煤矿地震的产生原因探究与预测

吴谦亮　　软件 01 2020012357

目录

1 研究背景

煤矿地球物理站的主要任务之一是确定当前地下采矿场所的所谓微震危险程度（特别是称为岩爆的高能破坏性震荡的危险）。岩爆对矿工是一个严重的威胁，并能破坏长墙与设备。地震波灾害是最难探测和预测的自然灾害，在这方面，它可以与地震相媲美。越来越先进的地震和地震声学监测系统使

人们能够更好地了解岩体过程和定义地震灾害预测方法。然而，迄今为止创建的方法的准确性还远远不够完美。地震过程的复杂性和低能量地震事件的数量与高能量现象（如 >10^4J）的数量之间的巨大反差导致统计技术不足以预测地震灾害。

因此，本文通过对某矿区的观测数据进行研究分析，研究与发生地震相关的因子信息，提出如下几个问题：1. 是否存在影响地震发生的重要因子并找到它？2. 利用已有的数据集，根据观测数据预测地震是否会发生。

2 数据介绍和研究思路

2.1 数据介绍

本研究所用的数据来自网站 (https://archive.ics.uci.edu/ml/datasets/seismic-bumps), 提供的来自波兰一个煤矿的两个采用长臂开采法的矿区的高能地震的相关数据，一共包含 2584 条数据，有个变量和一个结果，如下表 1 所示：

变量名	含义	变量名	含义
seismic	通过地震法获得的矿山工作中的轮班地震危险评估结果	nbumps2	前一班次内记录到的地震波的数量 (能量范围 $[10^{2,10}3)$)
seismoacoustic	通过地震声学方法获得的矿山工作中的地震危险评估结果	nbumps3	前一班次内记录到的地震波的数量 (能量范围 $[10^{3,10}4)$)
shift	轮班信息	nbumps4	前一班次内记录到的地震波的数量 (能量范围 $[10^{4,10}45)$)

变量名	含义	变量名	含义
genergy	监测长堤的地震检波器中，最活跃的检波器（GMax）在前一班次记录的地震能量	nbumps5	前一班次内记录到的地震波的数量 (能量范围 $[10^{5,10}6)$)
gpuls	GMax 在上一班次中记录的脉冲数	nbumps6	前一班次内记录到的地震波的数量 (能量范围 $[10^{6,10}7)$)
gdenergy	GMax 在前一班次记录的能量与前八班次记录的平均能量之间的偏差	nbumps7	前一班次内记录到的地震波的数量 (能量范围 $[10^{7,10}8)$)
gdpuls	前一班次记录的脉冲数与前八个班次记录的平均脉冲数之间的 GMax 偏差	nbumps89	前一班次内记录到的地震波的数量 (能量范围 $[10^{8,10}10)$)
ghazard	基于 GMax 登记表的地震声学方法获得的矿井工作中的地震危险评估结果	energy	前一班次内登记的地震波的总能量
nbumps	前一班次内记录到的地震波的数量	maxenergy	前一班次内登记登记的最大的地震能量

其中，seismic、seismoacoustic、shift、ghazard 为类别型变量，将其转化为整数型变量，当做连续变量进行处理 (该操作对后续结果有影响，如后续要

进行更精确实验可以进行其他处理方法)；其余变量均为数值型变量，为了方便实验，将变量 nbumps2 到 nbumps89 移除，只剩下 nbumps。数据没有缺失值。

2.2 研究思路

本研究在对数据进行初步的探索性分析后，首先通过主成分分析对数据进行降维，并提取主成分得分用于可视化。之后，由于数据集已经提供了是否发生高能地震的结果，故可以对这个变量进行判别分析，找到一个相对准确的判别准则。最后，根据地震预测评估准则，对所用样本进行聚类分析，探究不同矿区的地震可能性，并比较它们之间的特征。

3 数据建模和实验结果

3.1 探索性数据分析和数据变换

先观察整体数据分布的特征，如下图所示。结果显示地震评估风险大多集中在 1 和 2 即没有风险或较低风险。同时，虽然观测到的能量和脉冲数都较大，但其与 8 次观测的平均值差值的很小，几乎可以认为没有偏差，说明观测数据在一段时间内都比较稳定。

```
 seismic  seismoacoustic shift       genergy           gpuls
 1:1682   1:1580         0: 921   Min.   :    100   Min.   :   2.0
 2: 902   2: 956         1:1663   1st Qu.:  11660   1st Qu.: 190.0
          3:  48                  Median :  25485   Median : 379.0
                                  Mean   :  90243   Mean   : 538.6
                                  3rd Qu.:  52833   3rd Qu.: 669.0
                                  Max.   :2595650   Max.   :4518.0
    gdenergy           gdpuls         ghazard      nbumps
 Min.   : -96.00   Min.   :-96.000   1:2342   Min.   :0.0000
 1st Qu.: -37.00   1st Qu.:-36.000   2: 212   1st Qu.:0.0000
 Median :  -6.00   Median : -6.000   3:  30   Median :0.0000
 Mean   :  12.38   Mean   :  4.509            Mean   :0.8595
 3rd Qu.:  38.00   3rd Qu.: 30.250            3rd Qu.:1.0000
 Max.   :1245.00   Max.   :838.000            Max.   :9.0000
     energy         maxenergy          class
 Min.   :     0   Min.   :     0   Min.   :0.00000
 1st Qu.:     0   1st Qu.:     0   1st Qu.:0.00000
 Median :     0   Median :     0   Median :0.00000
 Mean   :  4975   Mean   :  4279   Mean   :0.06579
 3rd Qu.:  2600   3rd Qu.:  2000   3rd Qu.:0.00000
 Max.   :402000   Max.   :400000   Max.   :1.00000
```

我们绘制协方差图如下，可以观察到 gernergy 和 gpuls 相关性较强，gdenergy 和 gdpuls 相关性较强，energy 和 maxenergy 相关性较强。

对部分连续性变量绘制直方图观察分布。对于 gernergy 和 gpuls 变量绘制直方图发现数据比较集中在左侧，故进行对数变换后得到如下直方图，可以观测到对数变换后的数据可以大致近似为正态分布，方便后续分析的进行。

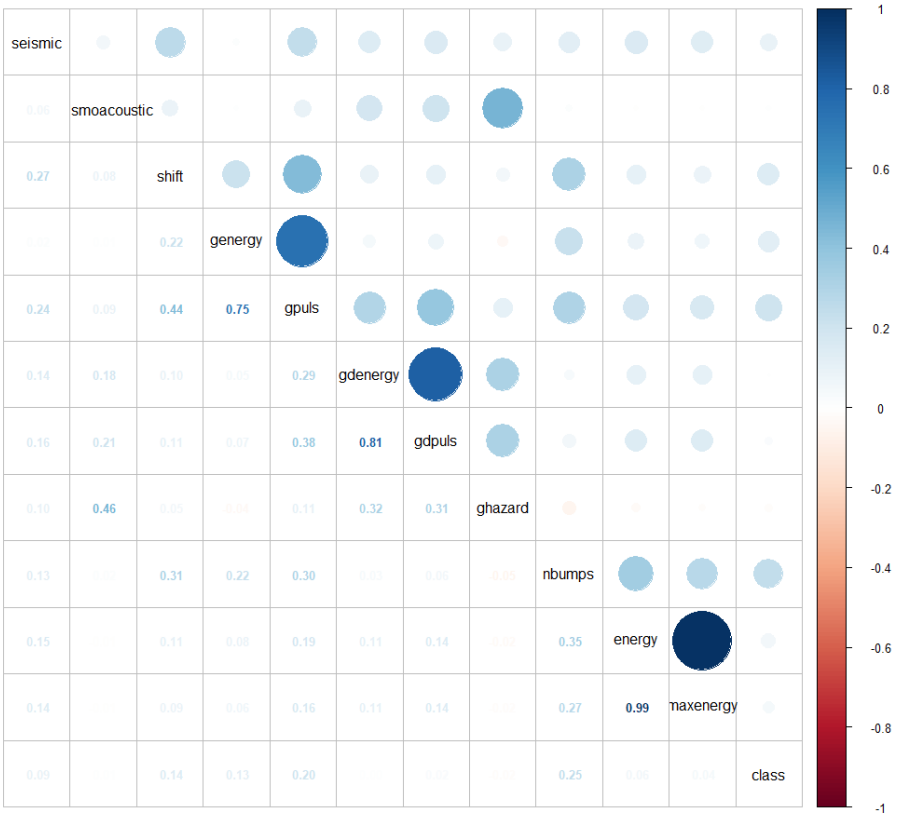
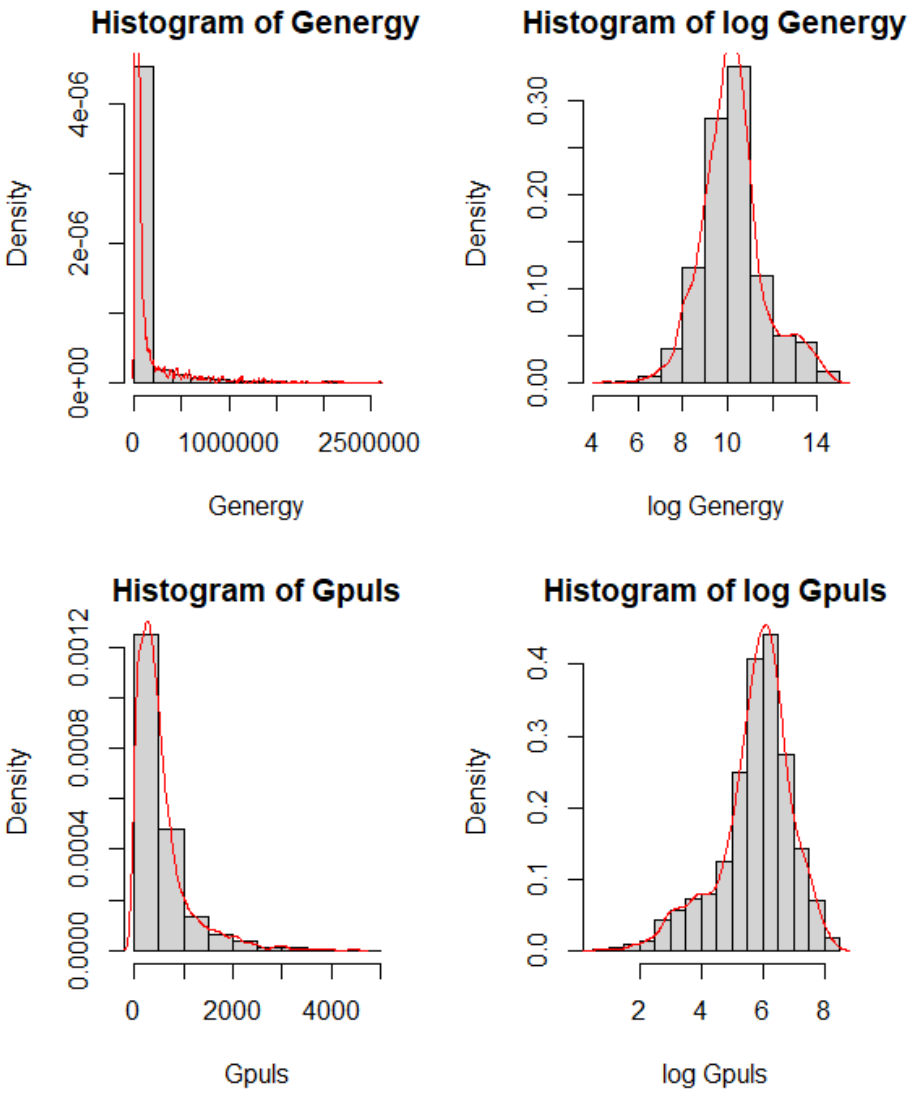
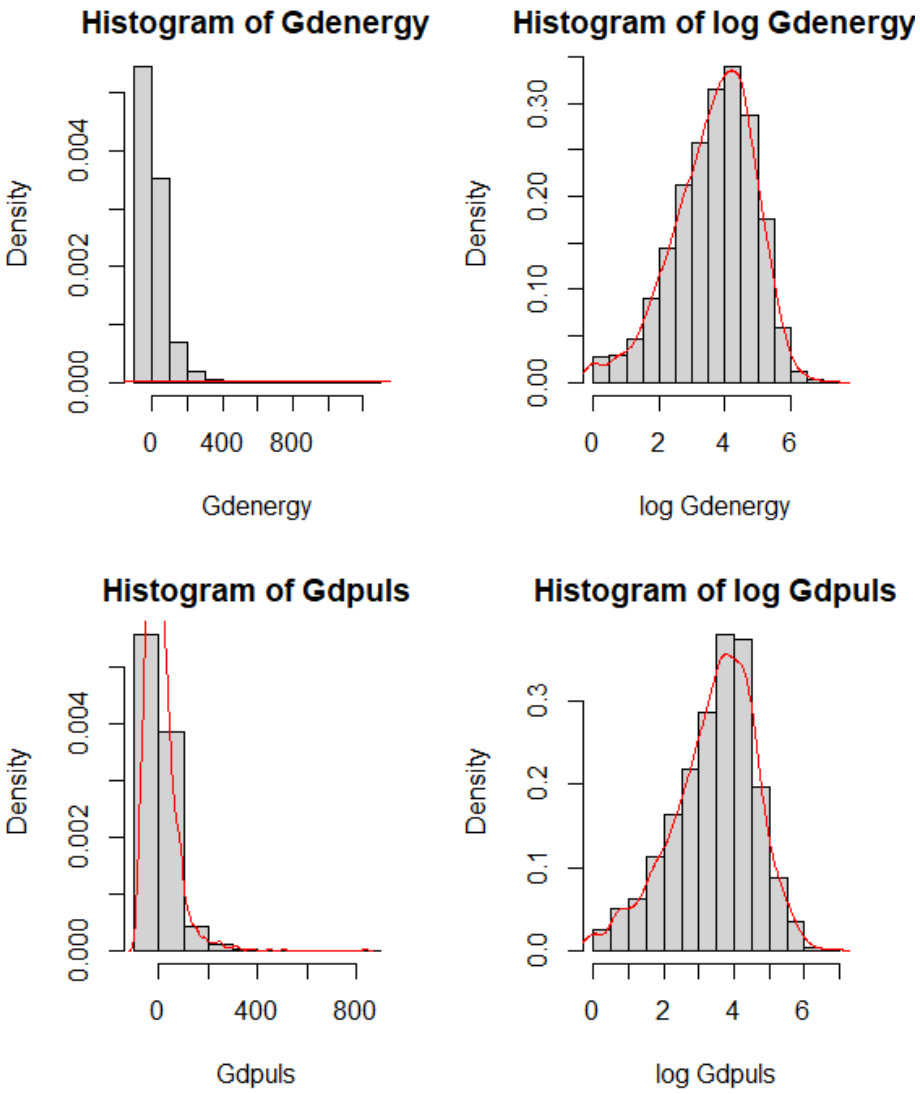


图 1: corr

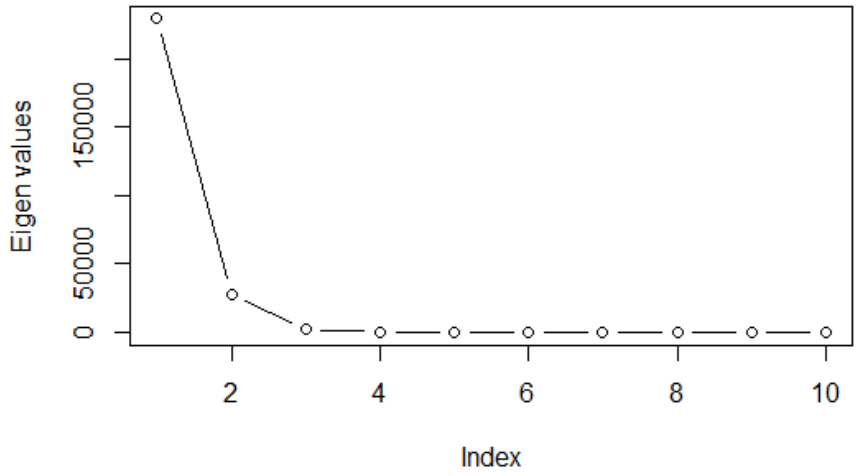


类似的，我们绘制出 gdenergy 和 gdpuls 变量的直方图后发现数据集中在左侧，在进行对数变化后可以近似认为服从正态分布。



3.2 主成分分析与降维

利用 prcomp 函数对连续变量进行主成分分析 (剔除 01 变量 shift)，绘制崖底碎石图如下，我们发现前两个主成分解释占比为 99%, 故只保留前两个主



成分。

得到前两个主成分的得分为:

	PC1	PC2
seismic	0.000	0.000
seismoacoustic	0.000	0.000
genergy	-1.000	0.009
gpuls	-0.002	-0.002
gdenergy	0.000	0.000
gdpuls	0.000	0.000
ghazard	0.000	0.000
nbumps	0.000	0.000
energy	-0.007	-0.726
maxenergy	-0.006	-0.688

通过以上系数可以看出，第一个主成分主要表示了变量 genergy, 即最活跃的检波器（GMax）在前一班次记录的地震能量，这说明了最近时段的检测的最大能量能够很好地反应这个时段可能发生的地震强度；第二个主成分主要为 energy 和 maxenergy, 即前一班次观测的总能量，说明前一时段的总体观测表现也很好地反应了这段时间内地质活动的特征。

3.3 判别分析

在本部分，我们主要关心变量 class 过拟合作为类别标签进行判别分析，出判别分析准确率，混淆矩阵等方面进行评价。我们将数据集按照 8:2 分为训练集和测试集，分别计算训练集和测试集的 APER 作为模型评价指标。

3.3.1 LDA 与 QDA

LDA 与 QDA 是最经典的判别分析方式，LDA 是一类线性判别器，而 QDA 则得到非线性结构解。

由于 LDA 模型需要正态性假设，而我们通过 EDA 知道变量 genergy,gpuls,gdenergy,gdpuls,energy,maxenergy 为连续变量，且经过对数变换后近似服从正态分布，故我们对这 6 个变量调用 MASS 包的 lda 函数进行线性判别分析，再通过 predict 函数利用已经训练的模型对测试集进行预测，

我们绘制出 LDA 模型下的混淆矩阵如下，可以观察到分类情况较好

train	true	
pred	0	1
0	491	24
1	1	1

对于本数据集，我们进行 cross-validation 验证，计算关于 LDA 和 QDA 的 APER 表格如下：

APER	训练集	测试集
LDA	0.048	0.074
QDA	0.088	0.100

可以观察到无论在训练集还是测试集，LDA 模型下 APER 的值都要明显小于 QDA 模型，说明用 LDA 模型可以更好地对该数据集进行预测，但理论

上 QDA 需要的假设更少，该结果与理论相反，具体原因有待进一步分析。

为将判别分析的结果直观的展示出来，以点的形状表示判别分析给出的类别，以点的颜色表示真实类型，得到前两个主成分得分的散点图如下：

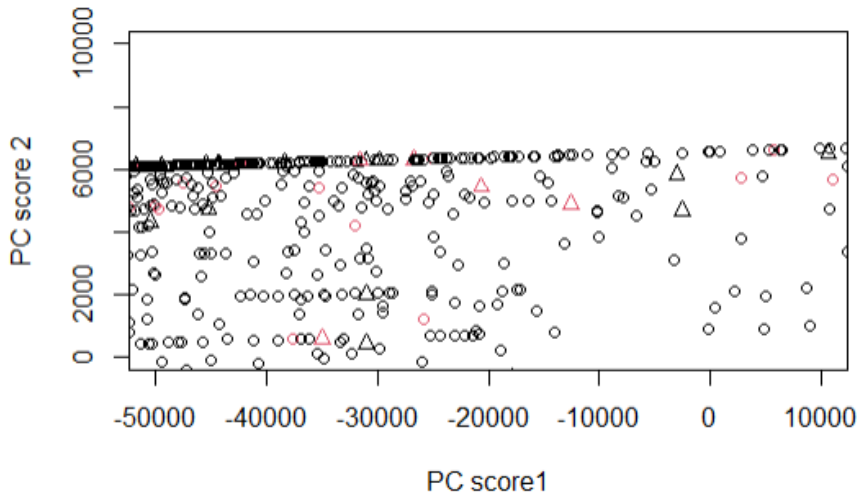


图 2: lda_plot

可以看到由于数据集中两类标签的分布并不相同，而是以 class=0 即不发生高能地震的样本为主，因此模型大部分都预测为 0。同时注意到由于两类样本分布并没有很明显的分界线，因此我们选择其他判别分析的方法来进行分类。

3.3.2 其他判别分析

根据以上分析，知道该模型使用 lda 并不能达到一个很好的效果，因此，我们尝试更为灵活的随机森林算法，并与 lda 进行比较。

随机森林 (Random Forest, RF) 算法是在决策树 (Decision Tree, DT) 的基础上，引入随机采样生成的多棵树组成的结果来进行分类，解决了单棵树分析力不足的问题。树的个数选取对于 RF 模型来说十分重要，本文使用交叉验证的方式，选取验证集表现最佳之处作为模型中树的个数。

SVM 则引入最大间隔的思想，使用 hinge 损失函数进行优化。kernel 的引入与构造则突破了线性框架，大大提升了 SVM 模型的灵活性与拟合能力。这里采用广为使用的高斯核函数带入模型，得到测试集和训练集的 APER 分别为 0.05996132 和 0.06724722

我们绘制两个模型的 AUC 曲线如下，可见 RF 的表现好于 SVM，SVM 预测的结果几乎等效于随机猜的结果，我认为原因可能为正负例比例不同导致模型训练拟合较差。

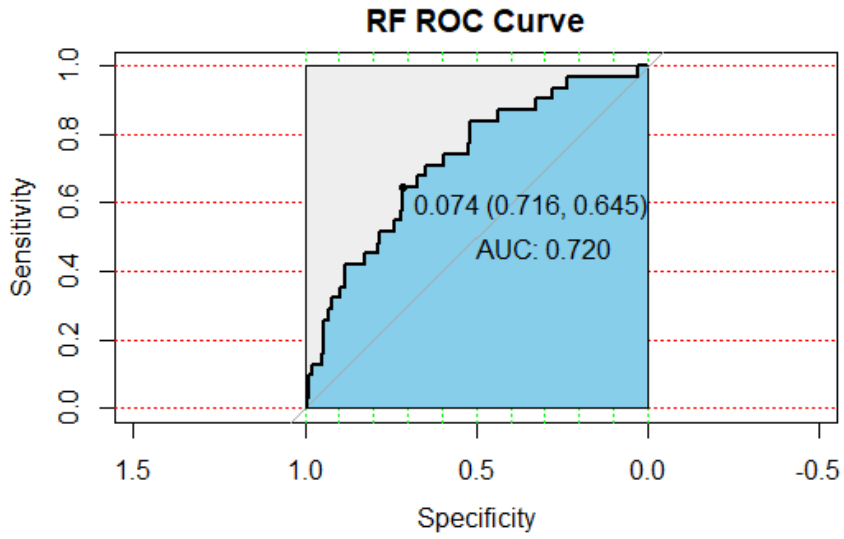


图 3: rf_roc

4 结论与改进

本研究通过主成分分析、判别分析探究了波兰某矿场的矿区地震因素特征，得到了两类主要成分，并利用已有数据集训练了 LDA、QDA 和额外的随机森林和 SVM 模型，对于是否发生强震这一标签进行了预测分析，效果能够接受。

本研究存在的不足在于样本数据集分布不均，并且相关变量个数过多，导致实验初期的预处理和数据变换工作较多。并且由于样本分布相关于是否发

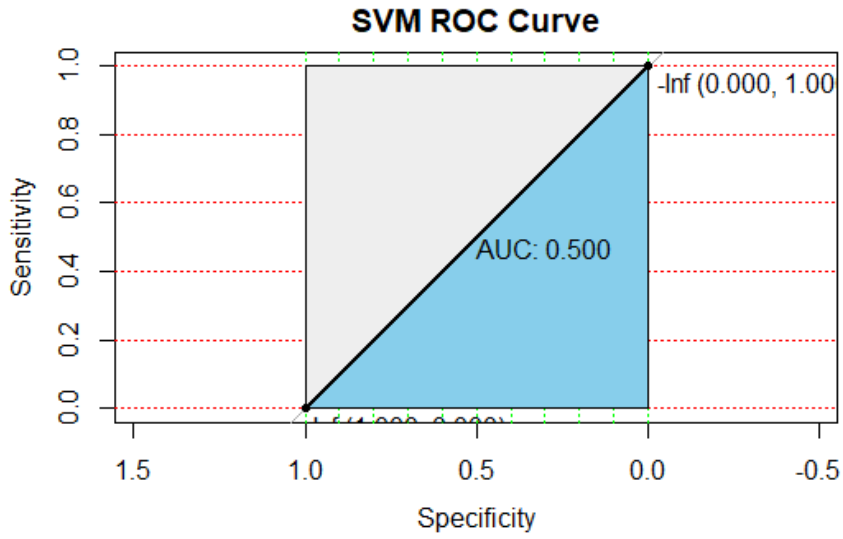


图 4: svm_roc

生强震来说没有没有明显区分，因此对于判别分析来说不是一个良好的数据集，因此如果能够得到更好的数据集，本实验得到的分类结果也许会更加精准。

5 引用

[1]MarekSikora, Induction and pruning of classification rules for prediction of microseismic hazards in coal mines. [2] 何其, 淮河流域夏季水质的聚类及判别分析.

6 附录

R 语言代码

```
#读取数据
library(readr)
```

```
library(CCA)
library(CCP)
library(MASS)
library(ggplot2)
library(randomForest)
library(e1071)
#library(gg3D)
library(pROC)
setwd("E:/Code/R/MSA_homework/project")
raw_data <- read_delim("seismic-bumps.csv", delim=",", col_names =F)
colnames(raw_data) <- c("seismic","seismoacoustic","shift","genergy","gpuls","gdenergy"
                        "gdpuls","ghazard","nbumps","nbumps2","nbumps3","nbumps4",
                        "nbumps5","nbumps6","nbumps7","nbumps89","energy","maxenergy",
                        "class")

new_data <- raw_data[,-c(10:16)]

# data preprocess
#seismic
v1 <- c(unique(new_data$seismic))
v2 <- c(1,2,3,4)
x <- data.frame(v1,v2)
len <- length(x$v1)
for(i in 1:len){
  new_data$seismic[which(new_data$seismic==x$v1[i])]=x$v2[i]
}
new_data$seismic <- as.numeric(new_data$seismic)
#seismoacoustic
v1 <- c(unique(new_data$seismoacoustic))
v2 <- c(1,2,3)
x <- data.frame(v1,v2)
len <- length(x$v1)
for(i in 1:len){
```

```
  new_data$seismoacoustic[which(new_data$seismoacoustic==x$v1[i])]=x$v2[i]
}
new_data$seismoacoustic <- as.numeric(new_data$seismoacoustic)
#shift
v1 <- c(unique(new_data$shift))
v2 <- c(0,1)
x <- data.frame(v1,v2)
len <- length(x$v1)
for(i in 1:len){
  new_data$shift[which(new_data$shift==x$v1[i])]=x$v2[i]
}
new_data$shift <- as.numeric(new_data$shift)
#ghazard
v1 <- c(unique(new_data$ghazard))
v2 <- c(1,2,3)
x <- data.frame(v1,v2)
len <- length(x$v1)
for(i in 1:len){
  new_data$ghazard[which(new_data$ghazard==x$v1[i])]=x$v2[i]
}
new_data$ghazard <- as.numeric(new_data$ghazard)

# 因子化
new_data$seismic <- as.factor(new_data$seismic)
new_data$seismoacoustic <- as.factor(new_data$seismoacoustic)
new_data$shift <- as.factor(new_data$shift)
new_data$ghazard <- as.factor(new_data$ghazard)

# EDA
summary(new_data)
dev.new()
par(mfrow=c(1,2))
#genergy
```

```
hist(new_data$genergy, xlab="Genergy", main="Histogram of Genergy", prob=T)
lines(density((new_data$genergy)),col="red")
new_genergy <- log(new_data$genergy)
hist(new_genergy, xlab="log Genergy", main="Histogram of log Genergy", prob=T)
lines(density((new_genergy)),col="red")
#gpuls
hist(new_data$gpuls, xlab="Gpuls", main="Histogram of Gpuls", prob=T)
lines(density((new_data$gpuls)),col="red")
new_gpuls <- log(new_data$gpuls)
hist(new_gpuls, xlab="log Gpuls", main="Histogram of log Gpuls", prob=T)
lines(density((new_gpuls)),col="red")
#gdenergy
hist(new_data$gdenergy, xlab="Gdenergy", main="Histogram of Gdenergy", prob=T)
lines(density((new_data$genergy)),col="red")
new_gdenergy <- log(abs(new_data$gdenergy))
hist(new_gdenergy, xlab="log Gdenergy", main="Histogram of log Gdenergy", prob=T)
lines(density((new_gdenergy),na.rm = TRUE),col="red")
#gdpuls
hist(new_data$gdpuls, xlab="Gdpuls", main="Histogram of Gdpuls", prob=T)
lines(density((new_data$gdpuls)),col="red")
new_gdpuls <- log(abs(new_data$gdpuls))
hist(new_gdpuls, xlab="log Gdpuls", main="Histogram of log Gdpuls", prob=T)
lines(density((new_gdpuls),na.rm = TRUE),col="red")

#PCA
pc <- prcomp(new_data[,c(4:7,10:11)])
sdev <- pc$sdev
plot(1:6,sdev,xlab="Index",ylab="Eigen values", type="b")
sum(sdev[1]/sum(sdev))
coef <- pc$rotation[,1:2]
round(coef,3)
pcscore <- pc$x[,1:2]
```

```
#DA
#分为训练集和测试集
final_data <- new_data
final_data$genergy <- new_genergy
final_data$gpuls <- new_gpuls
final_data$gdenergy <- new_gdenergy
final_data$gdpuls <- new_gdpuls
final_data$energy <- log(final_data$energy)
final_data$maxenergy <- log(final_data$maxenergy)
final_data$gdenergy[(final_data$gdenergy==-Inf)]=0
final_data$gdpuls[(final_data$gdpuls==-Inf)]=0
final_data$energy[(final_data$energy==-Inf)]=0
final_data$maxenergy[(final_data$maxenergy==-Inf)]=0

train_sub <- sample(nrow(final_data),4/5*nrow(final_data))
train_data <- final_data[train_sub, c(4:7,10:12)]
test_data <- final_data[-train_sub, c(4:7,10:12)]

#LDA
L <- lda(class~.,data=train_data)
pre_ran = predict(L,newdata=test_data)$class
tabcv = table(pred=pre_ran, true=test_data$class);tabcv
cverr = sum(tabcv[row(tabcv)!=col(tabcv)])/sum(tabcv); cverr
pre_train = predict(L, newdata=train_data)$class
tabcv = table(pred=pre_train, true=train_data$class);tabcv
cverr = sum(tabcv[row(tabcv)!=col(tabcv)])/sum(tabcv); cverr

# QDA
Q = qda(class~., data=train_data)
pre_ran <- predict(Q, newdata=test_data)$class
tabcv = table(pred=pre_ran, true=test_data$class);tabcv
cverr = sum(tabcv[row(tabcv)!=col(tabcv)])/sum(tabcv); cverr
pre_train = predict(Q, newdata=train_data)$class
```

```
tabcv = table(pred=pre_train, true=train_data$class);tabcv
cverr = sum(tabcv[row(tabcv)!=col(tabcv)])/sum(tabcv); cverr

#plot
plot(pcscore[,1],pcscore[,2],pch=as.numeric(pre_ran),col=as.factor(test_data$class),xla

# RF
testerr = c(); trainerr = c()
for (i in 1:10){
  rf = randomForest(class~., data=train_data, ntree = 25 * i)
  pre_ran_test = predict(rf, newdata=test_data)
  tabcv_test = table(pred=pre_ran_test, true=test_data$class)
  cverr_test = sum(tabcv_test[row(tabcv_test)!=col(tabcv_test)])/sum(tabcv_test)
  testerr = c(testerr, cverr_test)
  pre_ran_train = predict(rf, newdata=train_data)
  tabcv_train = table(pred=pre_ran_train, true=train_data$class)
  cverr_train =
  sum(tabcv_train[row(tabcv_train)!=col(tabcv_train)])/sum(tabcv_train)
  trainerr = c(trainerr, cverr_train)
}
testerr; trainerr

tree_num = c(25 * (1:10), 25 * (1:10))
err = c(trainerr, testerr)
type = c(rep('train', 10), rep('test', 10))
ggplot(data.frame(tree_num, err), aes(x=tree_num, y=err, color=type,
shape=type)) + geom_line() + geom_point(size=3)

rf = randomForest(class~., data=train_data, ntree = 175)
varImpPlot(rf, main = "variable importance")
pre_ran <- predict(rf, newdata=test_data)
tabcv = table(pred=pre_ran, true=test_data$class);tabcv
cverr = sum(tabcv[row(tabcv)!=col(tabcv)])/sum(tabcv); cverr
```

```
ran_roc <- roc(test_data$class, as.numeric(pre_ran))
plot(ran_roc, print.auc=TRUE, auc.polygon=TRUE, grid=c(0.1, 0.2),grid.col
=c("green", "red"), max.auc.polygon=TRUE, auc.polygon.col="skyblue",
print.thres=TRUE, main='RF ROC Curve')

# SVM
svm = svm(class~., data=train_data, type = 'C',kernel = 'radial')
pre_ran_test <- predict(svm, newdata=test_data)
tabcv_test = table(pred=pre_ran_test, true=test_data$class)
cverr_test =
sum(tabcv_test[row(tabcv_test)!=col(tabcv_test)])/sum(tabcv_test);
cverr_test
pre_ran_train <- predict(svm, newdata=train_data)
tabcv_train = table(pred=pre_ran_train, true=train_data$class)
cverr_train = sum(tabcv_train[row(tabcv_train)!=col(tabcv_train)])/sum(tabcv_train);cve
ran_roc <- roc(test_data$class, as.numeric(pre_ran_test))
plot(ran_roc, print.auc=TRUE, auc.polygon=TRUE, grid=c(0.1, 0.2), grid.col =c("green",
print.thres=TRUE, main='SVM ROC Curve')
```