中按照变量重要性从高到低的顺序逐个添加自变量，直至模型准确率稳定地高于阈值；
- 贪婪算法筛除初始可行解中的变量：在得到的可行解中，逐个尝试删去变量，使得模型准确率依旧稳定的高于阈值，直至每个变量都不能被删除为止，得到一个可行解；
- 随机采样迭代：以归一化后的 CE 值为权重，随机地从有关变量里选取数目和当前可行解一样的变量，若能够使得模型准确率稳定提高，则更新可行解，重复多次直至迭代多次后可行解不变或达到最大迭代次数，结束迭代；
- 基于贪婪方法对迭代后的解进行筛除：最后，对随机采样后的可行解利用贪婪算法进行变量筛除，最后得到最优解，最优解的数目即为求出的建立模型所需最少变量数。

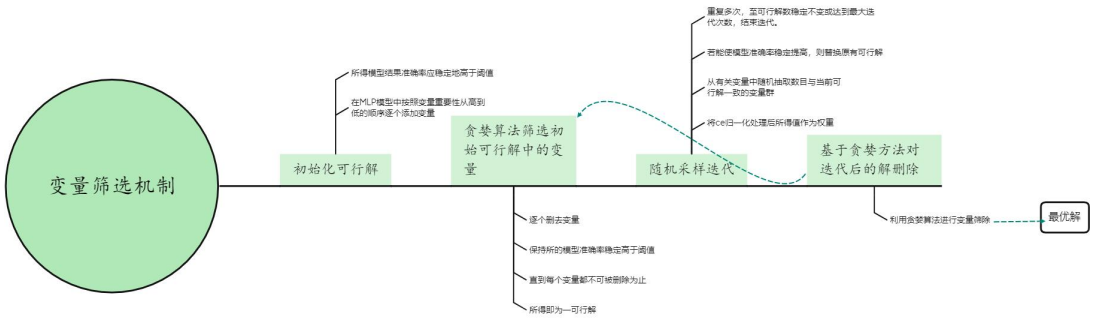

图 6: 变量筛选机制步骤示意图

5.3.2 初始化可行解的生成与第一次变量剔除

首先，我们按照变量重要性排序问题中求得的次序，按批次将基因变量向 MLP 模型中添加并训练，直至得到的模型均方误差 MSE 达到 0.15 以下。在以 10 个变量为单位向模型中添加变量过程中，均方误差随变量个数变化的情况如图 7 所示：

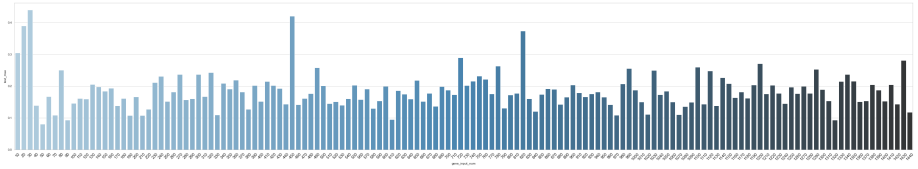

图 7: R^2 随加入模型变量个数的变化图

可以观察到在 40 个变量时模型的 MSE 首次跌入 0.15 内，于是只需考虑 30-50 变量时模型的 MSE 何时最先跌入 0.15 即可，绘制变化图如图 8 所示，可以看到，恰好是 40 个变量时，模型的 MSE 进入 0.15 以下，达到了 0.139

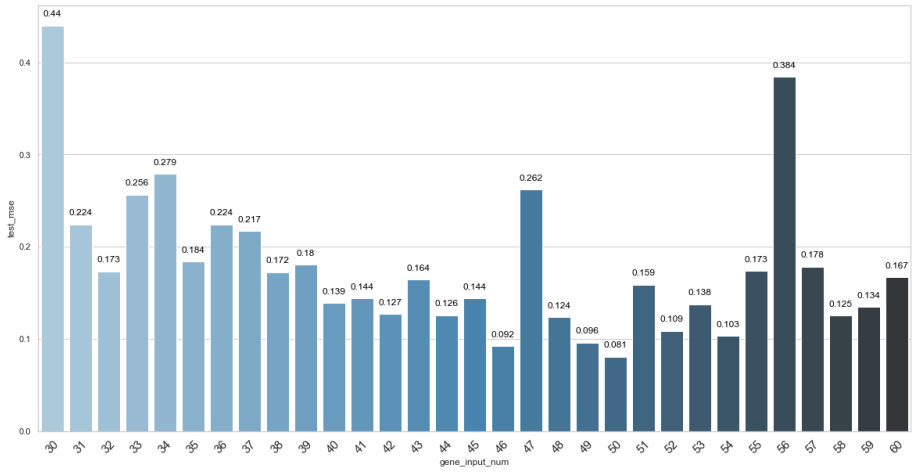

图 8: 30-50 个变量模型 MSE 的变化图

随后，我们得到了以重要性排前 40 的基因作为自变量的 MLP 模型，达到了 0.139 的模型 MSE

而后进行变量剔除，逐步筛去了 1367491_at,1368318_at 等 **27 个**变量，使得模型 MSE 改变至 0.104，同时模型变量个数下降为 13，筛去变量过程中得到的 MSE 如表 6 所示：

表 6: 第一次变量剔除过程

已删除变量个数	删除变量名	新的 R^2 值
1	1367491_at	0.010
2	1368318_at	0.127
3	1367845_at	0.112
...		
27	1368212_at	0.104

5.3.3 随机迭代采样与第二次变量剔除

按照生成初始可行解中变量个数 40，我们以概率 P_i 对相关基因进行随机采样，其中概率 P_i 满足 (3)：

$$P_i = \tilde{CE}_i / \sum_{n=1}^{N} \tilde{CE}_n \tag{3}$$

其中 $\tilde{CE}_i$ 为相关基因 i 的 Copula 熵归一化后的值

按照上述概率重复进行随机采样，直至采样出的变量组合在模型中能表现出更低的 MSE，则采取新变量组合，迭代步数与 MSE 的关系如图 9 所示：

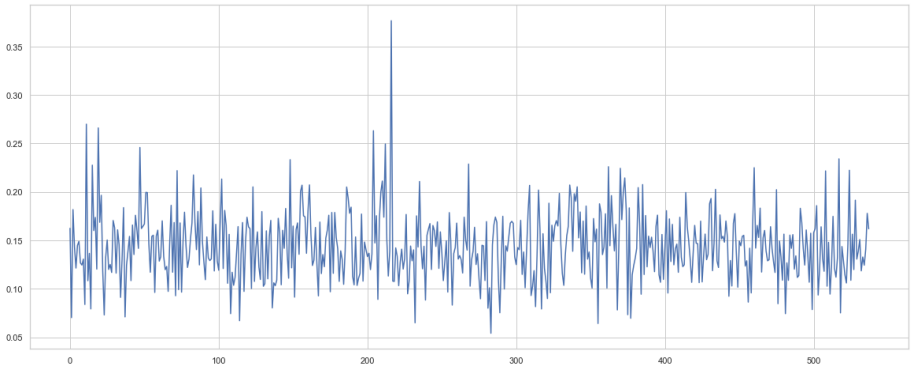

图 9: 随机采样过程中 MSE 变化情况

随后同样进行变量剔除，逐步筛去变量，使得模型 MSE 变为 0.146，同时模型变量个数下降为 7 个，筛去变量过程中得到的 MSE 如表 7 所示：

表 7: 第二次变量剔除过程

已删除变量个数	删除变量名	新的 MSE 值
1	1383444_at	0.104
2	1368470_at	0.110
3	1389265_at	0.096
...		
7	1382579_at	0.146

最后得到含有 7 个变量的 MLP 模型，训练优度 R^2 值为 0.996，在测试集具有 0.146 的 MSE 值。认为得到的值即为对于目标基因建立一个模型所需要的最少自变量数 7。

5.3.4 变量筛选模型的稳定性分析

考虑到算法具有随机性，我们尝试了重复八次变量筛选，得到的每一次变量个数如图 10 所示：

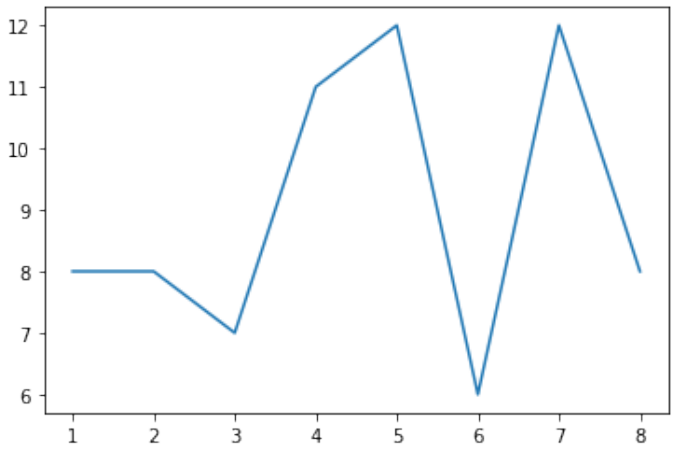

图 10: 八次变量筛选得到的最小自变量个数

可以观察到算法的结果还是相对稳定的，取 7 作为最少自变量个数建立目标基因相关模型并没有什么问题。

6 模型评价与推广

6.1 模型的优缺点

6.1.1 模型的优点

- Copula 熵法相比于其他求解相关性手段具有明显优越性。该法物理可解释性好，变量无关，有充分坚实的数学理论基础，对于基因分析这种没有明确模型的问题有着极高的效率与较好的效果，十分适合大规模变量筛选问题；
- 采用了参考嵌入法的变量排序机制，运行速度快。同时这样的变量排序机制也具有极强的直观性和可解释性，是一种通过实验的得到的排序方式，与更加理论化的 Copula 熵法对应互补，且更具有说服力，适合要求更精细的重要性排序工作；
- 在利用 Copula 熵法进行筛选时，模拟了无关变量 CE 值服从的分布及其概率密度函数(正态分布)，再以置信区间筛选变量，计算代价小且具有说服力；
- 在刻画自变量基因与目标基因的关系时采取多重感知机模型，较好的解决了未知非线性关系难以处理的问题；
- 在最少自变量个数问题中，本文采用贪婪算法迅速的得出局部最优解，进而利用随机采样方法增加了求出的局部最优解接近全局最优解的可能性，保证了解的优度的同时不会带来过大的运算量和代码运行时间。

6.1.2 模型的缺点

- 在最少自变量个数问题中所给出的变量筛选机制，本质上是一种随机迭代方法，但未能从理论上证明其收敛性和收敛速度；
- 以无关变量 CE 值分布的 99.9% 置信区间筛选 CE 值过于严格，可能筛去过多的变量，但如若减小置信程度，则剩余变量过多，后续算法运行时间可能过长；
- 模型重要性排序模型中，对于自变量的随机扰动次数不足，导致重要性顺序可能不稳定；
- 文中较多使用随机性算法，得到的结论可能不具有复现性，

6.2 模型的应用背景及推广

- **基于 Copula 熵的变量筛选模型**

此模型在理论上具有诸多优点且便于实际操作与实现，同时文中给出的估计方法并不会带来很大的运算量。更加一般地，本文提出的基于 Copula 熵的变量筛选机制，对于大规模数据集的变量筛选和模型降维问题均能够有广泛的应用；

- **变量重要性模型**

  此模型具有极强的实际意义，其实在高维神经网络中的表现并不是特别好。但对于可解释性更强的模型，如决策树、支持向量机等初等机器学习模型，这种方法的表现将更加值得期待；

- **以低自变量数目为目标的变量筛选模型**

  此模型借鉴了现代优化算法的思想，相比于传统优化算法更擅长于解决具有多个极值点的非凸问题，随机采样的格式也使得模型能够进一步的推广改进。

参考文献

[1] Scheetz, T. E., Kim, K. Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L.Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proceedings of the National Academy of Sciences 103, 14429–14434.

[2] Chiang, A. P., Beck, J. S., Yen, H. J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J., et al. (2006). Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardet–biedl syndrome gene (bbs11). Proceedings of the National Academy of Sciences 103, 6287–6292.

[3] 马健.(2022).Copula 熵：理论和应用.[ChinaXiv:202105.00070].

[4] 周志华. 机器学习 [M]. 北京. 清华大学出版社.2016.

附录

A　附件及支撑材料目录

- ”main.ipynb”：数据分析主程序文件
- ”data” 文件夹：存放有源数据及数据分析过程中产生的部分中间结果
- ”fig” 文件夹：存放有论文中使用到的可视化结果