An Introduction to Data Mining with R

Yanchang Zhao

http://www.RDataMining.com

6 September 2013
Questions

- Do you know data mining and techniques for it?
Questions

- Do you know data mining and techniques for it?
- Have you used R before?
Questions

- Do you know data mining and techniques for it?
- Have you used R before?
- Have you used R in your data mining research or projects?
Outline

Introduction

Classification with R

Clustering with R

Association Rule Mining with R

Text Mining with R

Time Series Analysis with R

Social Network Analysis with R

R and Hadoop

Online Resources
What is R?

- R \(^1\) is a free software environment for statistical computing and graphics.
- R can be easily extended with 4,728 packages available on CRAN\(^2\) (as of Sept 6, 2013).
- Many other packages provided on Bioconductor\(^3\), R-Forge\(^4\), GitHub\(^5\), etc.
- R manuals on CRAN\(^6\)
  - *An Introduction to R*
  - *The R Language Definition*
  - *R Data Import/Export*
  - …

---

\(^1\)http://www.r-project.org/
\(^2\)http://cran.r-project.org/
\(^3\)http://www.bioconductor.org/
\(^4\)http://r-forge.r-project.org/
\(^5\)https://github.com/
\(^6\)http://cran.r-project.org/manuals.html
Why R?

- R is widely used in both academia and industry.
- R is ranked no. 1 again in the KDnuggets 2013 poll on Top Languages for analytics, data mining, data science[^7].
- The CRAN Task Views[^8] provide collections of packages for different tasks.
  - Machine learning & statistical learning
  - Cluster analysis & finite mixture models
  - Time series analysis
  - Multivariate statistics
  - Analysis of spatial data
  - . . .

[^8]: [http://cran.r-project.org/web/views/](http://cran.r-project.org/web/views/)
Classification with R

- Decision trees: `rpart`, `party`
- Random forest: `randomForest`, `party`
- SVM: `e1071`, `kernlab`
- Neural networks: `nnet`, `neuralnet`, `RSNNS`
- Performance evaluation: `ROCR`
The Iris Dataset

# iris data
str(iris)

## 'data.frame': 150 obs. of 5 variables:
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ... 
## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ... 
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ... 
## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ... 
## $ Species : Factor w/ 3 levels "setosa","versicolor",...

# split into training and test datasets
set.seed(1234)
ind <- sample(2, nrow(iris), replace=T, prob=c(0.7, 0.3))
iris.train <- iris[ind==1, ]
iris.test <- iris[ind==2, ]
Build a Decision Tree

```r
# build a decision tree
library(party)
iris.formula <- Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
iris.ctree <- ctree(iris.formula, data=iris.train)
```
plot(iris.ctree)
Prediction

# predict on test data
pred <- predict(iris.ctree, newdata = iris.test)
# check prediction result
table(pred, iris.test$Species)

##
## pred setosa versicolor virginica
## setosa 10 0 0 0
## versicolor 0 12 2
## virginica 0 0 14
Outline

Introduction

Classification with R

Clustering with R

Association Rule Mining with R

Text Mining with R

Time Series Analysis with R

Social Network Analysis with R

R and Hadoop

Online Resources
Clustering with R

- **k-means**: `kmeans()`, `kmeansruns()\(^9\)
- **k-medoids**: `pam()`, `pamk()`
- **Hierarchical clustering**: `hclust()`, `agnes()`, `diana()`
- **DBSCAN**: `fpc`
- **BIRCH**: `birch`

\(^9\)Functions are followed with “()”, and others are packages.
**k-means Clustering**

```r
set.seed(8953)
iris2 <- iris
# remove class IDs
iris2$Species <- NULL
# k-means clustering
iris.kmeans <- kmeans(iris2, 3)
# check result
table(iris$Species, iris.kmeans$cluster)

##
## 1 2 3
## setosa 0 50 0
## versicolor 2 0 48
## virginica 36 0 14
```
# plot clusters and their centers
plot(iris2[, c("Sepal.Length", "Sepal.Width")], col=iris.kmeans$cluster)
points(iris.kmeans$centers[, c("Sepal.Length", "Sepal.Width")],
     col=1:3, pch="*", cex=5)
Density-based Clustering

```r
library(fpc)
iris2 <- iris[-5]  # remove class IDs
# DBSCAN clustering
ds <- dbscan(iris2, eps = 0.42, MinPts = 5)
# compare clusters with original class IDs
table(ds$cluster, iris$Species)
```

```text
##
## setosa  versicolor virginica
## 0   2    10    17
## 1  48     0     0
## 2   0   37     0
## 3   0     3   33
```
# 1-3: clusters; 0: outliers or noise

`plotcluster(iris2, ds$cluster)`
Outline

Introduction

Classification with R

Clustering with R

Association Rule Mining with R

Text Mining with R

Time Series Analysis with R

Social Network Analysis with R

R and Hadoop

Online Resources
Association Rule Mining with R

- Association rules: `apriori()`, `eclat()` in package `arules`
- Sequential patterns: `arulesSequence`
- Visualisation of associations: `arulesViz`
The Titanic Dataset

```r
load("./data/titanic.raw.rdata")
dim(titanic.raw)

## [1] 2201  4

idx <- sample(1:nrow(titanic.raw), 8)
titanic.raw[idx, ]
```

```
##   Class Sex  Age Survived
## 501  3rd Male Adult   No
## 477  3rd Male Adult   No
## 674  3rd Male Adult   No
## 766  Crew Male Adult  No
##1485  3rd Female Adult No
##1388  2nd Female Adult No
## 448  3rd Male Adult   No
## 590  3rd Male Adult   No
```
# find association rules with the APRIORI algorithm
library(arules)

rules <- apriori(titanic.raw, control=list(verbose=F),
                 parameter=list(minlen=2, supp=0.005, conf=0.8),
                 appearance=list(rhs=c("Survived=No", "Survived=Yes"),
                                  default="lhs"))

# sort rules
quality(rules) <- round(quality(rules), digits=3)

rules.sorted <- sort(rules, by="lift")

# have a look at rules
# inspect(rules.sorted)
<table>
<thead>
<tr>
<th>#</th>
<th>lhs</th>
<th>rhs</th>
<th>support</th>
<th>confidence</th>
<th>lift</th>
</tr>
</thead>
<tbody>
<tr>
<td># 1</td>
<td>{Class=2nd, Age=Child} =&gt; {Survived=Yes}</td>
<td>0.011</td>
<td>1.000</td>
<td>3.096</td>
<td></td>
</tr>
<tr>
<td># 2</td>
<td>{Class=2nd, Sex=Female, Age=Child} =&gt; {Survived=Yes}</td>
<td>0.006</td>
<td>1.000</td>
<td>3.096</td>
<td></td>
</tr>
<tr>
<td># 3</td>
<td>{Class=1st, Sex=Female} =&gt; {Survived=Yes}</td>
<td>0.064</td>
<td>0.972</td>
<td>3.010</td>
<td></td>
</tr>
<tr>
<td># 4</td>
<td>{Class=1st, Sex=Female, Age=Adult} =&gt; {Survived=Yes}</td>
<td>0.064</td>
<td>0.972</td>
<td>3.010</td>
<td></td>
</tr>
<tr>
<td># 5</td>
<td>{Class=2nd, Sex=Male, Age=Adult} =&gt; {Survived=No}</td>
<td>0.070</td>
<td>0.917</td>
<td>1.354</td>
<td></td>
</tr>
<tr>
<td># 6</td>
<td>{Class=2nd, Sex=Female} =&gt; {Survived=Yes}</td>
<td>0.042</td>
<td>0.877</td>
<td>2.716</td>
<td></td>
</tr>
<tr>
<td># 7</td>
<td>{Class=Crew, Sex=Female} =&gt; {Survived=Yes}</td>
<td>0.009</td>
<td>0.870</td>
<td>2.692</td>
<td></td>
</tr>
<tr>
<td># 8</td>
<td>{Class=Crew, Sex=Female, Age=Adult} =&gt; {Survived=Yes}</td>
<td>0.009</td>
<td>0.870</td>
<td>2.692</td>
<td></td>
</tr>
<tr>
<td># 9</td>
<td>{Class=2nd, Sex=Male} =&gt; {Survived=No}</td>
<td>0.070</td>
<td>0.860</td>
<td>1.271</td>
<td></td>
</tr>
<tr>
<td># 10</td>
<td>{Class=2nd, Age=Child} =&gt; {Survived=Yes}</td>
<td>0.011</td>
<td>1.000</td>
<td>3.096</td>
<td></td>
</tr>
</tbody>
</table>
library(arulesViz)
plot(rules, method = "graph")
Outline

Introduction

Classification with R

Clustering with R

Association Rule Mining with R

Text Mining with R

Time Series Analysis with R

Social Network Analysis with R

R and Hadoop

Online Resources
Text Mining with R

- Text mining: *tm*
- Topic modelling: *topicmodels, lda*
- Word cloud: *wordcloud*
- Twitter data access: *twitterR*
Fetch Twitter Data

## retrieve tweets from the user timeline of @rdatamining
library(twitteR)
# tweets <- userTimeline('rdatamining')
load(file = './data/rdmTweets.RData')
(nDocs <- length(tweets))

## [1] 320

strwrap(tweets[[320]]$text, width = 50)

## [1] "An R Reference Card for Data Mining is now"
## [2] "available on CRAN. It lists many useful R"
## [3] "functions and packages for data mining"

# convert tweets to a data frame
df <- do.call("rbind", lapply(tweets, as.data.frame))
library(tm)
# build a corpus
myCorpus <- Corpus(VectorSource(df$text))
# convert to lower case
myCorpus <- tm_map(myCorpus, tolower)
# remove punctuation & numbers
myCorpus <- tm_map(myCorpus, removePunctuation)
myCorpus <- tm_map(myCorpus, removeNumbers)
# remove URLs
removeURL <- function(x) gsub("http\[[[:alnum:]]\]*", "", x)
myCorpus <- tm_map(myCorpus, removeURL)
# remove 'r' and 'big' from stopwords
myStopwords <- setdiff(stopwords("english"), c("r", "big"))
# remove stopwords
myCorpus <- tm_map(myCorpus, removeWords, myStopwords)
```r
# keep a copy of corpus
myCorpusCopy <- myCorpus

# stem words
myCorpus <- tm_map(myCorpus, stemDocument)

# stem completion
myCorpus <- tm_map(myCorpus, stemCompletion,
                   dictionary = myCorpusCopy)

# replace "miners" with "mining", because "mining" was
# first stemmed to "mine" and then completed to "miners"
myCorpus <- tm_map(myCorpus, gsub, pattern="miners",
                   replacement="mining")

strwrap(myCorpus[320], width=50)

## [1] "r reference card data mining now available cran"
## [2] "list used r functions package data mining"
## [3] "applications"
```
Frequent Terms

```
myTdm <- TermDocumentMatrix(myCorpus,
control=list(wordLengths=c(1,Inf)))

# inspect frequent words
(freq.terms <- findFreqTerms(myTdm, lowfreq=20))

## [1] "analysis"  "big"      "computing"
## [4] "data"      "examples" "mining"
## [7] "network"   "package"  "position"
## [10] "postdoctoral" "r"       "research"
## [13] "slides"    "social"   "tutorial"
## [16] "university" "used"
```
Associations

```r
# which words are associated with 'r'?  
findAssocs(myTdm, "r", 0.2)

## examples code package
## 0.32 0.29 0.20

# which words are associated with 'mining'?  
findAssocs(myTdm, "mining", 0.25)

## data mahout recommendation sets
## 0.47 0.30 0.30 0.30
## supports frequent itemset
## 0.30 0.26 0.26
```
library(graph)
library(Rgraphviz)
plot(myTdm, term=freq.terms, corThreshold=0.1, weighting=T)
```r
library(wordcloud)
m <- as.matrix(myTdm)
freq <- sort(rowSums(m), decreasing=T)
wordcloud(words=names(freq), freq=freq, min.freq=4, random.order=F)
```
library(topicmodels)
set.seed(123)
myLda <- LDA(as.DocumentTermMatrix(myTdm), k=8)

## Topic 1          Topic 2          Topic 3          Topic 4
## [1,] "data"      "r"            "r"            "research"
## [2,] "mining"    "package"     "time"         "position"
## [3,] "big"       "examples"    "series"       "data"
## [4,] "association" "used"         "users"        "university"
## [5,] "rules"     "code"         "talk"         "postdoctoral"

## Topic 5          Topic 6          Topic 7          Topic 8
## [1,] "mining"    "group"        "data"         "analysis"
## [2,] "data"      "data"         "r"            "network"
## [3,] "slides"    "used"         "mining"       "social"
## [4,] "modelling" "software"     "analysis"     "text"
## [5,] "tools"     "kdnuggets"   "book"         "slides"
Outline

Introduction

Classification with R

Clustering with R

Association Rule Mining with R

Text Mining with R

Time Series Analysis with R

Social Network Analysis with R

R and Hadoop

Online Resources
Time Series Analysis with R

- Time series decomposition: `decomp()`, `decompose()`, `arima()`, `stl()`
- Time series forecasting: `forecast`
- Time Series Clustering: `TSclust`
- Dynamic Time Warping (DTW): `dtw`
Outline

Introduction
Classification with R
Clustering with R
Association Rule Mining with R
Text Mining with R
Time Series Analysis with R
Social Network Analysis with R
R and Hadoop
Online Resources
Social Network Analysis with R

- **Packages:** `igraph`, `sna`
- **Centrality measures:** `degree()`, `betweenness()`, `closeness()`, `transitivity()`
- **Clusters:** `clusters()`, `no.clusters()`
- **Cliques:** `cliques()`, `largest.cliques()`, `maximal.cliques()`, `clique.number()`
- **Community detection:** `fastgreedy.community()`, `spinglass.community()`
Outline

Introduction

Classification with R

Clustering with R

Association Rule Mining with R

Text Mining with R

Time Series Analysis with R

Social Network Analysis with R

R and Hadoop

Online Resources
R and Hadoop

- Packages: RHadoop, RHive
- RHadoop\(^\text{10}\) is a collection of 3 R packages:
  - `rmr2` - perform data analysis with R via MapReduce on a Hadoop cluster
  - `rhdfs` - connect to Hadoop Distributed File System (HDFS)
  - `rhbase` - connect to the NoSQL HBase database
- You can play with it on a single PC (in standalone or pseudo-distributed mode), and your code developed on that will be able to work on a cluster of PCs (in full-distributed mode)!
- Step by step to set up my first R Hadoop system
  http://www.rdatamining.com/tutorials/rhadoop

\(^{10}\)https://github.com/RevolutionAnalytics/RHadoop/wiki
library(rmr2)
map <- function(k, lines) {
  words.list <- strsplit(lines, "\\s")
  words <- unlist(words.list)
  return(keyval(words, 1))
}
reduce <- function(word, counts) {
  keyval(word, sum(counts))
}
wordcount <- function(input, output = NULL) {
  mapreduce(input = input, output = output, input.format = "text",
            map = map, reduce = reduce)
}
## Submit job
out <- wordcount(in.file.path, out.file.path)

11

---

11 From Jeffrey Breen’s presentation on *Using R with Hadoop* 
Outline

Introduction

Classification with R

Clustering with R

Association Rule Mining with R

Text Mining with R

Time Series Analysis with R

Social Network Analysis with R

R and Hadoop

Online Resources
Online Resources

- RDataMining website
  http://www.rdatamining.com
    - R Reference Card for Data Mining
    - R and Data Mining: Examples and Case Studies

- RDataMining Group on LinkedIn (3100+ members)
  http://group.rdatamining.com

- RDataMining on Twitter (1200+ followers)
  http://twitter.com/rdatamining

- Free online courses
  http://www.rdatamining.com/resources/courses

- Online documents
  http://www.rdatamining.com/resources/onlinedocs
The End

Thanks!

Email: yanchang(at)rdatamining.com