

# “深圳杯”数模竞赛

代谢综合长征风险、趋势预测和干预模型的构建



庾鹏

马力群

丁建勋

# 代谢综合征风险、趋势预测和干预模型的构建

## 摘要

人与人之间基因序列的差异，即基因变异，影响着每个人罹患代谢综合疾病的风险高低，也影响着不同非遗传因素在每个人身上的具体作用效果。因此，本文基于人体基因信息以及与基因密切相关的蛋白质信息来对代谢综合征进行分析。

首先本文利用所给样本的 RNA 表达数据，筛选对代谢综合症有重要贡献的基因序列。鉴于所给基因组基因种类较多，文章首先采用主成分分析法对所给数据进行降维处理，其次对降维的数据进行线性回归，寻得对代谢综合症有突出贡献的基因，即代谢综合症的关键通路。

最后，本文通过利用反向传播算法训练了一个三层的人工神经网络，将处理后的三组数据作为网络的输入，患病程度作为网络的输出，通过所给数据训练后的网络模型便可以根据所给数据判断出病人的患病程度。

关键字：代谢综合征 基因序列分析 主成分分析 线性回归 BP 神经网络 反向传播算法

## 一、问题重述

随着生物技术的发展，人们对自身的探索与发展日新月异，我们日渐发现生命技术已经可是的我们可以解析人类复杂的遗产密码，换言之，一个小小的人体基因片段就有可能影响我们身体中健康的动态稳定环境，对人身体造成损害。另一方面，各类社会环境因素，生活方式等也会对其特征产生短期到长期的影响，最终形成多种外部表现特征。本题目就是基于人类基因序列引发的代谢综合症患病率的分析，给定特定随机人群的代谢综合症相关的临床检测数据，基因组数据，表观基因组数据，转录组数据和蛋白质组数据以及代谢组数据为参考，分析代谢综合症的关键患病因素，将其关键通路列出并构建人类生命量化的动态模型，最后需要我们借助临床监测数据以及相关分析所得参数确定除实验群体外人群的代谢综合症的发病情况，并给以可行性结果分析。

## 二、模型假设

1. 忽略基因在转录、翻译过程中的缺损与突变。
2. 忽略每组数据所抽取出来的特征序列之间的内在相关的生物联系。

## 三、符号说明

$m$	所给人群样本中患病者的人数
$n$	所给人群样本中非患病者的人数
$A_i(i=1,2,3,\dots)$	患病人群的第 $i$ 个特征序列
$B_i(i=1,2,3,\dots)$	非患病人群的第 $i$ 个特征序列
$P(A_i)$	$A_i$ 出现的概率
$P(B_i)$	$B_i$ 出现的概率
$P$	患病概率
LONG	字符串数据
$t_j$	类 $C_i(i=1, 2)$ 的 TEAM 中的概率
$ C $	训练集中类的数目
$ D_{C_i} $	为训练集中属于类 $C_i$ 的 TEAM 数

$q_i, q_j$	为 TEAM 的属性值所组成的特征向量
$outputs$	网络输出单元的集合
$E$	误差
$t_{kd}$	训练样例 $d$ 与第 $k$ 个输出单元相关的预期输出值
$o_{kd}$	训练样例 $d$ 与第 $k$ 个输出单元相关的实际输出值
$D$	训练样例 $d$ 的集合
$d$	训练样例
$x_{ji}$	单元 $j$ 的第 $i$ 个输入
$w_{ji}$	与单元 $j$ 的第 $i$ 个输入相关联的权重
$\sigma(x)$	$sigmoid$ 函数
$\eta$	学习率
$net_j$	$net_j = \sum_i w_{ji} x_{ji}$ ，即单元 $j$ 的输入的加权和
$Downstream(j)$	单元的直接输入中包含单元 $j$ 的输出的单元的集合

#### 四、问题分析

为找到对解决该题最为重要的代谢综合征的关键患病因素以及对样本是否患病的判断依据，我们从以下几个方面对问题进行分析：

首先，对题目本身进行分析。由题目可知基因是罹患代谢综合征的主要内因，所以模型的构建应从基因入手。

其次，对与题目相关的生物知识进行分析。根据中心法则（图 1），DNA（对应基因组和表观基因组数据）、RNA（对应转录组数据）与蛋白质（对应蛋白质组数据）之间存在存在着直接对应的关系。因此可以将这一联系视为所给数据间的主要联系并且以上提到几组数据可以采用相类似的方法进行分析处理。



图 1 中心法则

文章首先将患病程度等级和每位的样本数据，性别，变异程度等建立对应关系，采用降维处理，主成分分析，神经网络分析等算法建立数据间的对应关系。

文章弱化未标明患病程度的样本，因此对数据的处理并不全面。

最后，从解决方法角度进行分析。虽然，由中心法则等相关生物知识可以知道所给数据间以及数据对所需求解的问题的大致联系，但是人体是一个复杂的非线性系统，各种因素间的联系错综复杂，难以构建一个模型将所有因素之间的联系都考虑进去。因此我们先尝试弱化不同因素间的联系，直接建立起统计数据与结果的对应关系，而神经网络算法恰能满足这一要求。文章用反向传播算法，运用输出单元的权重训练法则，求解输入新的样本数据后其患病程度。

综上所述，我们将围绕对字符串数据类型的处理、主成分分析，线性回归分析与神经网络算法来构建模型解决问题。

## 五、模型建立

### 一、利用方差与主成分分析对数据进行处理找出关键通路

由于本文所给 RNA 数据较多，首先对 20000 多组 RNA 数据进行筛选，选取具有样本中具有显著差异的 RNA 序列作为研究对象。论文首先对变异组数据进行筛选，对所给的 200 余样本的各个 RNA 序列表达情况计算方差，计算不同 RNA 序列在不同样本中表达数目的离散程度，对于离散程度较小的 RNA 序列，近似认为对代谢综合征的贡献较小，在之后的分析处理中不作计算。

$$\text{平均数: } M = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \quad (n \text{ 表示这组数据个数, } x_1, x_2, x_3, \dots, x_n \text{ 表示这组数据具体数值})$$

$$\text{方差公式: } s^2 = \frac{(M - x_1)^2 + (M - x_2)^2 + (M - x_3)^2 + \dots + (M - x_n)^2}{n}$$

下表为方差计算后，所寻得的离散程度最大的 40 组 RNA 在所给样本中的表达情况。

#	DH	DI	DJ	DK	DL	DM	DN	DO	DP	DQ	DR	DS
1	V112	V113	V114	V115	V116	V117	V118	V119	V120	V121	方差	@##probe
2	6508.63183594	83726.53125000	#####	#####	#####	91044.22662650	771774.31250000	5089.87011719	338.74829102	23.74959946	164795.09734048	SFTPB6439
3	#####	7172.83398438	9182.80468750	12723.98437500	12979.120111720	59315.96875000	10052.78515620	#####	#####	29621.80664060	134611.77026244	COL1A111277
4	3651.82519531	8.12269874	2.67849994	30.45689964	47.72560120	153.31700134	11.22189999	982.25561523	#####	#####	119126.25884621	KRT6A3853
5	2128.96630869	141898.46875000	#####	#####	#####	#####	15317.53906250	7967.75146484	369.64050293	38.50310136	103630.31990101	SFTPA2729238
6	37458.08984380	24894.21093750	#####	#####	4011.86284180	10465.03710940	62505.50000000	#####	#####	85004.69375000	95972.81194189	ADAM68755
7	#####	10277.52734380	#####	12179.58335940	11258.01680160	89249.14082500	9071.07228560	#####	#####	22424.97265620	94707.25294411	COL1A21278
8	#####	9767.90525000	#####	13274.74809380	8912.00585938	74799.97568250	9239.81738231	#####	#####	26829.47070310	91592.67924946	COL3A11261
9	#####	18993.23046880	#####	45925.76171880	10280.38769530	#####	20313.79882810	#####	#####	11712.84657970	87925.62253731	FN12335
10	53.22420120	19.40430069	2693.98681641	13.95940018	12.67710018	157.73969351	6.23439980	5.61289978	#####	#####	87131.19689809	KRT53852
11	14.73900032	10.37909085	149.99330139	0.00000000	3.72860003	40.29479980	4.15600000	64.29309845	#####	13067.29003910	83240.95725156	KRT143861
12	1.63769996	0.00000000	8.57100010	10.15229988	2.23709989	0.00000000	0.00000000	51029998	79890001	0.00000000	68002.17780314	CPB11360
13	103.17298652	512.63537598	1.07140005	0.00000000	115.58540344	89.92630005	21.61260033	69.39569855	910.25299072	43823.67578120	64047.68805727	KRT163868
14	270.48791604	1071.75085449	#####	353.67388916	102.16259766	918.42749023	43.64089966	749.06561279	1893.52331543	848.00646973	69465.60908608	IGF23481
15	25158.14648440	7528.66406250	#####	62447.48046880	2708.23268602	28912.00000000	16148.97363280	#####	#####	54305.94531250	59300.01873563	LOC8861096810
16	157.48890588	52.96229563	4513.19140625	446.70050049	210.29080200	275.87589970	57.39580172	475.05450440	3479.89355469	363.08020020	57892.51403738	H19283120
17	86357.42187500	21241.42578120	#####	#####	36269.20312500	#####	19168.25195310	#####	#####	62259.44531250	57178.51279907	CDY4972
18	771.88671875	43751.35646880	#####	87248.73437500	63751.67968750	84667.81250000	8261.42968750	3628.47783320	145.67239380	61.17309952	54634.67688981	SFTPA1163509
19	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	54340.83312747	SCGB2A24250
20	153.39469910	3789.26000977	2534.88671875	66.62439728	2143.17675781	272.23590088	413.54949951	85.72409821	#####	93928.39082500	49286.25148476	KRT173872
21	#####	16430.05468750	#####	27182.74023440	47794.92968750	46098.28125000	31332.08937500	#####	#####	48266.37369773	GAPDH2697	
22	37646.94531250	15917.41804530	#####	#####	77723.34375000	59402.45703120	31868.24609380	#####	#####	78806.76562500	46903.40526909	FTL2512
23	64478.47265620	12930.05468750	8483.72851562	14471.44628910	12141.68554690	33494.34765620	12267.24804690	#####	#####	22356.88281250	45382.63705953	SPARC6678
24	1331.42272949	5140.79443359	#####	5575.50781250	2103.65405273	792.13757324	246.88279724	5.10260010	34.62049866	908.60021973	43942.72278326	LTF4067
25	#####	73652.96875000	#####	91605.96093750	#####	#####	78024.52343750	#####	#####	#####	42417.12708654	ACTB80
26	28056.13671880	19121.84179630	#####	57713.19321890	25349.73820120	36289.43353090	90632.58937500	#####	7536.75097656	18164.08789060	41676.19672953	CTSD14509
27	24007.36914060	121.84120178	751.03790283	1904.18774414	527.96417236	1458.46947236	336.24270630	941.94485626	5512.51660156	2906.1022539	40221.02973903	S100A96280
28	262.84639795	21171.02929690	#####	73564.71875000	56923.19140620	37401.95484380	8976.89355459	5.61289978	5.85890007	2.15910006	39978.47116521	SFTPC8440
29	68861.17187500	63484.20703120	#####	56714.46875000	69489.18750000	77882.55468750	83192.43750000	#####	#####	52708.60937500	39413.53443371	EEF1A111915
30	84734.77343750	48504.96484380	#####	#####	68590.60156250	63237.34765620	40429.75781250	#####	#####	69480.39062500	38480.64968704	B2M567
31	1325.14501953	9.47649956	10.17809963	3.80710006	429.53021240	20.63879967	83130002	65.31359863	#####	30821.87890620	37524.63718637	KRT6B3854
32	1289.29101562	8428.24902344	8389.44726562	4521.26904297	4647.56884766	8283.39550781	3847.70581055	2519.67700195	1063.05187988	733.96911621	37459.75299234	C4A4720

针对前期粗糙的筛选结果，论文确定了 40 组在不同患病程度样本中表达数量差异性较大的 40 组 RNA，分别为：

SFTPB|6439; COL1A1|1277; KRT6A|3853; SFTPA2|729238; SFTPA2|729238; COL1A2|1278; COL3A1|1281; FN1|2335; KRT5|3852; KRT14|3861; CPB1|1360; KRT16|3868 ; IGF2|3481 ; LOC96610|96610 ; H19|283120 ; CD74|972 ; SFTPA1|653509; SCGB2A2|4250; KRT17|3872; GAPDH|2597; FTL|2512; SPARC|6678; LTF|4057; ACTB|60; CTSD|1509; S100A9|6280; SFTPC|6440; EEF1A1|1915; B2M|567; KRT6B|3854; C4A|720。

由于基因之间存在相互关联和组合关系，为排除变量之间可能存在的相互关联关系，论文在确定关键基因之前，首先对上述 40 组数据进行主成分分析，以获得各个基因对数据的贡献率，并通过计算影响因子，对上述 40 组数据进行降维处理，使得文章的关键变量进一步减少。

### (1) 主成分分析的原理

1、原始指标数据（关键 40 组 RNA 信息的表达情况）的标准化采集  $p$  维随机向量  $x = (x_1, x_2, \dots, x_p)^T$   $n$  个样品  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ ,  $i=1, 2, \dots, n$ 。  $n > p$ ，构造样本阵，对样本阵元进行如下标准化变换：

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i = 1, 2, \dots, n; j = 1, 2, \dots, p$$

$$\text{其中 } \bar{x}_j = \frac{\sum_{i=1}^n x_{ij}}{n}, s_j^2 = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}, \text{得标准化阵 } Z。$$

2、对标准化阵  $Z$  求相关系数矩阵

$$R = [r_{ij}]_{p \times p} = \frac{Z^T Z}{n-1}$$

$$\text{其中 } r_{ij} = \frac{\sum z_{kj} \cdot z_{ki}}{n-1}, i, j = 1, 2, \dots, p。$$

3、解样本相关矩阵  $R$  的特征方程  $|R - \lambda I_p| = 0$  得  $p$  个特征根，确定主成分

$$\frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^p \lambda_j} \geq 0.85$$

按

确定  $m$  值, 使信息的利用率达 85% 以上, 对每个  $\lambda_j, j=1, 2, \dots, m$ , 解方程组  $Rb = \lambda_j b$  得单位特征向量  $b_j^o$ 。

4、将标准化后的指标变量转换为主成分

$$U_{ij} = z_i^T b_j^o, j = 1, 2, \dots, m$$

$U_1$  称为第一主成分,  $U_2$  称为第二主成分, ...,  $U_p$  称为第  $p$  主成分。

5、对  $m$  个主成分进行综合评价

对  $m$  个主成分进行加权求和, 即得最终评价值, 权数为每个主成分的方差贡献率。

## (2) spss 对数据的主成分分析处理

KMO 與 Bartlett 檢定

Kaiser-Meyer-Olkin 測量取樣適當性。	.759
Bartlett 的球形檢定 大約 卡方	4256.715
df	561
顯著性	.000

首先对实验数据进行 **KMO 与 Bartlett 球形检定**。KMO (Kaiser-Meyer-Olkin) 检验统计量是用于比较变量间简单相关系数和偏相关系数的指标。主要应用于多元统计的因子分析。论文中的数据检测值为 0.759, 变量间的相关性强, 原有变量适合作因子分析。如果变量间彼此独立, 则无法从中提取公因子, 也就无法应用因子分析法。Bartlett 球形检验判断如果相关阵是单位阵, 则各变量独立因子分析法无效。由 SPSS 检验结果显示 Sig. < 0.05 (即  $p$  值 < 0.05) 时, 说明各变量间具有相关性, 因子分析有效。对论文中数据主成分分析, 获得 sig=0.000,

故本问题中应用主成分分析，能获得较为理想的实验效果。

检测原理及公式如下：

假设有  $r$  个分组，每组的标准差为由这些标准差还可以计算出  $MSe$ ，统计量近似服从  $r-1$  的卡方分布

$$K^2 = \frac{1}{c} \left[ (n-r) \ln MSe - \sum_{i=1}^r (n_i - 1) \ln \hat{s}_i^2 \right], \text{ 其中 } c = 1 + \frac{1}{3(r-1)} \left[ \sum_{i=1}^r \frac{1}{n_i - 1} - \frac{1}{n-r} \right]$$

对数据进行主成分分析获得的因子系数表



元件	起始特徵值			擷取平方和載入		
	總計	變異的 %	累加 %	總計	變異的 %	累加 %
1	6.153	18.097	18.097	6.153	18.097	18.097
2	5.125	15.075	33.172	5.125	15.075	33.172
3	3.998	11.758	44.929	3.998	11.758	44.929
4	2.341	6.886	51.816	2.341	6.886	51.816
5	1.747	5.139	56.955	1.747	5.139	56.955
6	1.628	4.789	61.744	1.628	4.789	61.744
7	1.399	4.114	65.858	1.399	4.114	65.858
8	1.301	3.826	69.683	1.301	3.826	69.683
9	1.208	3.553	73.236	1.208	3.553	73.236
10	1.031	3.032	76.268	1.031	3.032	76.268
11	.998	2.934	79.202			
12	.963	2.833	82.035			
13	.861	2.532	84.567			
14	.773	2.273	86.840			
15	.749	2.203	89.044			
16	.666	1.959	91.002			
17	.542	1.595	92.597			
18	.458	1.348	93.945			
19	.403	1.185	95.131			
20	.351	1.032	96.163			
21	.287	.845	97.008			
22	.228	.670	97.677			
23	.199	.587	98.264			
24	.157	.461	98.725			
25	.114	.334	99.059			
26	.086	.254	99.313			
27	.072	.212	99.525			
28	.056	.166	99.690			
29	.045	.132	99.822			
30	.024	.070	99.892			
31	.020	.057	99.950			
32	.008	.023	99.973			
33	.007	.019	99.992			
34	.003	.008	100.000			



KRT14 3861	.921							
KRT17 3872	.827							
COL1A2 127		.974						
8								
COL3A1 128		.962						
1								
COL1A1 127		.949						
7								
SPARC 6678		.926						
FN1 2335		.804						
SFTPA1 6535			.963					
09								
SFTPA2 7292			.961					
38								
FTL 2512								
SFTPC 6440								
HLA-B 3106				.900				
B2M 567				.854				
CD74 972				.708				
IGF2 3481								
ADAM6 8755					.894			
LOC96610 96					.874			
610								
CEACAM6 46						.894		
80								
SFTPB 6439						.778		
CTSD 1509							.844	
SCGB2A2 42								.816
50								

C4A 720										
H19 283120										
ACTB 60										
GAPDH 2597										
PIP 5304										
EEF1A1 1915								.852		
S100A9 6280										
CPB1 1360									.683	
LTF 4057										

擷取方法：主體元件分析。

轉軸方法：具有 Kaiser 正規化的最大變異法。<sup>a</sup>

a. 在 9 疊代中收斂循環。

上表首行的 1-10 为降维处理后的 10 个变量，由实验结果可知各个原有的 RNA 表达情况对新变量的数据贡献率平均在 85%左右，对数据的利用率较高。

根据样本信息文件，论文将信息文件中的患病等级与样本的 RNA 表达情况建立对应关系的新数据表。同时，在该数据中，只保留论文经过降维处理后得到的 10 组变量信息，且只选用样本信息文件中给出具体患病等级的对应关系。得到的具体数据如下图：

1	编号	FAC1	FAC2	FAC3	FAC4	FAC5	FAC6	FAC7	FAC8	FAC9	FAC10	病情
2	73	0.58322	2.8101	0.02665	-0.27649	-0.32361	-0.39776	-0.29127	0.17465	0.15006	-0.21815	4.3
3	83	-0.18897	-0.59745	0.13734	-0.4218	-0.18005	-0.14191	2.21122	0.86658	-0.31422	-0.30677	1
4	84	-0.25404	-0.44069	-0.48126	-0.81461	-0.4118	-0.44575	1.0301	-0.48769	0.12344	-0.48063	1
5	99	-0.59692	-0.00176	-0.37969	-0.91088	-0.21288	0.4235	-0.53944	0.18601	-0.33501	0.02132	1.7
6	100	-0.05931	-0.18394	-0.02204	0.40535	-0.36155	5.03084	0.12369	0.07502	-0.49622	0.40015	2.3
7	101	-0.72065	3.37239	-0.07883	0.84684	-0.08795	-0.43334	1.40283	0.68287	-0.44609	-0.08742	2.3
8	102	-0.19187	-0.67596	-0.54689	1.39901	-0.23283	0.36306	1.0721	0.62447	-0.19896	0.71203	1.3
9	103	-0.00729	-0.5977	-0.06212	0.41348	5.64107	0.07876	0.28124	0.13416	-0.62776	0.71458	1.7
10	105	-0.52702	1.12434	0.22924	0.02993	-0.43782	-0.39312	0.75691	1.14774	0.1603	0.34427	2.7
11	106	-0.53863	0.59535	-0.2817	-0.07183	1.36248	-0.52065	-0.62745	0.137	-0.31923	-0.21879	3.3
12	107	-0.06442	-0.60484	-0.75838	0.59708	0.88105	-0.5433	0.59483	0.1823	2.92883	1.73534	1.3
13	108	-0.51012	0.26352	0.20743	0.3978	0.93302	-0.7857	-1.22704	1.25489	-0.17846	-3.77461	1.3
14	109	-0.3415	-0.19701	-0.48839	0.01575	0.45073	-0.76903	-0.55504	-0.24598	1.26865	-1.15229	1.3
15	110	0.01693	-0.5092	0.01296	4.99884	-0.7167	0.30459	-0.2524	-0.88585	-0.93651	-1.29065	1.3
16	111	-0.33262	1.27191	0.22484	-0.52342	0.30985	3.12289	-0.23291	0.18359	0.73404	-0.51791	1.7
17	112	-0.51504	-0.36361	0.16231	0.76169	-0.47147	-1.31216	-0.82986	0.05721	0.27656	3.2913	3.3
18	113	-0.31638	0.49651	0.29112	0.95867	-0.2838	-0.15879	-1.15853	-0.15281	0.98014	1.69783	3.3
19	114	-0.48548	-0.02192	-0.6072	-0.14984	0.08972	-0.26982	-0.90762	-0.67394	-0.72171	-0.69567	3.3
20	115	-0.40354	0.1263	-0.25496	0.18603	0.97117	-0.20034	-0.06001	0.66173	0.55854	0.08678	2.3
21	116	-0.28095	-0.40496	-0.42422	-0.75936	-0.45196	-0.34254	-0.52433	-1.30504	-0.52304	-0.25971	1.7
22	117	-0.13569	-0.50836	-0.77125	0.2055	-0.87462	0.69988	-0.15525	0.73733	3.24685	0.2363	1.3
23	132	1.7367	-0.73855	-0.02589	-0.31448	-0.56791	0.10397	-1.02269	-0.54671	-1.37229	0.56833	3
24	133	4.23398	-0.22815	1.17482	-0.23659	0.32123	0.03367	0.57039	-2.80408	0.97658	-0.43528	4.3
25	134	0.20203	-1.27808	-1.32603	-0.75946	-0.73391	-0.35318	-1.32958	1.91384	-2.3196	0.88319	4.3
26	135	1.42371	-1.19671	-0.54313	-0.32699	-0.62129	-0.23858	-0.38229	1.81136	-0.56507	0.50897	2
27	136	-0.0178	-0.72309	-0.71671	-0.73095	-0.18549	-0.56689	-0.58744	-0.20147	-0.40709	-0.37904	4.3
28	137	4.03376	1.72842	-0.86393	0.30848	0.04572	-0.25263	-0.04799	2.29984	0.01571	-0.07697	3
29	160	-0.43948	0.25218	2.95923	-0.12568	0.11402	-0.219	-0.23831	1.573	0.0464	0.21013	4
30	161	-0.25699	-0.44708	4.80318	0.07769	-0.18751	-0.16651	-0.42306	-0.08879	-0.20042	0.17945	4
31	162	-0.25352	-0.45517	-0.46497	-0.49954	-0.55733	-0.36895	3.95968	0.35459	-0.33835	-0.26931	4
32	163	-0.23668	-0.64667	1.07403	-0.42241	-0.37626	-0.38371	0.84358	-0.62615	-0.27925	-0.12508	3
33	164	-0.15838	-0.46513	0.58143	-0.63763	-0.28524	-0.29869	1.60011	-0.18242	-0.24603	-0.12618	3
34	167	-0.12379	-0.84271	0.56531	-0.80203	-0.37009	-0.21542	1.33123	0.02269	-0.17045	-0.43153	3
35	202	-0.32185	-0.45955	0.11446	0.3744	-0.32256	0.07873	0.19353	-0.15567	-0.52795	0.44511	4
36	203	-0.44457	0.28745	0.03426	-0.34061	-0.36332	-0.06525	-1.06818	-0.79695	0.99669	0.48389	1.7
37	204	-0.4996	0.75684	0.29012	-0.54684	-0.31956	-0.46949	-0.64938	0.58425	-0.12594	0.12625	1.7
38	206	-0.32463	-0.69945	-0.67506	-0.41045	-0.78515	-0.41043	-0.61043	-0.40745	1.97244	-1.90806	2.7
39	207	-0.16425	-0.84003	-0.7115	-0.59112	1.36137	0.34365	0.25101	-1.09423	0.34183	-0.10506	1.3

首行中的 FAC1-10，为降维处理后的可代替原有 RNA 表达情况的数据。

针对上图所分析处理的数据条件，文章建立关键 RNA 序列的表达情况与患病程度的关系，此部分论文采用回归分析的方法，探究各个 RNA 序列在决定患病程度上的权重。

$$Y = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

论文确定 10 个变量的关系矩阵，同时 y 值为患病程度。

利用 spss，对上述问题进行回归分析，分析结果如下：

模型摘要<sup>b</sup>

模型	R	R 平方	調整後 R 平方	標準偏斜度錯誤	變更統計資料				
					R 平方變更	F 值變更	df1	df2	顯著性 F 值變更
1	.592 <sup>a</sup>	.351	.160	.9655	.351	1.837	10	34	.091

a. 預測值: (常數), [%1:, FAC10:

b. 應變數\ : 病情

係數<sup>a</sup>

模型	非標準化係數		標準化係數	T	顯著性	相關			共線性統計資料	
	B	標準錯誤	Beta			零階	部分	部分	允差	VIF
1 (常數)	2.550	.144		17.697	.000					
FAC1	.242	.145	.231	1.673	.103	.234	.276	.231	.999	1.001
FAC2	.131	.145	.125	.902	.374	.125	.153	.125	.999	1.001
FAC3	.336	.145	.321	2.326	.026	.325	.370	.321	.999	1.001
FAC4	-.167	.145	-.160	-1.154	.257	-.161	-.194	-.159	.999	1.001
FAC5	-.209	.144	-.200	-1.450	.156	-.201	-.241	-.200	.999	1.001
FAC6	-.189	.144	-.181	-1.313	.198	-.182	-.220	-.181	1.000	1.000
FAC7	-.045	.144	-.043	-.310	.758	-.043	-.053	-.043	1.000	1.000
FAC8	-.038	.147	-.035	-.255	.800	-.036	-.044	-.035	.998	1.002
FAC9	-.205	.144	-.197	-1.427	.163	-.197	-.238	-.197	1.000	1.000
FAC10	.197	.144	.189	1.364	.181	.190	.228	.189	1.000	1.000

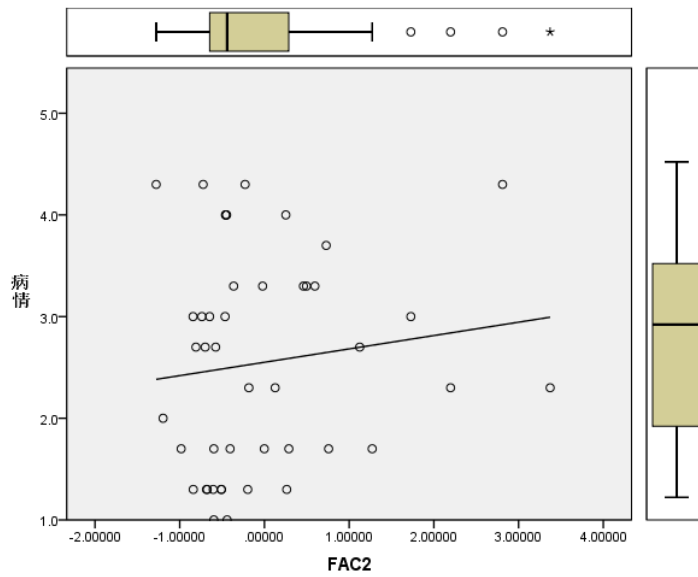
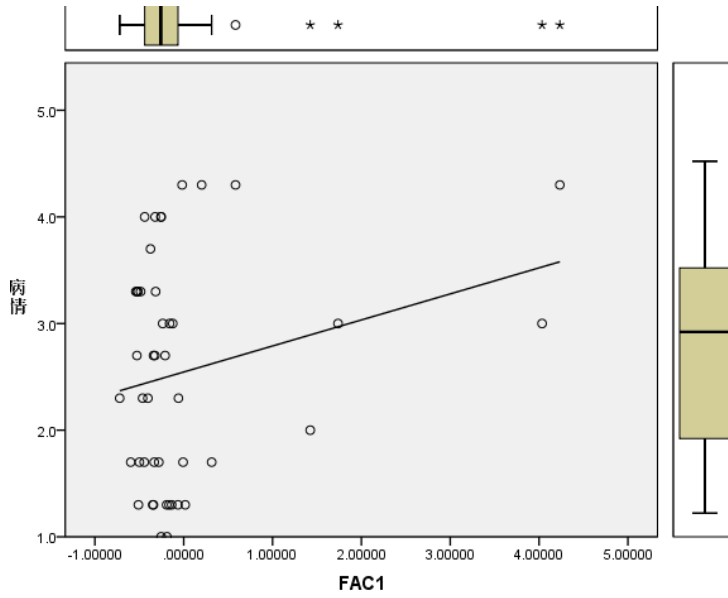
a. 應變數\ : 病情

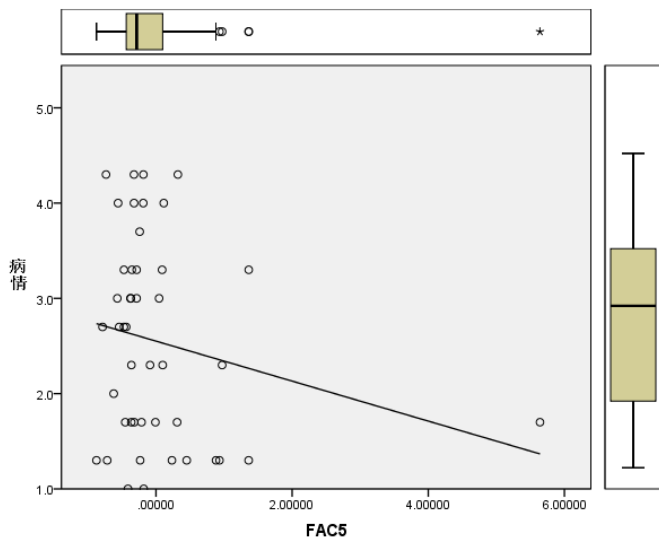
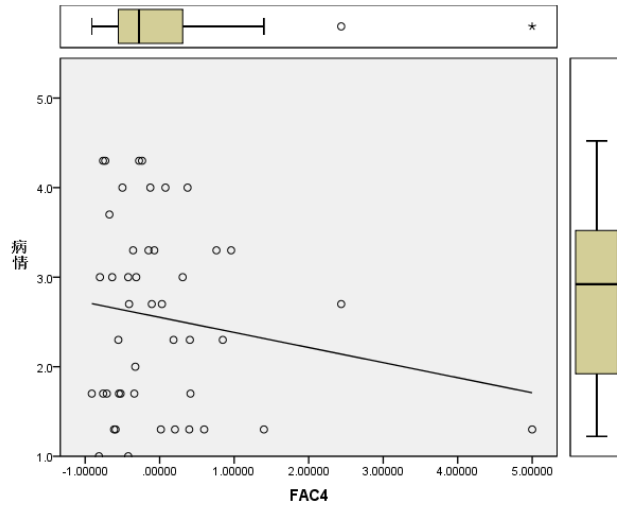
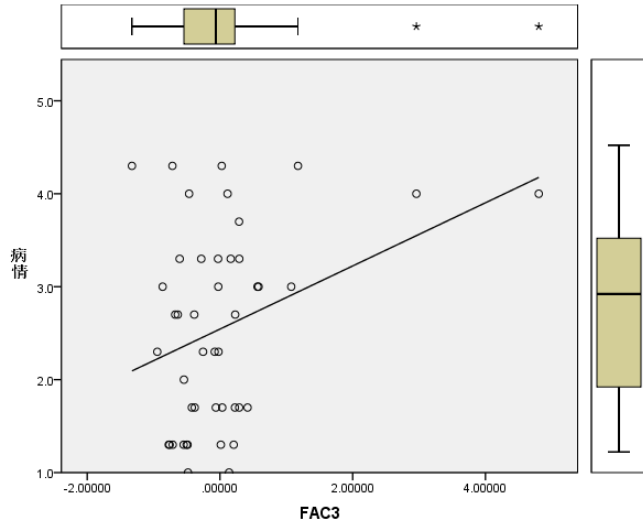
由结果知, 设患病程度为 Y, FAC1,FCA2,⋯,FCA10 分别为  $x_1, x_2, \dots, x_{10}$ , 则

$$Y=0.242x_1+0.131x_2+0.336x_3-0.167x_4-0.209x_5-0.189x_6-0.045x_7-0.38x_8-0.205x_9+0.197x_{10}+2.550$$

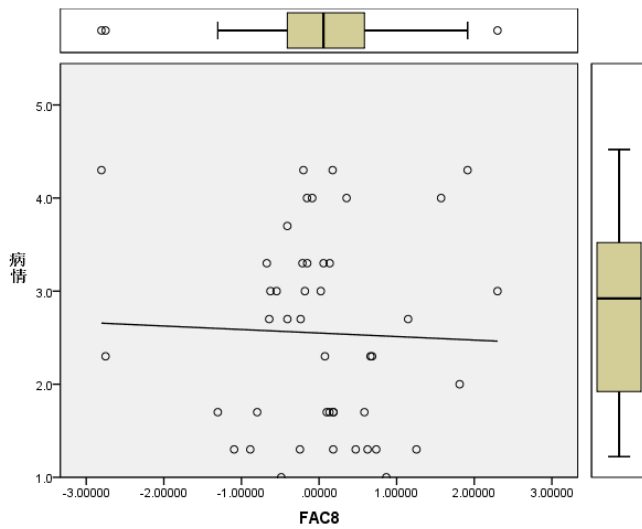
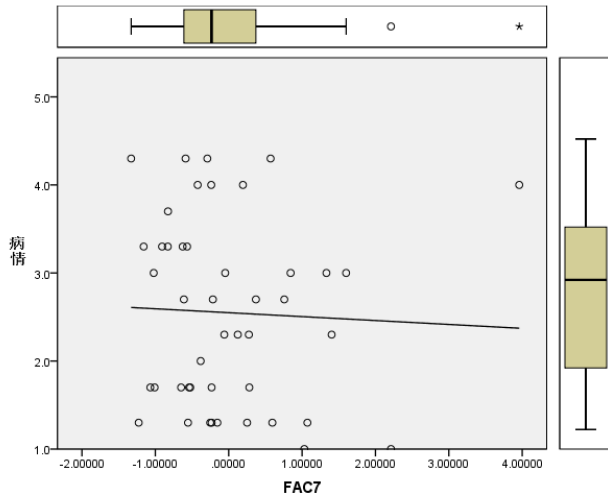
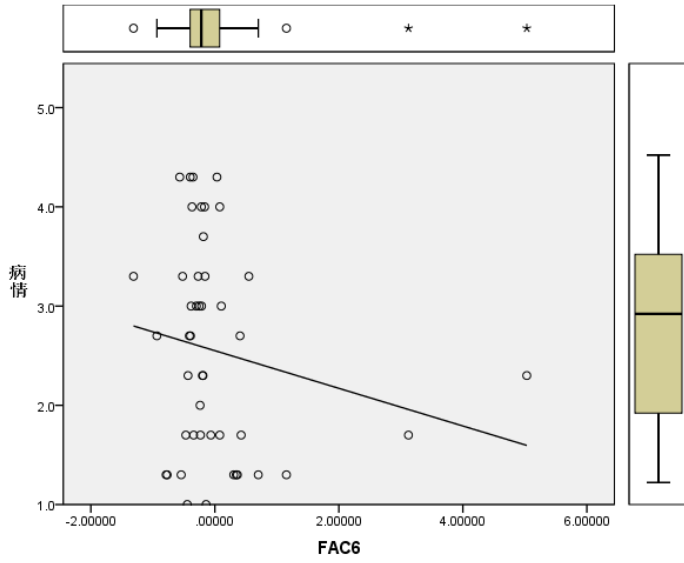
由上式知 FAC1, FAC2, FAC3, FAC10 的表达对患病有促进作用, 而 FAC4,FAC5, FAC6, FAC7, FAC8, FAC9 的表达对患病有抑制作用。其中 FAC3 促进作用最大, FAC5 的抑制作用最大。

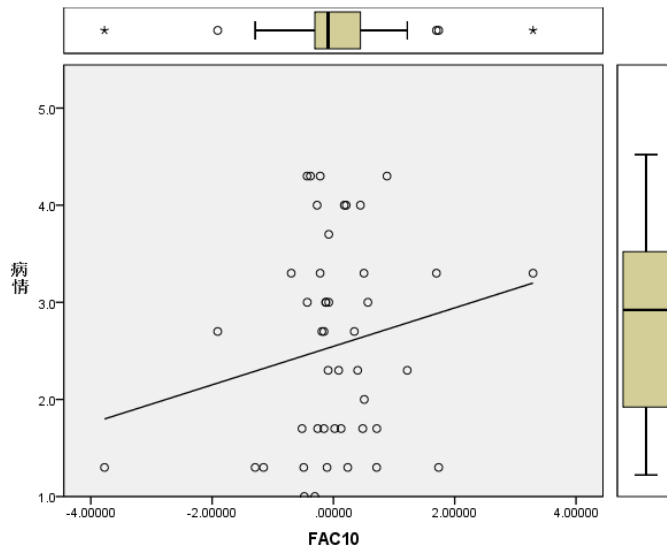
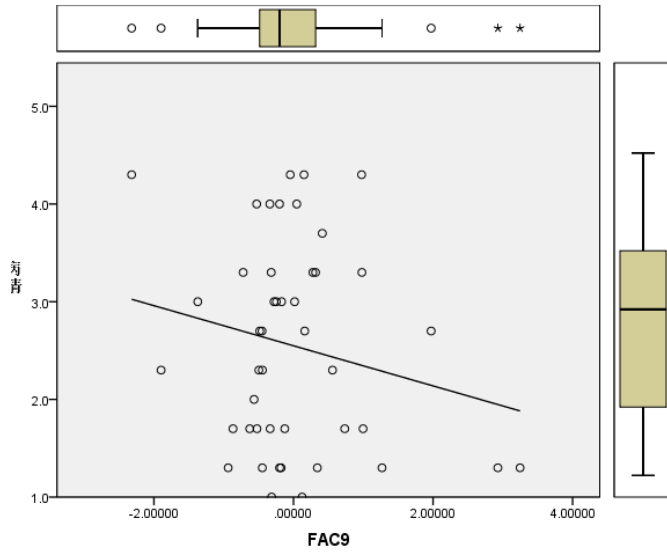
相关单个变量对患病影响的回归曲线如下图









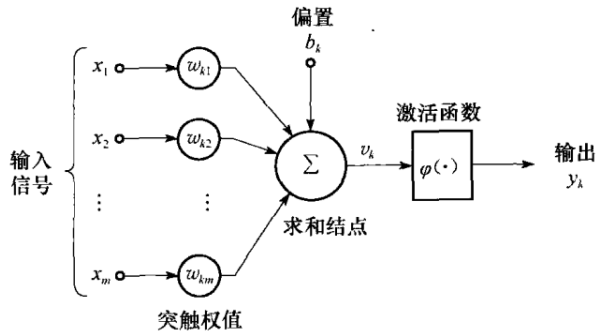


## 二、建立 BP 神经网络模型对病人患病程度进行预测

患者的患病程度应与其各组数据呈函数关系，而经过训练的人工神经网络可以对任意函数进行拟合，因此以三组数据作为输入，训练后的神经网络的输出即为预测的患者患病程度。在搭建针对该问题的具体神经网络模型之前，先对基本的多层神经网络及反向传播算法进行介绍。

### (一) 多层神经网络

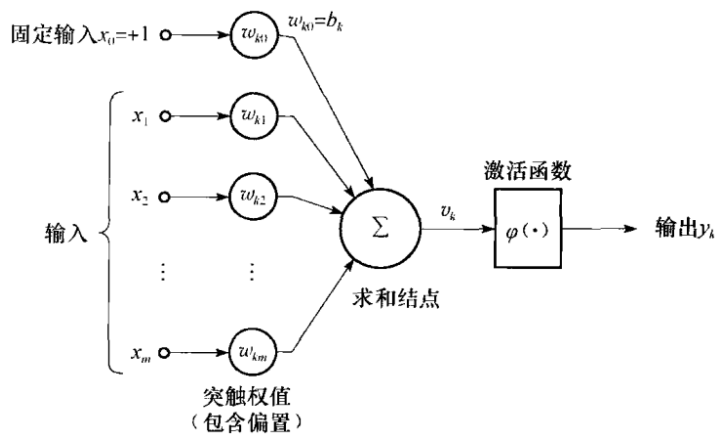
人工神经网络的基本单元是神经元模型，如图所示



其输出为：

$$y_k = \varphi(\mathbf{x}_k \cdot \mathbf{w}_k + b_k)$$

可改进为：

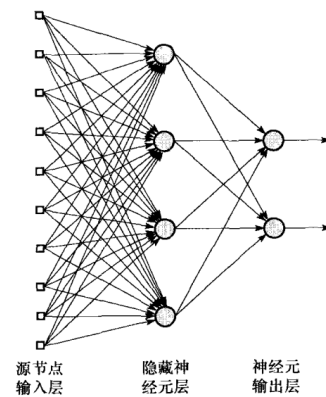


其输出为：

$$y_k = \varphi(\mathbf{x}_k \cdot \mathbf{w}_k)$$

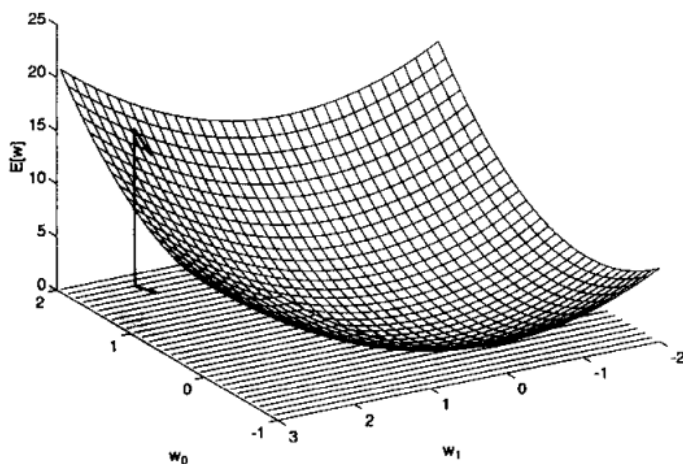
多层神经网络为多个神经元模型的组合，有一个输入层，中间有一个或多个隐含层，有一个输出层，每一层的输入为上一层的输出的组合。

神经网络的学习为有监督学习即训练数据的每组输入都给出一个预期输出，每次学习时，以该次的输出与预期输出之间的误差大小为标准来调整各个连接边的权重，使网络更加拟合训练数据的特征，输出不断接近预期输出，从而实现网络相应的功能。为定量地分析与调整权重，在此定义误差函数：



$$E(\bar{w}) = \frac{1}{2} \sum_{d \in D} \sum_{k \in \text{outputs}} (t_{kd} - o_{kd})^2$$

误差  $E$  是关于网络各边权重的函数，当只有两个权重时，误差曲面如图所示。由此可引伸至更高维的误差曲面。



当误差函数的值为全局最小值时网络对训练数据的拟合达到最佳。为使误差达到最小并减少计算量，我们采用随机梯度下降算法，每次随机选取一个训练样例  $d$ ，计算关于  $d$  的误差  $E$ ，此时的误差  $E$  为：

$$E_d(\mathbf{w}) = \frac{1}{2} \sum_{k \in \text{outputs}} (t_{kd} - o_{kd})^2$$

对  $w_i$  求偏导  $\frac{\partial E}{\partial w_i}$ ，则第  $i$  个权重更新为：

$$w_i = w_i - \eta \frac{\partial E}{\partial w_i}$$

随机梯度下降算法要求误差函数可微，因此，选取 *sigmoid* 函数作为神经元的激活函数。记  $\sigma(x)$  为 *sigmoid* 函数：

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

其导数为：

$$\frac{d\sigma(x)}{d(x)} = \sigma(x) \cdot [1 - \sigma(x)]$$

## (二)反向传播算法

首先对叙述与推导中用到的变量及其下标做出说明：

$x_{ji}$ : 单元  $j$  的第  $i$  个输入

$w_{ji}$ : 与单元  $j$  的第  $i$  个输入相关联的权重

$net_j$ :  $net_j = \sum_i w_{ji} x_{ji}$ , 即单元  $j$  的输入的加权和

$o_j$ : 单元  $j$  计算出的输出

$t_j$ : 单元  $j$  的目标输出

$\sigma$ : *sigmoid* 函数

*outputs*: 网络输出单元的集合

*Downstream(j)*: 单元的直接输入中包含单元  $j$  的输出的单元的集合

由于训练样例仅对网络的输出提供了目标值  $t_k$ , 缺少直接的目标值来计算隐藏单元的误差值, 从而无法更新隐藏层与隐藏层、隐藏层与输出层之间的权重。为此采用反向传播算法间接计算隐藏单元的误差。

在此引入误差项  $\delta_i$ , 其意义类似于预期输出与实际输出之差( $t_i - o_i$ ), 具体表述及推导在后面说明, 在此以三层神经网络叙述利用反向传播算法与随机梯度下降算法训练多层神经网络的步骤。其中, 记每一个训练样例的形式为序偶  $\langle \mathbf{x}, t \rangle$ ,  $\mathbf{x}$  是网络输入值向量,  $t$  是期望输出值。

训练步骤:

步 1: 创建具有  $n_{in}$  个输入,  $n_{hidden}$  个隐藏单元,  $n_{out}$  个输出单元的网络

步 2: 初始化所有的网络权重为较小的随机值

步 3: 若达到训练结束条件, 训练结束。若未达到, 则转步 4

步 4: 随机选取训练样例  $\langle \mathbf{x}_d, t_d \rangle$ , 将  $\mathbf{x}$  输入网络, 计算网络中每一个单元  $u$  的输出  $o_u$

步 5: 对于网络的每一个输出单元  $k$ , 计算它的误差项  $\delta_k$ :

$$\delta_k \leftarrow o_k(1 - o_k)(t_k - o_k)$$

步 6: 对于网络中的每个隐藏单元  $h$ , 计算它的误差项  $\delta_h$ :

$$\delta_h \leftarrow o_h(1 - o_h) \sum_{k \in \text{outputs}} w_{kh} \delta_k$$

步 7: 更新每个网络权重  $w_{ji}$ :

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji}$$

其中

$$\Delta w_{ij} = \eta \delta_j x_{ji}$$

步 8: 转步 3

反向传播算法的推导:

用随机梯度下降算法迭代处理训练样例时需计算

$$w_{ji} = w_{ji} - \eta \frac{\partial E_d}{\partial w_{ji}}$$

现在开始推导  $\frac{\partial E_d}{\partial w_{ji}}$

首先, 注意权值  $w_{ji}$  仅能通过  $net_j$  影响网络的其他部分。所以我们可以使用链式法则得到

$$\frac{\partial E_d}{\partial w_{ij}} = \frac{\partial E_d}{\partial net_j} \frac{\partial net_j}{\partial w_{ij}} = \frac{\partial E_d}{\partial net_j} x_{ji}$$

此时需推导  $\frac{\partial E_d}{\partial net_j}$ , 需分两种情况, 一种情况是单元  $j$  是网络的一个输出单

元, 另一种情况是  $j$  是网络中的隐藏层中的单元。

### 情况 1: 输出单元的权重训练法则

注意  $net_j$  仅能通过  $o_j$  影响网络, 利用链式法则得:

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j} \frac{\partial o_j}{\partial net_j}$$

首先考虑上式第一项:

$$\frac{\partial E_d}{\partial o_j} = \frac{\partial}{\partial o_j} \frac{1}{2} \sum_{k \in \text{outputs}} (t_k - o_k)^2 = \frac{\partial}{\partial o_j} \frac{1}{2} (t_j - o_j)^2 = -(t_j - o_j)$$

第二项:

$$\frac{\partial o_j}{\partial net_j} = \frac{\partial \sigma(net_j)}{\partial net_j} = o_j(1 - o_j)$$

所以:

$$\frac{\partial E_d}{\partial net_j} = -(t_j - o_j) o_j(1 - o_j)$$

在之后，以误差项  $\delta_i$  代表任意单元  $i$  的  $-\frac{\partial E_d}{\partial net_i}$ ，即

$$\delta_i = -\frac{\partial E_d}{\partial net_i} = (t_i - o_i)o_i(1 - o_i)$$

## 情况 2：隐藏单元的权重训练法则

对于网络中的内部单元或者说隐藏单元的额情况，推导  $w_{ji}$  必须考虑  $w_{ji}$  间接地影响网络输出，从而影响  $E_d$ 。

$net_j$  只能通过  $Downstream(j)$  中的单元影响网络输出再影响  $E_d$ ，所以有：

$$\begin{aligned} \frac{\partial E_d}{\partial net_j} &= \sum_{k \in Downstream(j)} \frac{\partial E_d}{\partial net_k} \frac{\partial net_k}{\partial net_j} \\ &= \sum_{k \in Downstream(j)} -\delta_k \frac{\partial net_k}{\partial net_j} \\ &= \sum_{k \in Downstream(j)} -\delta_k \frac{\partial net_k}{\partial o_j} \frac{\partial o_j}{\partial net_j} \\ &= \sum_{k \in Downstream(j)} -\delta_k w_{kj} \frac{\partial o_j}{\partial net_j} \\ &= \sum_{k \in Downstream(j)} -\delta_k w_{kj} o_j (1 - o_j) \end{aligned}$$

以  $\delta_j$  表示的  $-\frac{\partial E_d}{\partial net_j}$ ，即

$$\delta_j = o_j(1 - o_j) \sum_{k \in Downstream(j)} \delta_k w_{kj}$$

综上所述，便得到任意权值的更新法则，即：

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji}$$

其中

$$\Delta w_{ij} = \eta \delta_j x_{ji}$$

### (三) 搭建针对该问题的具体神经网络模型

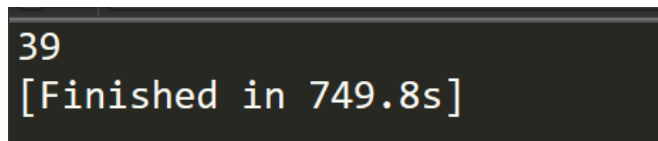
针对该问题，我们建立了一个三层的神经网络，输入层，一层隐藏层及输出层。

选取 RNA 表达量差异最大，即方差最大的前 600 组数据及性别作为输入，则输入层有 601 个输入，隐藏层选取 50 个单元。患病程度由数据可知共有 11 种，

故输出层选 11 个单元，将预期输出设置为当患病程度为第 k 种时，第 k 个输出单元输出 1，其他单元输出 0。

由上述要求，得到完整的训练数据，进行训练。采用随机梯度下降法，随机抽取 10000 次数据训练。

由于只有一份数据，而模型训练与模型验证需用两份数据。故在此采用交叉验证的方式进行验证。从数据中随机抽取 10 组作为验证组，其余组作为训练组，用训练组数据训练完之后用 10 组训练组的数据进行验证，如上重复十次，输出这累计 10 次的验证组数据（即 100 组）中网络输出结果正确的个数。如图：



```
39
[Finished in 749.8s]
```

故准确率为 39%

最后再以全部数据进行训练得到最终的模型。

#### （四）模型预估样本患病性能评价体系

为了对预估样本患病系统的效果做分析，我们需要一个评价体系来进行评估。论文以此借用文本分类的相关指标。

首先做如下假设：待测试样本集合中共有 N 组数据 TEAM, A、C：实际为患病 TEAM；B、D：实际为正常 TEAM。则  $N=A+B+C+D$ ； $N_s=A+C$  为实际患病 TEAM 的数目； $N_t=B+D$  为实际的正常 TEAM 数目。而预估样本患病系统对其作出的判定结果为：判定为患病 TEAM 的是 A、B；判定为正常 TEAM 的是：C、D。

根据以上假设，定义如下几个评价指标来衡量预估样本患病系统的性能。

##### 1. 召回率 (RccaU)：

$R=A/N_s=A/(A+C)$ ，即患病 TEAM 的检出率。这个指标反映了系统发现患病 TEAM 的能力，召回率越高，则说明患病 TEAM 的“漏网率”越低；

##### 2. 正确率 (Precision)：

$P=P/(A+B)$ ，即是患病 TEAM 的检对率。该指标反映了系统找对的能力，正确率越高，则将正常 TEAM 误判为患病 TEAM 的数量就越小；

##### 3. 精确率 (Accuracy)：

$Accur=(A+D)/N$ ，即对所有 TEAM (包括患病 TEAM 和正常 TEAM) 的判对率；

##### 4. 错误率 (ErrorRate)：



$Err=(B+C)/N$ ，即对所有 TEAM(包括患病 TEAM 和正常 TEAM)的判错率；

5. F 值：

$F=2*P*R/(R+P)$ ，该值实际上是召回率和正确率的调和平均，它将召回率 和正确率综合成一个判定指标。

对预估样本患病性能好坏的评估通常有两个指标：准确率和查全率。准确率是所有判断的文本中与人工分类结果吻合的文本所占的比率，其数学公式表示如下：

准确率=分类正确的文本数 / 实际分类的文本数

查全率是指人工分类结果应有的文本中分类系统吻合的文本所占的比率，其数学公式如下：

查全率=分类正确的文本数 / 应有的文本数

准确率和查全率反映了分类质量的两个方面，两者必须综合考虑，故在实际应用中，人们通常还采用 F1 指标值来评价分类器的好坏，其数学公式如下：

$F1 \text{ 测试值}=(\text{准确率} \times \text{查全率}+2) / (\text{准确率}+\text{查全率})$

根据实验结果，当 F1 达至  $U_{so}\%$  以上时，就可看作对 TEAM 进行了比较准确的判断。

## 六、模型改进

神经网络经过大规模数据训练可以得到很好的拟合效果，但是该问题给出的数据规模较小，拟合效果达不到最佳，如若可以得到规模更大、更多样化的数据，该模型的质量将大大提升。

此外，该神经网络中训练次数为随机定的比较适中的值，但是不同的值可能会对结果产生影响，当训练次数过多时，网络将会对训练数据产生过拟合现象，即针对训练数据网络均能得到基本正确的输出，但是当输入为训练数据以外的数据时，网络将得到不正确的输出。而当训练次数较少，网络的拟合效果将会很差。举例来说，我们需要训练一个网络来判断一个物体是否为树叶，假若所有的训练数据都是绿色的树叶，并且大部分树叶的边缘是带锯齿的。当训练次数过多时，网络会认为凡是树叶，其边缘一定有锯齿，边缘不带锯齿的树叶不会被判定为树叶；训练次数较少时，网络会认为只要是绿色的都是树叶。因此，还需要更多的

实验来确定最佳的训练程度使网络准确率得到提升。

## 七、模型推广

本文所提到的模型可以推广至用来进行病理分析与疾病判断。因为每个人的患病与否都是与体内各种指标、基因、个人作息、饮食等相关的，所以可以将人的患病看做一个关于上述各种因子的函数。而本文所提出的模型就是基于人体内各种基因的情况来分析疾病原因并作出患病程度的判断。故只需稍加改进，便可很容易推广至上述领域中。

### 参考文献：

- [1] 索伦森. 《蛋白质与蛋白质组学》 2007
- [2] 邢凯. 《整合转录组及全基因组重测序方法鉴定影响猪脂肪沉积的关键基因及其变异》 《中国农业大学》 2015
- [3] 威廉·费勒 著；胡迪鹤 译 《概率论及其应用》
- [4] 徐野《复杂互联系统与网络鲁棒性研究》 2015年9月
- [5] David Freedman 等著，魏宗舒，施锡铨等译 《统计学》
- [6] 张晓辉，李莹，王华勇，赵宏 应用特征聚合进行中文文本分类的改进 KNN 算法
- [7] Mitchell, Tom M., J. G. Carbonell, and R. S. Michalski. *Machine learning.. Machine Learning.* Springer US, 1986:417-433.