



北京邮电大学

BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS



相随相伴、谓之关联





艾新波 / 2018 • 北京






课程体系

R语言数据分析

上部：论道

-  第1章 气象万千、数以等观
-  第2章 所谓学习、归类而已
-  第3章 格言联璧话学习
-  第4章 源于数学、归于工程

中部：执具

-  第5章 工欲善其事必先利其器
-  第6章 基础编程
-  第7章 数据对象

 第8章 人人都爱tidyverse

 第9章 最美不过数据框

下部 博术

 第10章 观数以形

 **第11章 相随相伴、谓之关联**

 第12章 既是世间法、自当有分别

 第13章 方以类聚、物以群分

 第14章 庐山烟雨浙江潮

Task Views

*始终不要忘了, **Task Views**提供了最基本的导航*

Task Views → Machine Learning → Association Rules

Association Rules : Package arules provides both data structures for efficient handling of sparse binary data as well as interfaces to implementations of **Apriori** and **Eclat** for mining frequent itemsets, maximal frequent itemsets, closed frequent itemsets and association rules.

加载数据

```
library(tidyverse)
```

```
library(readr)
```

```
cjb_url <-
```

```
"https://github.com/byaxb/RDataAnalytics/raw/master/data/cjb.csv"
```

```
cjb <- read_csv(cjb_url,
```

```
               locale = locale(encoding = "CP936"))
```

数据预处理

arules包只能对离散数据进行关联规则挖掘，离散化有专用的包discretization

对于大部分的任务而言，cut()函数已经够用了

#定义一个百分制转成五分制成绩的函数

```
as_five_grade_scores <- function(x) {  
  cut(x,  
    breaks = c(0, seq(60, 100, 10)),  
    include.lowest = TRUE, right = FALSE,  
    ordered_result = TRUE,  
    labels = c("不及格", "及格", "中", "良", "优"))  
}
```

数据预处理

#对数据进行预处理

```
cjb %<>%
```

```
mutate_at(vars(xb, wlfk), factor) %>% #类型转换
```

```
mutate_at(vars(yw:sw), as_five_grade_scores) %>% #数据分箱
```

```
select(-(1:2)) #前两列姓名、班级两列不参与规则挖掘
```

```
library(arules)
```

#转换为transaction

```
cjb_trans <- as(cjb, "transactions")
```

数据预处理

cjb_trans

```
#> transactions in sparse format with
```

```
#> 775 transactions (rows) and
```

```
#> 49 items (columns)
```

```
inspect(head(cjb_trans))
```

```
#> items
```

transactionID

```
#> [1] {xb=女,yw=优,sx=良,wy=优,zz=优,ls=优,d1=优,w1=优,hx=优,sw=良,wlfk=文科} 1
```

```
#> [2] {xb=男,yw=良,sx=优,wy=良,zz=优,ls=优,d1=优,w1=良,hx=良,sw=良,wlfk=文科} 2
```

```
#> [3] {xb=男,yw=优,sx=中,wy=良,zz=优,ls=优,d1=优,w1=良,hx=优,sw=良,wlfk=文科} 3
```

```
#> [4] {xb=女,yw=优,sx=良,wy=优,zz=优,ls=优,d1=优,w1=良,hx=良,sw=良,wlfk=文科} 4
```

```
#> [5] {xb=男,yw=良,sx=优,wy=良,zz=优,ls=良,d1=良,w1=优,hx=优,sw=优,wlfk=文科} 5
```

```
#> [6] {xb=女,yw=优,sx=良,wy=良,zz=优,ls=良,d1=优,w1=良,hx=优,sw=良,wlfk=文科} 6
```

数据预处理

```
cjb_trans %>%
```

```
  as("list")
```

```
#> $`1`
```

```
#> [1] "xb=女"      "yw=优"      "sx=良"      "wy=优"
```

```
#> [5] "zz=优"      "ls=优"      "dl=优"      "wl=优"
```

```
#> [9] "hx=优"      "sw=良"      "wlfk=文科"
```

```
#>
```

```
#> $`2`
```

```
#> [1] "xb=男"      "yw=良"      "sx=优"      "wy=良"
```

```
#> [5] "zz=优"      "ls=优"      "dl=优"      "wl=良"
```

```
#> [9] "hx=良"      "sw=良"      "wlfk=文科"
```


规则挖掘

#算法实现，只是一句话的事儿

```
library(arulesViz)
```

```
irules_args_default <- apriori(cjb_trans)
```

```
#> Apriori
```

```
#>
```

```
#> Parameter specification:
```

```
#> confidence minval smax arem aval originalSupport
```

```
#> 0.8 0.1 1 none FALSE TRUE
```

```
#> maxtime support minlen maxlen target ext
```

```
#> 5 0.1 1 10 rules FALSE
```

规则挖掘

```
irules_args_default
```

```
#> set of 2097 rules
```

```
inspect(head(irules_args_default))
```

#>	lhs	rhs	support	confidence	lift	count
#> [1]	{w1=优}	=> {sx=优}	0.21	0.83	1.8	166
#> [2]	{w1=优}	=> {w1fk=理科}	0.21	0.81	1.6	162
#> [3]	{w1=优}	=> {hx=优}	0.24	0.94	1.5	188
#> [4]	{w1=优}	=> {d1=优}	0.24	0.92	1.3	185
#> [5]	{w1=优}	=> {zz=优}	0.21	0.81	1.1	162
#> [6]	{yw=优}	=> {d1=优}	0.24	0.89	1.2	183

规则挖掘

#设置支持度、置信度、最小长度等参数

```
irules <- apriori(  
  cjb_trans,  
  parameter = list(  
    minlen = 2,  
    supp = 50 / length(cjb_trans), #最小支持度, 减少偶然性  
    conf = 0.8 #最小置信度, 推断能力  
  ))  
  
inspectDT(irules)
```

规则挖掘

Viewer Zoom

Show 10 entries

Search:

	LHS	RHS	support	confidence	lift	count
	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>
[1]	{hx=中}	{wlfk=文科}	0.065	0.862	1.696	50.000
[2]	{wl=优}	{sx=优}	0.214	0.826	1.808	166.000
[3]	{wl=优}	{wlfk=理科}	0.209	0.806	1.639	162.000
[4]	{wl=优}	{hx=优}	0.243	0.935	1.510	188.000
[5]	{wl=优}	{dl=优}	0.239	0.920	1.267	185.000
[6]	{wl=优}	{zz=优}	0.209	0.806	1.096	162.000
[7]	{yw=优}	{dl=优}	0.236	0.888	1.223	183.000
[8]	{yw=优}	{zz=优}	0.244	0.917	1.247	189.000
[9]	{sw=优}	{hx=优}	0.334	0.918	1.483	259.000
[10]	{sw=优}	{dl=优}	0.326	0.897	1.235	253.000

Showing 1 to 10 of 5,584 entries

Previous 1 2 3 4 5 ... 559 Next

规则挖掘

#进一步设定前项和后项

```
irules <- apriori(  
  cjb_trans,  
  parameter = list(  
    minlen = 2,  
    supp = 50 / length(cjb_trans),  
    conf = 0.8  
  ),  
  appearance = list(rhs = paste0("wlfk=", c("文科", "理科")),  
                    default = "lhs"))
```

规则挖掘

Viewer Zoom

— □ ×

Show 10 entries

Search:

	LHS	RHS	support	confidence	lift	count
	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>
[16]	{ls=优,hx=良}	{wlfk=文科}	0.083	0.800	1.574	64.000
[74]	{xb=女,ls=优,sw=良}	{wlfk=文科}	0.083	0.800	1.574	64.000
[132]	{xb=女,sx=优,wl=优,hx=优}	{wlfk=理科}	0.072	0.800	1.627	56.000
[151]	{zz=优,dl=优,wl=优,hx=优}	{wlfk=理科}	0.155	0.800	1.627	120.000
[152]	{yw=良,zz=优,dl=优,wl=优}	{wlfk=理科}	0.088	0.800	1.627	68.000
[172]	{xb=男,zz=优,hx=优,sw=优}	{wlfk=理科}	0.124	0.800	1.627	96.000
[289]	{yw=良,wy=优,zz=优,dl=优,sw=优}	{wlfk=理科}	0.072	0.800	1.627	56.000
[299]	{vw=良,sx=优,ls=优,dl=	{wlfk=理科}	0.072	0.800	1.627	56.000

Showing 1 to 10 of 401 entries

Previous 1 2 3 4 5 ... 41 Next

规则挖掘

#对规则进行排序

```
irules_sorted <- sort(irules, by = "lift")
```

```
inspectDT(irules_sorted)
```

Viewer Zoom

Show 10 entries

Search:

	LHS	RHS	support	confidence	lift	count
	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>
[1]	{xb=男,sx=优,ls=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.074	0.934	1.901	57.000
[2]	{xb=男,sx=优,dl=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.090	0.933	1.899	70.000
[3]	{xb=男,sx=优,ls=优,dl=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.072	0.933	1.899	56.000
[4]	{xb=男,dl=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.099	0.928	1.887	77.000
[5]	{xb=男,sx=优,zz=优,ls=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.065	0.926	1.883	50.000
[6]	{xb=男,sx=优,zz=优,ls=优,dl=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.065	0.926	1.883	50.000

删除冗余规则

	LHS	RHS	support	confidence	lift
	All	All	All	All	All
[1]	{hx=中}	{wlfk=文科}	0.065	0.862	1.696
[2]	{wl=优}	{wlfk=理科}	0.209	0.806	1.639
[3]	{zz=优,wl=及格}	{wlfk=文科}	0.068	0.828	1.629
[4]	{xb=女,sx=中}	{wlfk=文科}	0.072	0.862	1.695
[5]	{sx=中,zz=优}	{wlfk=文科}	0.079	0.813	1.600
[6]	{xb=女,sw=中}	{wlfk=文科}	0.090	0.909	1.788
[7]	{zz=优,sw=中}	{wlfk=文科}	0.095	0.851	1.673
[8]	{sx=优,zz=良}	{wlfk=理科}	0.075	0.906	1.843
[9]	{wl=优,sw=优}	{wlfk=理科}	0.168	0.850	1.728
[10]	{sx=优,wl=优}	{wlfk=理科}	0.183	0.855	1.740



删除冗余规则

```
subset.matrix <-
```

```
  is.subset(irules_sorted, irules_sorted, sparse = FALSE)
```

```
subset.matrix[lower.tri(subset.matrix, diag = TRUE)] <- NA
```

```
redundant <- colSums(subset.matrix, na.rm = TRUE) >= 1
```

```
as.integer(which(redundant))
```

```
#> [1] 3 5 6 7 8 13 14 15 19 22 23 25
```

```
#> [13] 26 27 29 30 31 34 38 39 40 41 44 45
```

```
.....
```

```
#> [277] 391 392 393 394
```

删除冗余规则

```
(irules_pruned <- irules_sorted[!redundant])
```

```
#> set of 121 rules
```

```
inspectDT(irules_pruned)
```

[1]	{xb=男,sx=优,ls=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.074	0.934	1.901	57.000
[2]	{xb=男,sx=优,dl=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.090	0.933	1.899	70.000
[3]	{xb=男,dl=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.099	0.928	1.887	77.000
[4]	{xb=男,sx=优,wl=优,hx=优,sw=优}	{wlfk=理科}	0.092	0.922	1.876	71.000
[5]	{xb=男,sx=优,dl=优,wl=优,sw=优}	{wlfk=理科}	0.090	0.921	1.874	70.000
[6]	{xb=男,sx=优,dl=优,wl=优,hx=优}	{wlfk=理科}	0.105	0.920	1.872	81.000
[7]	{xb=男,sx=优,ls=优,wl=优,sw=优}	{wlfk=理科}	0.074	0.919	1.870	57.000
[8]	{xb=女,zz=优,sw=中}	{wlfk=文科}	0.070	0.947	1.863	54.000
[9]	{xb=男,sx=优,wy=优,wl=优,hx=优}	{wlfk=理科}	0.068	0.914	1.859	53.000
[10]	{xb=男,yw=良,sx=优,dl=优,wl=优}	{wlfk=理科}	0.068	0.914	1.859	53.000

规则规则可视化

```
library(arulesViz)
```

#最常用的一种方式

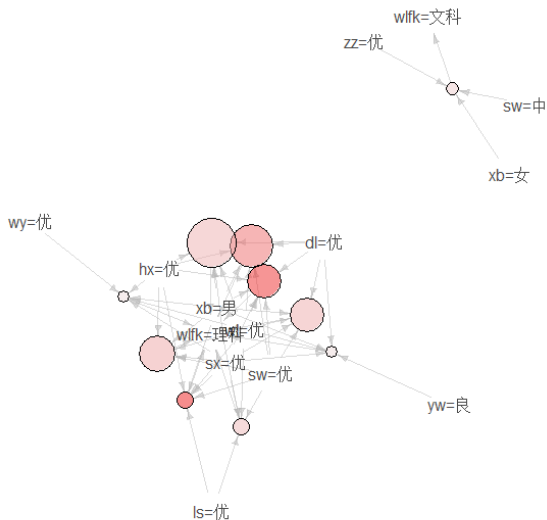
```
plot(irules_pruned[1:10],  
     method = "graph")
```

#交互式的规则可视化

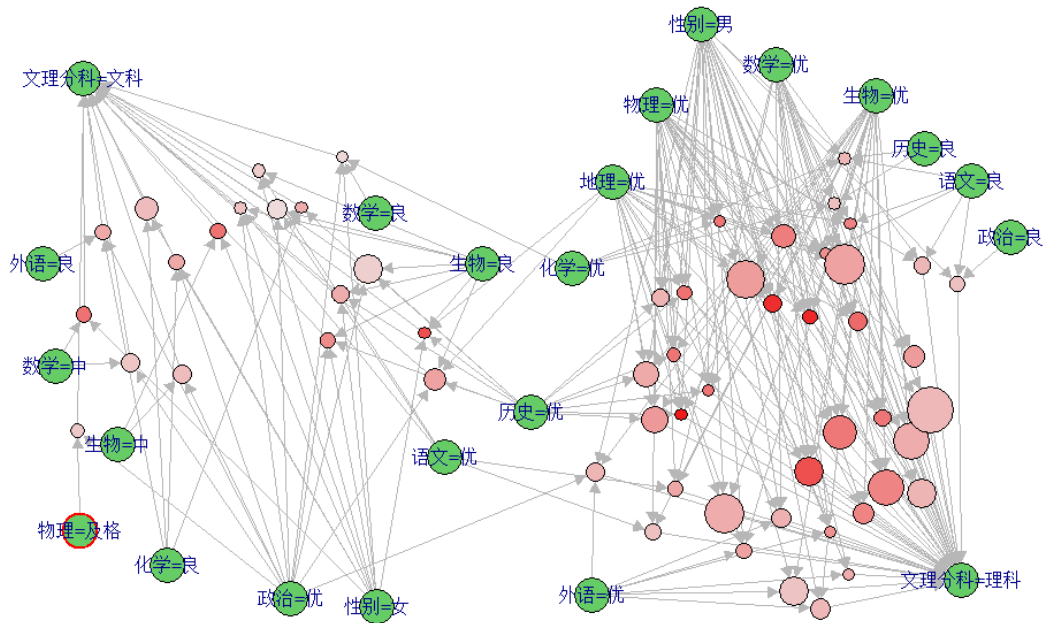
```
library(tcltk2)  
plot(irules_pruned,  
     method = "graph",  
     interactive = TRUE)
```

Graph for 10 rules

size: support (0.068 - 0.105)
color: lift (1.859 - 1.901)



规则规则可视化



规则评估

#查看评估指标

`quality(irules_pruned)`

#>	support	confidence	lift	count
#> 331	0.074	0.93	1.9	57
#> 334	0.090	0.93	1.9	70
#> 229	0.099	0.93	1.9	77
#> 212	0.092	0.92	1.9	71
#> 213	0.090	0.92	1.9	70

规则评估

$$\frac{\text{confidence}(X \Rightarrow Y) + \text{confidence}(! X \Rightarrow ! Y)}{2}$$

#更多评估指标

```
(more_measures <- interestMeasure(  
  irules_pruned,  
  measure = c("support", "confidence", "lift", "casualConfidence"),  
  transactions = cjb_trans))
```

#	support	confidence	lift	casualConfidence
# 1	0.073548	0.93443	1.9007	0.99990
# 2	0.090323	0.93333	1.8985	0.99990
# 3	0.099355	0.92771	1.8871	0.99989
# 4	0.091613	0.92208	1.8756	0.99988

规则提取

```
irules_sub <- subset(irules_pruned,  
                      lhs %pin% c("sw") &  
                      lift > 1.8)  
  
inspect(irules_sub)
```

#>	lhs	rhs	support	confidence
#> [1]	{xb=男, sx=优, ls=优, wl=优, hx=优, sw=优}	=> {wlfk=理科}	0.074	0.93
#> [2]	{xb=男, sx=优, dl=优, wl=优, hx=优, sw=优}	=> {wlfk=理科}	0.090	0.93
#> [6]	{xb=男, sx=优, ls=优, wl=优, sw=优}	=> {wlfk=理科}	0.074	0.92
#> [7]	{xb=女, zz=优, sw=中}	=> {wlfk=文科}	0.070	0.95

规则保存与导出

#规则如何保存呢？如果是为了复用，存为rda文件就好

```
save(irules_pruned,  
      file = "rules.rda")
```

#如果是要导出规则，最好是将规则转换成数据框，然后另存为csv文件

```
irules_pruned_in_df <- as(irules_pruned, "data.frame")  
View(irules_pruned_in_df)  
write.csv(irules_pruned_in_df,  
          file = "Rules.csv",  
          quote = TRUE,  
          row.names = FALSE)
```


规则保存与导出

rules	support	confidence	lift	count
{xb=男,sx=优,ls=优,wl=优,hx=优,sw=优} => {wlfk=理科}	0.07354839	0.9344262	1.900736	57
{xb=男,sx=优,dl=优,wl=优,hx=优,sw=优} => {wlfk=理科}	0.09032258	0.9333333	1.898513	70
{xb=男,dl=优,wl=优,hx=优,sw=优} => {wlfk=理科}	0.09935484	0.9277108	1.887076	77
{xb=男,sx=优,wl=优,hx=优,sw=优} => {wlfk=理科}	0.09161290	0.9220779	1.875618	71
{xb=男,sx=优,dl=优,wl=优,sw=优} => {wlfk=理科}	0.09032258	0.9210526	1.873532	70
{xb=男,sx=优,dl=优,wl=优,hx=优} => {wlfk=理科}	0.10451613	0.9204545	1.872316	81
{xb=男,sx=优,ls=优,wl=优,sw=优} => {wlfk=理科}	0.07354839	0.9193548	1.870079	57
{xb=女,zz=优,sw=中} => {wlfk=文科}	0.06967742	0.9473684	1.863478	54
{xb=男,sx=优,wy=优,wl=优,hx=优} => {wlfk=理科}	0.06838710	0.9137931	1.858765	53
{xb=男,yw=良,sx=优,dl=优,wl=优} => {wlfk=理科}	0.06838710	0.9137931	1.858765	53
{xb=男,sx=优,wl=优,hx=优} => {wlfk=理科}	0.10838710	0.9130435	1.857241	84
{xb=男,ls=优,wl=优,hx=优,sw=优} => {wlfk=理科}	0.08129032	0.9130435	1.857241	63

规则保存与导出

#当然，在另存为csv之前，也可以对规则进行必要的处理

```
irules_pruned_in_df %<>%  
  separate(  
    rules,  
    sep = "=>",  
    into = c("LHS", "RHS")) %>%  
  mutate_at(  
    vars("LHS", "RHS"),  
    funs(gsub("[\\{\\}] ", "", .)))
```

规则保存与导出

LHS	RHS	support	confidence	lift	count
xb=男,sx=优,ls=优,wl=优,hx=优,sw=优	wlfk=理科	0.07354839	0.9344262	1.900736	57
xb=男,sx=优,dl=优,wl=优,hx=优,sw=优	wlfk=理科	0.09032258	0.9333333	1.898513	70
xb=男,dl=优,wl=优,hx=优,sw=优	wlfk=理科	0.09935484	0.9277108	1.887076	77
xb=男,sx=优,wl=优,hx=优,sw=优	wlfk=理科	0.09161290	0.9220779	1.875618	71
xb=男,sx=优,dl=优,wl=优,sw=优	wlfk=理科	0.09032258	0.9210526	1.873532	70
xb=男,sx=优,dl=优,wl=优,hx=优	wlfk=理科	0.10451613	0.9204545	1.872316	81
xb=男,sx=优,ls=优,wl=优,sw=优	wlfk=理科	0.07354839	0.9193548	1.870079	57
xb=女,zz=优,sw=中	wlfk=文科	0.06967742	0.9473684	1.863478	54
xb=男,sx=优,wy=优,wl=优,hx=优	wlfk=理科	0.06838710	0.9137931	1.858765	53
xb=男,yw=良,sx=优,dl=优,wl=优	wlfk=理科	0.06838710	0.9137931	1.858765	53
xb=男,sx=优,wl=优,hx=优	wlfk=理科	0.10838710	0.9130435	1.857241	84
xb=男,ls=优,wl=优,hx=优,sw=优	wlfk=理科	0.08129032	0.9130435	1.857241	63

关联规则回顾

关联规则的挖掘以事务为单位，寻求的是项集与项集之间的联系

$A \Rightarrow B$ ：并非因果关系，只是伴随关系。后项伴随着前项的出现，也可能是前项引发、诱发了后项的出现

任何 $Lift < 1$ 的规则都不能显示是一个真正的内在伴随现象，无论他的支持度和置信度有多高

Apriori 算法基于先验原理：非频繁项集的超集必然是非频繁的

$A \Rightarrow B$ 置信度高和 $B \Rightarrow A$ 置信度高：共生关系

$A \Rightarrow B$ 置信度高和 $B \Rightarrow A$ 置信度低：寄生关系

$A \Rightarrow B$ 置信度低和 $B \Rightarrow A$ 置信度低：没有关系

关联规则回顾

用一句话来形容关联规则的挖掘

所谓关联规则的学习，其实就是观察历史记录

如果：

B总是频繁地和A一起出现（**支持度**）

当A出现时，B出现的概率很大（**置信度**），

甚至是更大（**提升度**）

那么：

很自然形成一条关联规则 $A \Rightarrow B$

那种算法更好

除了经典的Apriori

还有好的新的算法如：

FP-Growth、Eclat等

它们能挖出更多更好的规则么

（在支持度、置信度框架下）



延伸阅读

IDA@SMU

Intelligent Data Analysis Lab



Navigation

Home
Team
Projects
Software
Data sets
Contact us

Projects

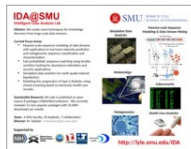
Hurricanes (PIIH)
Genomics (QuasiAlign)
Streams (TRACDS)
recommenderlab
arules
seriation

Affiliation

EMIS, CSE
Lyle, SMU

About the Intelligent Data Analysis Lab at SMU

At IDA@SMU we create novel techniques inspired by knowledge discovery, data mining, machine learning, artificial intelligence, data analytics and statistical learning to work with data from various sources. We currently focus on creating novel **techniques for knowledge discovery from large scale data streams** with applications in storm prediction, security metagenomics, health screening, and recommender systems. All projects are accompanied by code published as open source R packages (CRAN/BioConductor). We currently maintain 15 very popular packages. > read about our current projects



IDA@SMU is part of the Bobby B. Lyle School of Engineering, SMU. The lab was created in 2009 when Dr. Hahsler joined as director. It is the successor of SMU's Database Research Group founded by Dr. Dunham.

Acknowledgement of Support



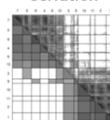
arulesVis



EMMSA



seriation



<http://lyle.smu.edu/IDA/arules/> [Accessed on 2018-2-1]

A decorative blue border with rounded corners frames the entire slide. Two thin blue lines, one horizontal and one vertical, intersect to form a crosshair in the upper right quadrant of the slide.

谢谢聆听
Thank you

教师个人联系方式

艾新波

手机: 13641159546

QQ: 23127789

微信: 13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址: 北京邮电大学科研楼917室

课程网址: <https://github.com/byaxb/RDataAnalytics>

