Data\_mining\_Cup 2017

# 经济学院经济系 何友鑫 15320161152320

## 经分析该问题比较适合用支持向量机进行分类,用R语言作为实验软件，用到的包有 "e1071" "xlsx"

## 读文件，确定自变量和因变量

dataTrain <- read.table("./竞赛实验数据2017/kddtrain2017.txt")  
x <- dataTrain[,-101]  
y <- dataTrain[,101]

## 计算SVM在2种分类机，4种核函数下模型的错误次数

type=c("C-classification","nu-classification")  
kernel=c("linear","polynomial","radial","sigmoid")  
accuracy=matrix(0,2,4)  
for (i in 1:2)  
{  
 for ( j in 1:4)   
 {  
 model <- svm(x,y,type=type[i],kernel = kernel[j])  
 pred\_temp=predict(model,x)  
 accuracy[i,j]=sum(pred\_temp!=y)  
 }  
}  
dimnames(accuracy)=list(type,kernel)  
accuracy

## linear polynomial radial sigmoid  
## C-classification 1212 16 31 2971  
## nu-classification 1399 105 276 3229

print(paste0("所有模型中最高的正确率为",(6270-16)/6270))

## [1] "所有模型中最高的正确率为0.997448165869218"

## 由以上结果可知，使用SVM进行实验，type="C-classification",kernel = "polynomial"的模型最优。

## 实验1用训练数据的前5770条作为训练集，后500条作为测试集，看看预测结果

model1 <- svm(x[1:5770,],y[1:5770],type="C-classification",kernel = "polynomial")  
pred1 <- predict(model1,x[5771:6270,])

table(pred1,y[5771:6270])

##   
## pred1 0 1 2  
## 0 134 1 9  
## 1 2 142 2  
## 2 13 8 189

## 实验2用训练数据的前6000条作为训练集，后270条作为测试集，看看预测结果

model2 <- svm(x[1:6000,],y[1:6000],type="C-classification",kernel = "polynomial")  
pred2 <- predict(model2,x[6001:6270,])

table(pred2,y[6001:6270])

##   
## pred2 0 1 2  
## 0 73 0 3  
## 1 0 84 0  
## 2 7 5 98

## 实验3使用全部训练样本展示预测结果，并与真实情况的比较。

model\_fitted <- svm(x,y,type="C-classification",kernel = "polynomial")  
summary(model\_fitted)

##   
## Call:  
## svm.default(x = x, y = y, type = "C-classification", kernel = "polynomial")  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: polynomial   
## cost: 1   
## degree: 3   
## gamma: 0.01   
## coef.0: 0   
##   
## Number of Support Vectors: 5115  
##   
## ( 1443 2101 1571 )  
##   
##   
## Number of Classes: 3   
##   
## Levels:   
## 0 1 2

pred <- predict(model\_fitted,x)

table(pred,y)

## y  
## pred 0 1 2  
## 0 1963 1 0  
## 1 6 1839 1  
## 2 7 1 2452

## 由实验的的结果来看，模型是可信的。

## 由于自变量有100个，所以不好进行权重优化，99.74%的准确率是可以接受的范围，故即用原始模型作为最终模型

## 读测试数据，并用模型进行预测，将结果写入excel文件中

dataTest <- read.table("./竞赛实验数据2017/kddtest2017.txt")  
pred\_test=predict(model\_fitted,dataTest)  
write.xlsx(pred\_test,"predict\_result.xlsx",col.names = F,row.names = F)