**周总结（2020.11.15）**

**1 代码的实现**

rm(list = ls())

source("functions0725.R")

# 导入全血样本数据

load("WB.Rdata")

# 导入真实比例

load("WB\_Truepro.Rdata")

# 取消科学计数法

options(scipen = 200) #scipen 表示在200个数字以内都不使用科学计数法

CD4T = WB[,7:12]

CD8T = WB[,13:18]

Monocyte = WB[,19:24]

B = WB[,25:30]

NK = WB[,31:36]

Neutrophil = WB[,37:42]

Eosinophil = WB[,43:48]

# 求平均值，导出每种细胞类型的列表数据

Mean\_cell <- function(x){

W\_1 = matrix(0,nrow = nrow(x),ncol = 1)

for(i in 1:nrow(x)){

W\_1[i,] <- mean(x[i,])

}

return(W\_1)

}

#得到处理后各细胞类型的数据

CD4T\_dell <- as.matrix(apply(CD4T,1,mean))

CD8T\_dell <- as.matrix(apply(CD8T,1,mean))

Monocyte\_dell <- as.matrix(apply(Monocyte,1,mean))

B\_dell <- as.matrix(apply(B,1,mean))

NK\_dell <- as.matrix(apply(NK,1,mean))

Neutrophil\_dell <- as.matrix(apply(Neutrophil,1,mean))

Eosinophil\_dell <- as.matrix(apply(Eosinophil,1,mean))

WB\_Y = WB[,1:6]

#feat = select\_feature(mat = as.matrix(WB\_Y),method = "cv",nmarker = 1000,startn = 0)

WB\_W = cbind(CD4T\_dell,CD8T\_dell,Monocyte\_dell,B\_dell,NK\_dell,Neutrophil\_dell,Eosinophil\_dell)

colnames(WB\_W) = c("CD4T","CD8T","Monocyte","B","NK","Neutrophil","Eosinophil")

rownames(WB\_W) = rownames(WB\_Y)

# Real data verification,

verAll <- function(Y,W,W1.index = 1:6,H,type = "ME",method = "all",featsize,samplenum = 6){

result = list()

# 提取特征位点

feat = select\_feature(mat = as.matrix(Y),method = "cv",nmarker = featsize,startn = 0)

Y = as.matrix(Y[feat,])

W = as.matrix(W[feat,])

W1 = as.matrix(W[feat,W1.index])

if(method == "all" | method == "RB"){

cat("Running RB ...\n")

RBout = c()

out <- RB(Y = Y, W = W1)

H1 = as.matrix(H)

H1 <- H1[,W1.index]

H\_AbsBias = mean(colSums(abs(t(H1\*0.01) - out$H)) / ncol(H1))

H\_corr = mean(diag(cor(H1\*0.01,t(out$H))))

H\_featnum = featsize

RBout = rbind(RBout,c(H\_AbsBias,H\_corr,H\_featnum))

colnames(RBout) = c("H\_AbsBias","H\_corr","H\_featnum")

result$RBout = RBout

}

if(method == "all" | method == "DCQ"){

cat("Running DCQ ...\n")

DCQout = c()

out <- estCellPercent.DCQ(refExpr = W1,geneExpr = Y)

H1 = as.matrix(H)

H1 <- H[,W1.index]

H\_AbsBias = mean(colSums(abs(t(H1\*0.01) - out[W1.index,]\*0.01)) / ncol(H1))#预测细胞比例矩阵与真实细胞比例矩阵得到的平均绝对误差

H\_corr = mean(diag(cor(H1\*0.01,t(out[W1.index,]\*0.01)))) #计算相关系数

H\_featnum = featsize

DCQout = rbind(DCQout,c(H\_AbsBias,H\_corr,H\_featnum))

colnames(DCQout) = c("H\_AbsBias","H\_corr","H\_featnum")

result$DCQout = DCQout

}

if(method == "all" | method == "RF"){

cat("\nRunning RF ...\n")

RFout = c()

out <- RF(Y = Y,W = W1,K = ncol(W1),type = "ME",iters = 1000)

H1 = as.matrix(H)

H1 <- H[,W1.index]

H\_AbsBias = mean(colSums(abs(t(H1\*0.01) - out$H)) / ncol(H1))

H\_corr = mean(diag(cor(H1\*0.01,t(out$H))))

# W\_corr

H\_featnum = featsize

RFout = rbind(RFout,c(H\_AbsBias,H\_corr,H\_featnum))

colnames(RFout) = c("H\_AbsBias","H\_corr","H\_featnum")

result$RFout = RFout

}

if(method == "all" | method == "PR"){

cat("\nRunning PR ...\n")

PRout = c()

out <- PR(Y = Y,W = W,W1 = W1,type = "ME",K = ncol(W),iters = 1000)

H1 = as.matrix(H)

H\_AbsBias = mean(colSums(abs(t(H1\*0.01) - out$H)) / ncol(H1))

H\_corr = mean(diag(cor(H1\*0.01,t(out$H))))

# W\_corr = mean(diag(as.matrix(cor(W.new[,W2.index],out$W[,W2.index]))))

H\_featnum = featsize

PRout = rbind(PRout,c(H\_AbsBias,H\_corr,H\_featnum))

colnames(PRout) = c("H\_AbsBias","H\_corr","H\_featnum")

result$PRout = PRout

}

if(method == "all" | method == "CBS"){

cat("\nRunning CBS ...\n")

CBSout = c()

out <- DoCBS(Y = Y,W = W1)

H1 = as.matrix(H)

H1 <- H1[,W1.index]

H\_AbsBias = mean(colSums(abs(t(H1\*0.01) - t(out))) / ncol(H1))

H\_corr = mean(diag(cor(H1\*0.01,out)))

H\_featnum = featsize

CBSout = rbind(CBSout,c(H\_AbsBias,H\_corr,H\_featnum))

colnames(CBSout) = c("H\_AbsBias","H\_corr","H\_featnum")

result$CBSout = CBSout

}

return(result)

}

out <- verAll(Y = WB\_Y,W = WB\_W,W1.index = 1:5,H = WB\_Truepro,type = "ME",method = "all",featsize = 10000)

**2 运行结果展示**

我目前选取了1000和10000条数据进行测试，其结果并不理想，但PR跑出来的结果固然是比RF要好的。其结果运行如图1和图2所示。

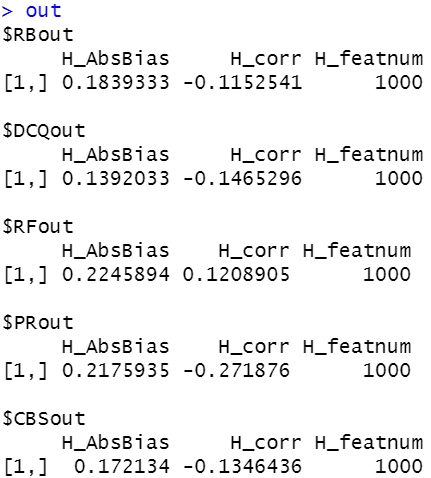


图1 1000条数据所得结果

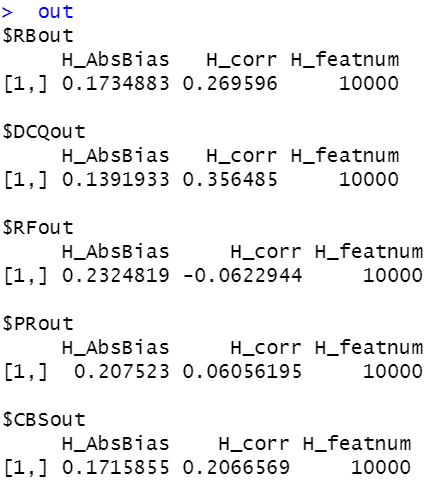


图2 10000条数据运行的结果

结果分析：

图中H\_AbsBias和H\_corr分别代表平均绝对误差和相关系数。图1得到的结果中PR的平均绝对误差排第四，比RF性能要好，而相关性却是最差的。图2得到的结果中PR的平均绝对误差仍然排第四，但相比1000条数据时，效果好了一些，也比RF性能要好，而相关性也提高了，排第四，比RF要好。

1. **问题分析与解决**

a.上述结果是取5个细胞类型的样本所运行出来的结果，接下来可以继续选取5个以上的细胞类型来运行代码。

b.目前虽然能够运行出结果了，但是代码的具体实现部分还有很多没看懂，需要继续学习相关知识以及阅读代码。

c.样本数只选择到了10000条，还可以继续选择20000、50000条数据去运行代码，甚至更高。