Datamining with R

Richard Skeggs

Wednesday 9 September, 2pm

www.BLGdataresearch.org

ESRC Business and Local Government Data Research Centre

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

UEA University of East Anglia

University of Essex

University of Kent

LSE

Housekeeping

- Using GoToWebinar software
- Expected length of Webinar: 1 hour (40 minutes of presentation, with 20 minutes for answering questions)
- Got any questions?
- Please type them into the ‘Questions’ box in the panel on your right and questions will be answered at the end (all audience is muted)

Introduction

Will cover pre-processing, anomaly detection, association rule learning, clustering, classification, regression and summarisation with R.

Data Mining

Data mining is the process to discover interesting knowledge from large amounts of data.

Han and Kamber, 2000

Pre Processing

```
# Read CSV into R
myData <- read.csv(file="c:/ReadIn.csv")

# Create a dataframe.  Either from a query
myData <- sqlQuery(channel, "SELECT …")

# or complete table
myData <- sqlFetch(channel, “tablename”)

# Read data from zip files
myData <- read.table('myFile.gz')
```

Pre Processing

```
# Split a string on a character
x <- unlist(strsplit(x, "\n"))

# Replace one character sequence with another
x <- gsub("\t","",x)

# Advanced Perl parsing is available.
x <- sub("^[[:space:]]*(.*?)[[:space:]]*$", "\\1", x,
perl=TRUE)
```

PreProcessing

```
#Cooerse one type into another
class(c("5", "6.7", "8", ""))
[1] "character"

as.numeric(c("5", "6.7", "8", ""))
[1] 5.0 6.7 8.0 NA
```

PreProcessing

```
# Coearse string to dates
mydates <- as.Date(c("2007-06-22", "2004-02-13"))

# change the date format
today<-Sys.Date()
today
[1] "2015-08-27"

format(today, format="%d/%m/%Y")
[1] "27/08/2015"
```

PreProcessing

Symbol	Meaning	Example
%d	day as a number (0-31)	01-31
%a **%A**	abbreviated weekday unabbreviated weekday	Mon Monday
%m	month (00-12)	00-12
%b **%B**	abbreviated month unabbreviated month	Jan January
%y **%Y**	2-digit year 4-digit year	15 2012

PreProcessing

Ignore missing values

- na.fail(object) returns the object if no missing values.
- na.omit(object) removes cases.

Replace missing values

- newiris$Species <- NULL
- is.na(x) <- value

Anomaly Detection

- 10 Packages in CRAN handling outliers.
  - Tsoutliers
  - outliers
- https://github.com/twitter/AnomalyDetection

Outliers

Tsoutliers uses time series.

- Detection of outliers in time series following the Chen and Liu (1993) procedure.
- Innovative outliers, additive outliers, level shifts, temporary changes and seasonal level shifts are considered.

Outliers

```
# load tsoutliers library
library("tsoutliers")
# load the Harmonised Indices of Consumer Prices data set
data("hicp")
y <- hicp[["011600"]]
y
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
1995 4.312409 4.320949 4.322011 4.331523 4.347306 4.370966
4.372986 4.345233 4.341074 4.346270 4.341074 4.346270
1996 4.352469 4.370966 4.381026 4.387014 4.432957 4.450853
4.439588 4.383026 4.364753 4.352469 4.356581 4.354527
```

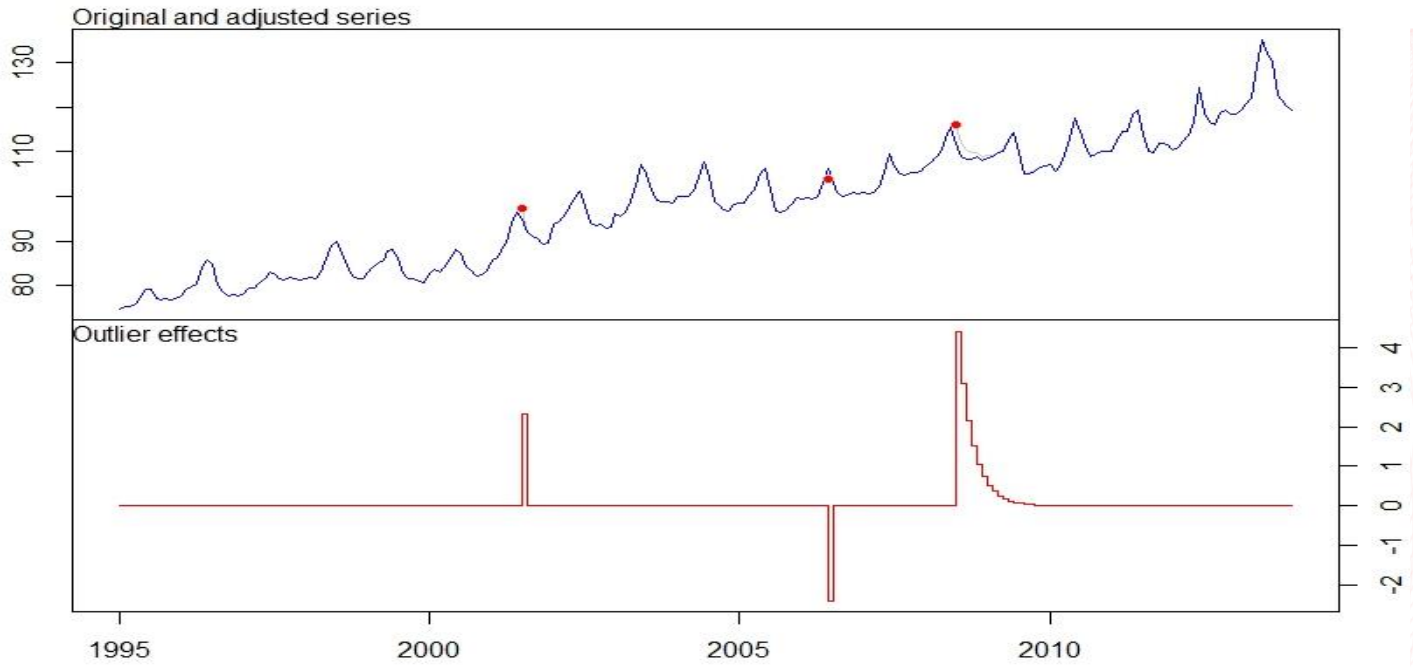
Outliers

```
# AO - additive outliers
# LS - level shift
# TC - temporary changes
out <- tsoutliers::tso(y,types = c("AO","LS","TC"),maxit.iloop=10)
out

Outliers:
  type   ind    time      coefhat    tstat
1 AO      79   2001:07     2.327     3.605
2 AO     138   2006:06    -2.400    -3.711
3 TC     163   2008:07     4.419     4.845

plot(out)
```

Outliers


Original and adjusted series
130
110
90
80
Outlier effects
4
3
2
1
0
-1
-2
1995
2000
2005
2010

Association Rule Learning

- Used to find association between variables.
- Often used as part of a recommendation engine.
- Arules & arulesViz packages available from CRAN.

Association Rule Learning

- data set contains the questionnaire data of the “Adult” database (originally called the “Census Income” Database)
- The Apriori algorithm employs level-wise search for frequent itemsets.

Association Rule Learning

```
> data("Adult")
> rules <- apriori(Adult, parameter = list(supp = 0.5, conf = 0.9, target =
"rules"))

Parameter specification:
confidence minval smax arem aval originalSupport support minlen maxlen target ext
0.9         0.1    1   none FALSE    TRUE           0.5     1      10 rules FALSE

Algorithmic control:
filter  tree  heap  memopt  load  sort  verbose
  0.1   TRUE  TRUE   FALSE  TRUE   2     TRUE
apriori - find association rules with the apriori algorithm version 4.21
(2004.05.09) (c) 1996-2004 Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[115 item(s), 48842 transaction(s)] done [0.02s].
sorting and recoding items ... [9 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [52 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].
```

Association Rule Learning

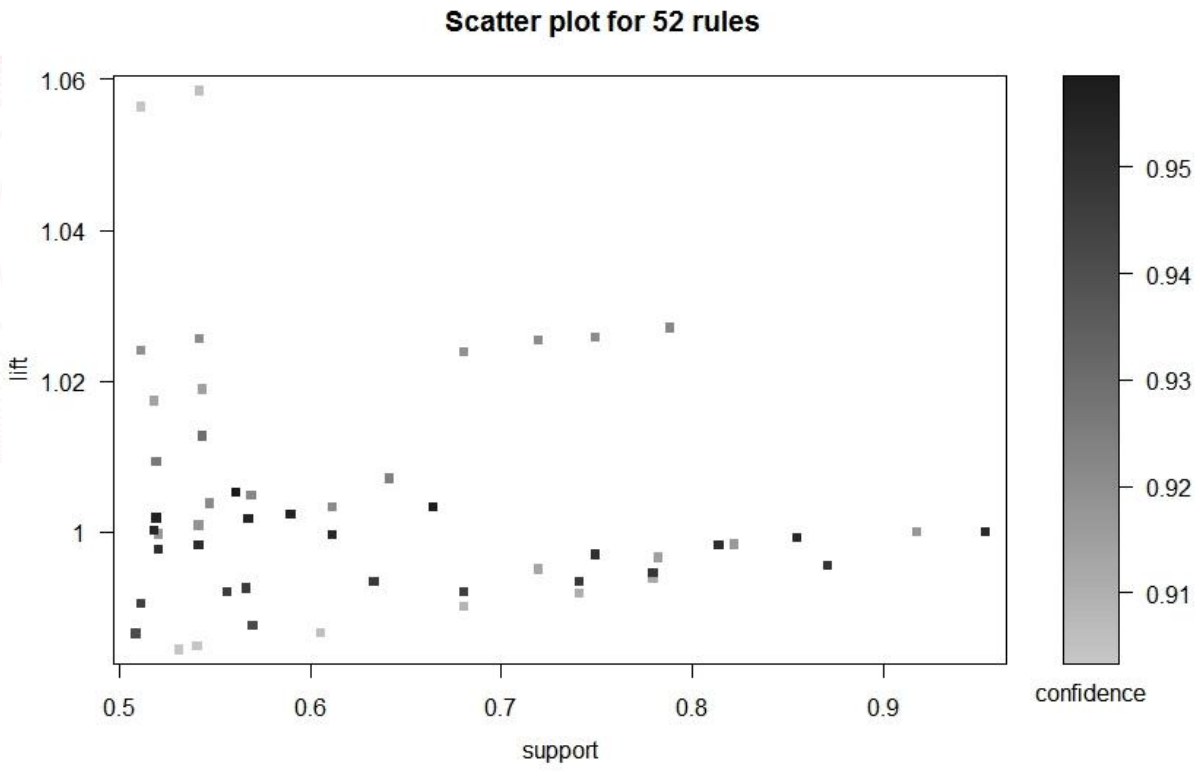
```
summary(rules)
set of 52 rules
rule length distribution (lhs + rhs):sizes
1  2  3  4
2 13 24 13

Min.  1st Qu. Median  Mean  3rd Qu. Max.
1.000 2.000   3.000   2.923 3.250   4.000
summary of quality measures:
   support         confidence          lift
Min.   :0.5084   Min.   :0.9031   Min.   :0.9844
1st Qu.:0.5415   1st Qu.:0.9155   1st Qu.:0.9937
Median :0.5974   Median :0.9229   Median :0.9997
Mean   :0.6436   Mean   :0.9308   Mean   :1.0036
3rd Qu.:0.7426   3rd Qu.:0.9494   3rd Qu.:1.0057
Max.   :0.9533   Max.   :0.9583   Max.   :1.0586

mining info:
data ntransactions support confidence
Adult         48842  0.5          0.9
```

Association Rule Learning

```
plot(rules, measure=c("support","lift"), shading="confidence")
```



Clustering

- K-Means
- Hybrid Hierarchical Clustering
- Expectation Maximization (EM)
- Dissimilarity Matrix Calculation
- Hierarchical Clustering
- Bayesian Hierarchical Clustering
- Density-Based Clustering
- K-Cores
- Fuzzy Clustering - Fuzzy C-means
- RockCluster
- Biclust
- Partitioning Around Medoids (PAM)
- CLUES
- Self-Organizing Maps (SOM)
- Proximus
- CLARA

Clustering

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	.2	setosa
4.9	3.0	1.4	.2	setosa
4.7	3.2	1.3	.2	setosa
4.6	3.1	1.5	.2	setosa
5.0	3.6	1.4	.2	setosa
5.4	3.9	1.7	.4	setosa
4.6	3.4	1.4	.3	setosa

Clustering

```
newiris <- iris
newiris$Species <- NULL
results <- kmeans(newiris, 3)

results K-means clustering with 3 clusters of sizes 62, 50, 38
Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1 5.901613     2.748387     4.393548     1.433871
2 5.006000     3.428000     1.462000     0.246000
3 6.850000     3.073684     5.742105     2.071053
Clustering vector:
[1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[49] 2 2 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[97] 1 1 1 1 3 1 3 3 3 3 1 3 3 3 3 3 3 1 1 3 3 3 3 1 3 1 3 1 3 3 1 1 3 3 3 3 3 1 3 3 3 3 1 3 3 3 1 3
[145] 3 3 1 3 3 1

Within cluster sum of squares by cluster:
[1] 39.82097 15.15100 23.87947
(between_SS / total_SS = 88.4 %)

Available components:
[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss"
[7] "size" "iter" "ifault"
```

Classification

- Support Vector Machines
- penalizedSVM
- K-Nearest Neighbours
- Outliers
- Decision Trees
- Naïve Bayes
- adaboost
- JRip

Classificiation

```
library(class)
library(e1071)

classifier<-naiveBayes(iris[,1:4], iris[,5])
table(predict(classifier, iris[,1:4]), iris[,5])
             setosa versicolor virginica
setosa           50          0         0
versicolor        0         47         3
virginica         0          3        47
```

Predict

Predict - generic function for predictions from the results of various model fitting functions.

- Arima
- glm
- HoltWinters
- Loess
- naive Bayes
- SVM