Chapter 5: Classification Using Decision Trees and Rules

Xi Liang

5/23/2017

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Identifying risky bank loans using C5.0 decision trees

Data Exploration

```r
credit <- read.csv("data/credit.csv")
```

```r
str(credit)
```

```
## 'data.frame':    1000 obs. of  21 variables:
##  $ checking_balance    : Factor w/ 4 levels "< 0 DM","> 200 DM",..: 1 3 4 1 1 4 4 3 4 3 ...
##  $ months_loan_duration: int  6 48 12 42 24 36 24 36 12 30 ...
##  $ credit_history      : Factor w/ 5 levels "critical","delayed",..: 1 5 1 5 2 5 5 5 5 1 ...
##  $ purpose             : Factor w/ 10 levels "business","car (new)",..: 8 8 5 6 2 5 6 3 8 2 ...
##  $ amount              : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
##  $ savings_balance     : Factor w/ 5 levels "< 100 DM","> 1000 DM",..: 5 1 1 1 1 5 4 1 2 1 ...
##  $ employment_length   : Factor w/ 5 levels "> 7 yrs","0 - 1 yrs",..: 1 3 4 4 3 3 1 3 4 5 ...
##  $ installment_rate    : int  4 2 2 2 3 2 3 2 2 4 ...
##  $ personal_status     : Factor w/ 4 levels "divorced male",..: 4 2 4 4 4 4 4 4 1 3 ...
##  $ other_debtors       : Factor w/ 3 levels "co-applicant",..: 3 3 3 2 3 3 3 3 3 3 ...
##  $ residence_history   : int  4 2 3 4 4 4 4 2 4 2 ...
##  $ property            : Factor w/ 4 levels "building society savings",..: 3 3 3 1 4 4 1 2 3 2 ...
##  $ age                 : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ installment_plan    : Factor w/ 3 levels "bank","none",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ housing             : Factor w/ 3 levels "for free","own",..: 2 2 2 1 1 1 2 3 2 2 ...
##  $ existing_credits    : int  2 1 1 1 2 1 1 1 1 2 ...
##  $ default             : int  1 2 1 1 2 1 1 1 1 2 ...
##  $ dependents          : int  1 1 2 2 2 2 1 1 1 1 ...
##  $ telephone           : Factor w/ 2 levels "none","yes": 2 1 1 1 1 2 1 2 1 1 ...
##  $ foreign_worker      : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ job                 : Factor w/ 4 levels "mangement self-employed",..: 2 2 4 2 2 4 2 1 4 1 ...
```

From here we will take a look at features that I believe that are likely to predict a loan default

```r
table(credit$checking_balance)
```

```
##
##     < 0 DM   > 200 DM 1 - 200 DM    unknown
##        274         63        269        394
```

```
table(credit$savings_balance)
```

```
##
##      < 100 DM     > 1000 DM  101 - 500 DM 501 - 1000 DM       unknown
##           603            48           103            63           183
```

```
summary(credit$months_loan_duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    12.0    18.0    20.9    24.0    72.0
```

```
summary(credit$amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     250    1366    2320    3271    3972   18420
```

```
credit$default <- ifelse(credit$default == 1, 'no', 'yes')
credit$default <- factor(credit$default)
table(credit$default)
```

```
##
##  no yes
## 700 300
```

Data Preparation

Creating random trainning and test datasets

```
set.seed(123)
train_sample <- sample(1000, 900)

credit_train <- credit[train_sample, ]
credit_test <- credit[-train_sample, ]
```

Training a model on the data

```
library(C50)
credit_model <- C5.0(credit_train[-17], credit_train$default)
credit_model
```

```
##
## Call:
## C5.0.default(x = credit_train[-17], y = credit_train$default)
##
## Classification Tree
## Number of samples: 900
## Number of predictors: 20
##
## Tree size: 54
##
## Non-standard options: attempt to group attributes
```

```r
summary(credit_model)
```

```
##
## Call:
## C5.0.default(x = credit_train[-17], y = credit_train$default)
##
##
## C5.0 [Release 2.07 GPL Edition]      Thu May 25 06:51:47 2017
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 900 cases (21 attributes) from undefined.data
##
## Decision tree:
##
## checking_balance in {> 200 DM,unknown}: no (412/50)
## checking_balance in {< 0 DM,1 - 200 DM}:
## :...other_debtors = guarantor:
##     :...months_loan_duration > 36: yes (4/1)
##     :   months_loan_duration <= 36:
##     :   :...installment_plan in {none,stores}: no (24)
##     :       installment_plan = bank:
##     :       :...purpose = car (new): yes (3)
##     :           purpose in {business,car (used),domestic appliances,education,
##     :                       furniture,others,radio/tv,repairs,
##     :                       retraining}: no (7/1)
##     other_debtors in {co-applicant,none}:
##     :...credit_history = critical: no (102/30)
##         credit_history = fully repaid: yes (27/6)
##         credit_history = fully repaid this bank:
##         :...other_debtors = co-applicant: no (2)
##         :   other_debtors = none: yes (26/8)
##         credit_history in {delayed,repaid}:
##         :...savings_balance in {> 1000 DM,501 - 1000 DM}: no (19/3)
##             savings_balance = 101 - 500 DM:
##             :...other_debtors = co-applicant: yes (3)
##             :   other_debtors = none:
##             :   :...personal_status in {divorced male,
##             :       :                   married male}: yes (6/1)
##             :       personal_status = female:
##             :       :...installment_rate <= 3: no (4/1)
##             :       :   installment_rate > 3: yes (4)
##             :       personal_status = single male:
##             :       :...age <= 41: no (15/2)
##             :           age > 41: yes (2)
##             savings_balance = unknown:
##             :...credit_history = delayed: no (8)
##             :   credit_history = repaid:
##             :   :...foreign_worker = no: no (2)
##             :       foreign_worker = yes:
##             :       :...checking_balance = < 0 DM:
##             :           :...telephone = none: yes (11/2)
##             :           :   telephone = yes:
```

```
##             :           :   :...amount <= 5045: no (5/1)
##             :           :       amount > 5045: yes (2)
##             :           checking_balance = 1 - 200 DM:
##             :           :...residence_history > 3: no (9)
##             :               residence_history <= 3: [S1]
##             savings_balance = < 100 DM:
##             :...months_loan_duration > 39:
##                 :...residence_history <= 1: no (2)
##                 :   residence_history > 1: yes (19/1)
##                 months_loan_duration <= 39:
##                 :...purpose in {car (new),retraining}: yes (47/16)
##                     purpose in {domestic appliances,others}: no (3)
##                     purpose = car (used):
##                     :...amount <= 8086: no (9/1)
##                     :   amount > 8086: yes (5)
##                     purpose = education:
##                     :...checking_balance = < 0 DM: yes (5)
##                     :   checking_balance = 1 - 200 DM: no (2)
##                     purpose = repairs:
##                     :...residence_history <= 3: yes (4/1)
##                     :   residence_history > 3: no (3)
##                     purpose = business:
##                     :...credit_history = delayed: yes (2)
##                     :   credit_history = repaid:
##                     :   :...age <= 34: no (5)
##                     :       age > 34: yes (2)
##                     purpose = radio/tv:
##                     :...employment_length in {0 - 1 yrs,
##                     :   :                     unemployed}: yes (14/5)
##                     :   employment_length = 4 - 7 yrs: no (3)
##                     :   employment_length = > 7 yrs:
##                     :   :...amount <= 932: yes (2)
##                     :   :   amount > 932: no (7)
##                     :   employment_length = 1 - 4 yrs:
##                     :   :...months_loan_duration <= 15: no (6)
##                     :       months_loan_duration > 15:
##                     :       :...amount <= 3275: yes (7)
##                     :           amount > 3275: no (2)
##                     purpose = furniture:
##                     :...residence_history <= 1: no (8/1)
##                         residence_history > 1:
##                         :...installment_plan in {bank,stores}: no (3/1)
##                             installment_plan = none:
##                             :...telephone = yes: yes (7/1)
##                                 telephone = none:
##                                 :...months_loan_duration > 27: yes (3)
##                                     months_loan_duration <= 27: [S2]
##
## SubTree [S1]
##
## property in {building society savings,unknown/none}: yes (4)
## property = other: no (6)
## property = real estate:
## :...job = skilled employee: yes (2)
##     job in {mangement self-employed,unemployed non-resident,
##             unskilled resident}: no (2)
##
## SubTree [S2]
##
## checking_balance = 1 - 200 DM: yes (5/2)
## checking_balance = < 0 DM:
## :...property in {building society savings,real estate,unknown/none}: no (8)
##     property = other:
##     :...installment_rate <= 1: no (2)
##         installment_rate > 1: yes (4)
##
##
## Evaluation on training data (900 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      54  135(15.0%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     589    44    (a): class no
##      91   176    (b): class yes
##
##
##  Attribute usage:
##
##  100.00% checking_balance
##   54.22% other_debtors
##   50.00% credit_history
##   32.56% savings_balance
##   25.22% months_loan_duration
##   19.78% purpose
##   10.11% residence_history
##    7.33% installment_plan
##    5.22% telephone
##    4.78% foreign_worker
##    4.56% employment_length
##    4.33% amount
##    3.44% personal_status
##    3.11% property
##    2.67% age
##    1.56% installment_rate
##    0.44% job
##
##
## Time: 0.0 secs
```

Improving model performance

```
credit_pred <- predict(credit_model, credit_test)
```

```
library(gmodels)
CrossTable(credit_test$default, credit_pred,
           prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
           dnn = c('actual default', 'predicted default'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                | predicted default
## actual default |        no |       yes | Row Total |
## ---------------|-----------|-----------|-----------|
##             no |        60 |         7 |        67 |
##                |     0.600 |     0.070 |           |
## ---------------|-----------|-----------|-----------|
##            yes |        19 |        14 |        33 |
##                |     0.190 |     0.140 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |        79 |        21 |       100 |
## ---------------|-----------|-----------|-----------|
##
##
```

Out of 100 applicant records, the model correctly predicted 60 applicants did not default, and 14 did default, resulting 74% accuracy and error rate 26% (higher than training data).

Improving model performance

Boosting accuracy of decision trees

We could try increasing the accuracy of the model through the addition of adaptive boosting

```
credit_boost10 <- C5.0(credit_train[-17], credit_train$default,
                       trails = 10)

credit_boost10
```

```
##
## Call:
## C5.0.default(x = credit_train[-17], y = credit_train$default, trails = 10)
##
## Classification Tree
## Number of samples: 900
## Number of predictors: 20
##
## Tree size: 54
##
## Non-standard options: attempt to group attributes
```

```r
summary(credit_boost10)
```

```
##
## Call:
## C5.0.default(x = credit_train[-17], y = credit_train$default, trails = 10)
##
##
## C5.0 [Release 2.07 GPL Edition]      Thu May 25 06:51:47 2017
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 900 cases (21 attributes) from undefined.data
##
## Decision tree:
##
## checking_balance in {> 200 DM,unknown}: no (412/50)
## checking_balance in {< 0 DM,1 - 200 DM}:
## :...other_debtors = guarantor:
##     :...months_loan_duration > 36: yes (4/1)
##     :   months_loan_duration <= 36:
##     :   :...installment_plan in {none,stores}: no (24)
##     :       installment_plan = bank:
##     :       :...purpose = car (new): yes (3)
##     :           purpose in {business,car (used),domestic appliances,education,
##     :                       furniture,others,radio/tv,repairs,
##     :                       retraining}: no (7/1)
##     other_debtors in {co-applicant,none}:
##     :...credit_history = critical: no (102/30)
##         credit_history = fully repaid: yes (27/6)
##         credit_history = fully repaid this bank:
##         :...other_debtors = co-applicant: no (2)
##         :   other_debtors = none: yes (26/8)
##         credit_history in {delayed,repaid}:
##         :...savings_balance in {> 1000 DM,501 - 1000 DM}: no (19/3)
##             savings_balance = 101 - 500 DM:
##             :...other_debtors = co-applicant: yes (3)
##             :   other_debtors = none:
##             :   :...personal_status in {divorced male,
##             :       :                   married male}: yes (6/1)
##             :       personal_status = female:
##             :       :...installment_rate <= 3: no (4/1)
##             :       :   installment_rate > 3: yes (4)
##             :       personal_status = single male:
##             :       :...age <= 41: no (15/2)
##             :           age > 41: yes (2)
##             savings_balance = unknown:
##             :...credit_history = delayed: no (8)
##             :   credit_history = repaid:
```

```
##             :   :...foreign_worker = no: no (2)
##             :       foreign_worker = yes:
##             :       :...checking_balance = < 0 DM:
##             :           :...telephone = none: yes (11/2)
##             :           :   telephone = yes:
##             :           :   :...amount <= 5045: no (5/1)
##             :           :       amount > 5045: yes (2)
##             :           checking_balance = 1 - 200 DM:
##             :           :...residence_history > 3: no (9)
##             :               residence_history <= 3: [S1]
##             savings_balance = < 100 DM:
##             :...months_loan_duration > 39:
##                 :...residence_history <= 1: no (2)
##                 :   residence_history > 1: yes (19/1)
##                 months_loan_duration <= 39:
##                 :...purpose in {car (new),retraining}: yes (47/16)
##                     purpose in {domestic appliances,others}: no (3)
##                     purpose = car (used):
##                     :...amount <= 8086: no (9/1)
##                     :   amount > 8086: yes (5)
##                     purpose = education:
##                     :...checking_balance = < 0 DM: yes (5)
##                     :   checking_balance = 1 - 200 DM: no (2)
##                     purpose = repairs:
##                     :...residence_history <= 3: yes (4/1)
##                     :   residence_history > 3: no (3)
##                     purpose = business:
##                     :...credit_history = delayed: yes (2)
##                     :   credit_history = repaid:
##                     :   :...age <= 34: no (5)
##                     :       age > 34: yes (2)
##                     purpose = radio/tv:
##                     :...employment_length in {0 - 1 yrs,
##                     :   :                     unemployed}: yes (14/5)
##                     :   employment_length = 4 - 7 yrs: no (3)
##                     :   employment_length = > 7 yrs:
##                     :   :...amount <= 932: yes (2)
##                     :   :   amount > 932: no (7)
##                     :   employment_length = 1 - 4 yrs:
##                     :   :...months_loan_duration <= 15: no (6)
##                     :       months_loan_duration > 15:
##                     :       :...amount <= 3275: yes (7)
##                     :           amount > 3275: no (2)
##                     purpose = furniture:
##                     :...residence_history <= 1: no (8/1)
##                         residence_history > 1:
##                         :...installment_plan in {bank,stores}: no (3/1)
##                             installment_plan = none:
##                             :...telephone = yes: yes (7/1)
##                                 telephone = none:
##                                 :...months_loan_duration > 27: yes (3)
##                                     months_loan_duration <= 27: [S2]
##
## SubTree [S1]
```

```
##
## property in {building society savings,unknown/none}: yes (4)
## property = other: no (6)
## property = real estate:
## :...job = skilled employee: yes (2)
##     job in {mangement self-employed,unemployed non-resident,
##             unskilled resident}: no (2)
##
## SubTree [S2]
##
## checking_balance = 1 - 200 DM: yes (5/2)
## checking_balance = < 0 DM:
## :...property in {building society savings,real estate,unknown/none}: no (8)
##     property = other:
##     :...installment_rate <= 1: no (2)
##         installment_rate > 1: yes (4)
##
##
## Evaluation on training data (900 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      54  135(15.0%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     589    44    (a): class no
##      91   176    (b): class yes
##
##
##  Attribute usage:
##
##  100.00% checking_balance
##   54.22% other_debtors
##   50.00% credit_history
##   32.56% savings_balance
##   25.22% months_loan_duration
##   19.78% purpose
##   10.11% residence_history
##    7.33% installment_plan
##    5.22% telephone
##    4.78% foreign_worker
##    4.56% employment_length
##    4.33% amount
##    3.44% personal_status
##    3.11% property
##    2.67% age
##    1.56% installment_rate
##    0.44% job
##
##
```

```
## Time: 0.0 secs
```

```
credit_boost_pred10 <- predict(credit_boost10, credit_test)

CrossTable(credit_test$default, credit_boost_pred10,
           prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
           dnn = c('actual default', 'predicted default'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                | predicted default
## actual default |        no |       yes | Row Total |
## ---------------|-----------|-----------|-----------|
##             no |        60 |         7 |        67 |
##                |     0.600 |     0.070 |           |
## ---------------|-----------|-----------|-----------|
##            yes |        19 |        14 |        33 |
##                |     0.190 |     0.140 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |        79 |        21 |       100 |
## ---------------|-----------|-----------|-----------|
##
##
```

Based on what I observed from the result, adaptive boosting did not improve the performance of the prediction.

Making mistakes more costlier than others

```
matrix_dimensions <- list(c("no", "yes"), c("no", "yes"))
names(matrix_dimensions) <- c("predicted", "actual")
```

```
error_cost <- matrix(c(0, 1, 4, 0), nrow = 2,
                     dimnames = matrix_dimensions)
```

In this case, we assume that a loan default costs the bank four times as much as a missed opportunity.

```
error_cost
```

```
##          actual
## predicted no yes
##       no   0   4
##       yes  1   0
```

```
credit_cost <- C5.0(credit_train[-17], credit_train$default,
                    costs = error_cost)
```

```r
credit_cost_pred <- predict(credit_cost, credit_test)
CrossTable(credit_test$default, credit_cost_pred,
           prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
           dnn = c('actual default', 'predicted default'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  100
##
##
##                 | predicted default
## actual default |        no |       yes | Row Total |
## ---------------|-----------|-----------|-----------|
##             no |        33 |        34 |        67 |
##                |     0.330 |     0.340 |           |
## ---------------|-----------|-----------|-----------|
##            yes |         7 |        26 |        33 |
##                |     0.070 |     0.260 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |        40 |        60 |       100 |
## ---------------|-----------|-----------|-----------|
##
##
```

Comparing to the boosted model,this model had 41% of error rate, while the boosting model only had 26%. However, boosting model had 19% of false postives (predicted 19% of applicants did not default while they did), the cost model effectively reducued the false postives with the trade off of reduction in false negatives. This may be acceptable if our cost estimates were accurate.