大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

北京理工大学计算机学院　孙新

2019年1月

4.5 Apriori关联规则算法

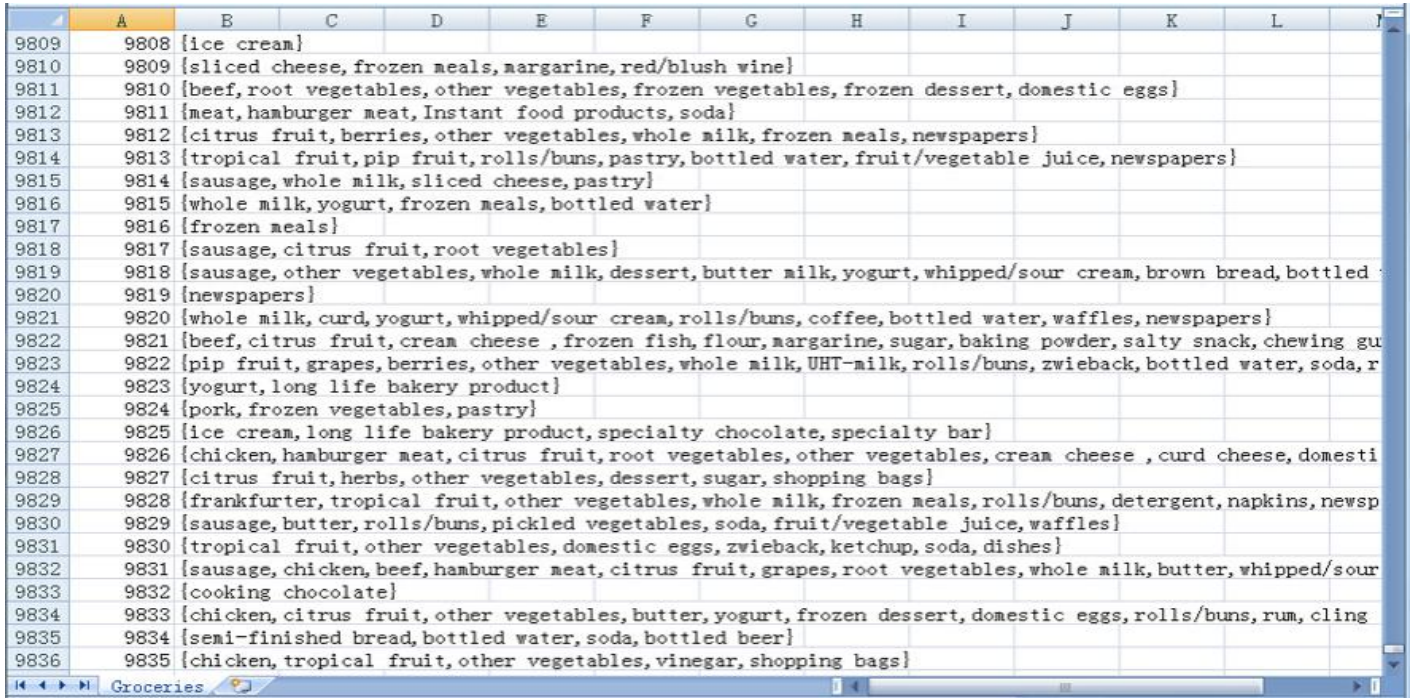
- 关联规则挖掘的基本思想
- Apriori 算法的基本过程
- Apriori关联规则算法步骤
- Apriori关联规则算法的特点
- R语言实现Apriori算法示例

- 利用超市购物篮Groceries数据进行关联规则分析，使用R语言中的arules包实现Apriori算法
- 导入包：library(arules)；加载数据源： Groceries数据集

```
> library(arules)  #加载 arules 包
> data(Groceries)
> Groceries
transactions in sparse format with
 9835 transactions (rows) and
 169 items (columns)
```

案例分析—数据源

- **Groceries数据集是来自一个现实世界中的超市经营一个月的购物数据，包含了9835次交易，以及169件商品。**



	A	B
9809	9808	{ice cream}
9810	9809	{sliced cheese, frozen meals, margarine, red/blush wine}
9811	9810	{beef, root vegetables, other vegetables, frozen vegetables, frozen dessert, domestic eggs}
9812	9811	{meat, hamburger meat, Instant food products, soda}
9813	9812	{citrus fruit, berries, other vegetables, whole milk, frozen meals, newspapers}
9814	9813	{tropical fruit, pip fruit, rolls/buns, pastry, bottled water, fruit/vegetable juice, newspapers}
9815	9814	{sausage, whole milk, sliced cheese, pastry}
9816	9815	{whole milk, yogurt, frozen meals, bottled water}
9817	9816	{frozen meals}
9818	9817	{sausage, citrus fruit, root vegetables}
9819	9818	{sausage, other vegetables, whole milk, dessert, butter milk, yogurt, whipped/sour cream, brown bread, bottled
9820	9819	{newspapers}
9821	9820	{whole milk, curd, yogurt, whipped/sour cream, rolls/buns, coffee, bottled water, waffles, newspapers}
9822	9821	{beef, citrus fruit, cream cheese , frozen fish, flour, margarine, sugar, baking powder, salty snack, chewing gu
9823	9822	{pip fruit, grapes, berries, other vegetables, whole milk, UHT-milk, rolls/buns, zwieback, bottled water, soda, r
9824	9823	{yogurt, long life bakery product}
9825	9824	{pork, frozen vegetables, pastry}
9826	9825	{ice cream, long life bakery product, specialty chocolate, specialty bar}
9827	9826	{chicken, hamburger meat, citrus fruit, root vegetables, other vegetables, cream cheese , curd cheese, domesti
9828	9827	{citrus fruit, herbs, other vegetables, dessert, sugar, shopping bags}
9829	9828	{frankfurter, tropical fruit, other vegetables, whole milk, frozen meals, rolls/buns, detergent, napkins, newsp
9830	9829	{sausage, butter, rolls/buns, pickled vegetables, soda, fruit/vegetable juice, waffles}
9831	9830	{tropical fruit, other vegetables, domestic eggs, zwieback, ketchup, soda, dishes}
9832	9831	{sausage, chicken, beef, hamburger meat, citrus fruit, grapes, root vegetables, whole milk, butter, whipped/sour
9833	9832	{cooking chocolate}
9834	9833	{chicken, citrus fruit, other vegetables, butter, yogurt, frozen dessert, domestic eggs, rolls/buns, rum, cling
9835	9834	{semi-finished bread, bottled water, soda, bottled beer}
9836	9835	{chicken, tropical fruit, other vegetables, vinegar, shopping bags}

Groceries

案例分析—加载数据集

- **数据转换：创建稀疏矩阵，每个Item一列，每一行代表一个transaction。1表示该transaction购买了该item，0表示没有购买。**

```
groceries <- read.transactions("groceries.csv", format="basket", sep=",")
```

ID	Whole milk	…	sausage
1	1	0	1
…	0	1	1
9835	1	0	0

稀疏矩阵

案例分析—探索数据

- **通过inspect（）函数可以看到超市的交易记录，每次交易的商品名称**

```
> inspect(Groceries[1:5])    #通过inspect函数查看Groceries数据集的前5次交易记录
  items
1 {citrus fruit,semi-finished bread,margarine,ready soups}
2 {tropical fruit,yogurt,coffee}
3 {whole milk}
4 {pip fruit,yogurt,cream cheese ,meat spreads}
5 {other vegetables,whole milk,condensed milk,long life bakery product}
```

案例分析—探索数据

通过summary（）函数可以查看该数据集的一些基本信息。

```
summary(groceries)
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:
      whole milk other vegetables       rolls/buns             soda
            2513             1903             1809             1715
          yogurt          (Other)
            1372            34055

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55
  16   17   18   19   20   21   22   23   24   26   27   28   29   32
  46   29   14   14    9   11    4    6    1    1    1    1    3    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   4.409   6.000  32.000

includes extended item information - examples:
            labels
1 abrasive cleaner
2 artif. sweetener
3   baby cosmetics
```

利用itemFrequency（）函数可以查看商品交易比列。也即，取数据集第100到800行，第1列到第3列，计算列代表的三个项目对应的支持度。当然，也可以把支持度itemFrequency排序，查看支持度的最大值，也即取前10个项目对应的支持度。

```
> itemFrequency(groceries[100:800,1:3])
abrasive cleaner artif. sweetener   baby cosmetics
      0.005706134       0.001426534       0.001426534
```

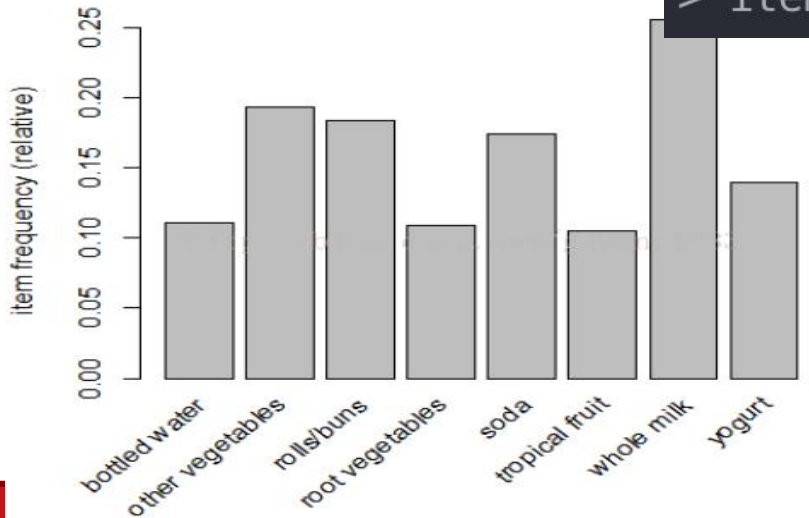
利用sort（）函数可以对支持度进行排序

```
> orderedItemFreq <- sort(itemFrequency(groceries), decreasing=T)
> orderedItemFreq[1:10]
      whole milk other vegetables       rolls/buns             soda           yogurt    bottled water
      0.25551601       0.19349263       0.18393493       0.17437722       0.13950178       0.11052364
 root vegetables   tropical fruit    shopping bags         sausage
      0.10899847       0.10493137       0.09852567       0.09395018
```

可视化商品的支持度——商品的频率图

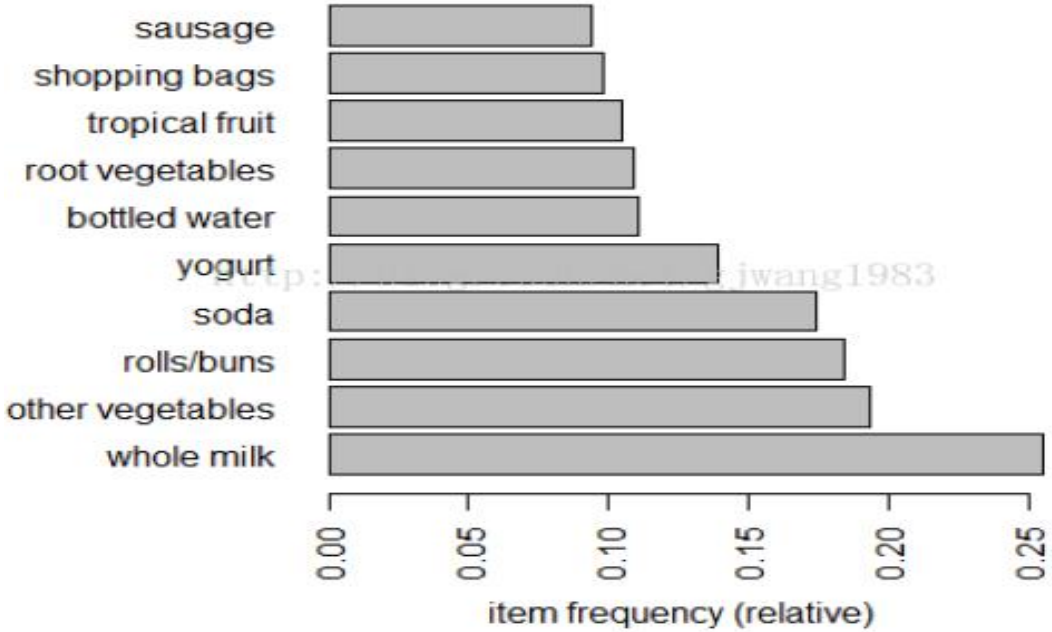
为了直观地呈现统计数据，可以使用itemFrequenctyPlot()函数生成一个用于描绘所包含的特定商品的交易比例的柱状图。因为包含很多种商品，不可能同时展现出来，因此可以通过support或者topN参数进行排除一部分商品进行展示。support = 0.1 表示支持度至少为0.1。

```
> itemFrequencyPlot(groceries, support=0.1)
```



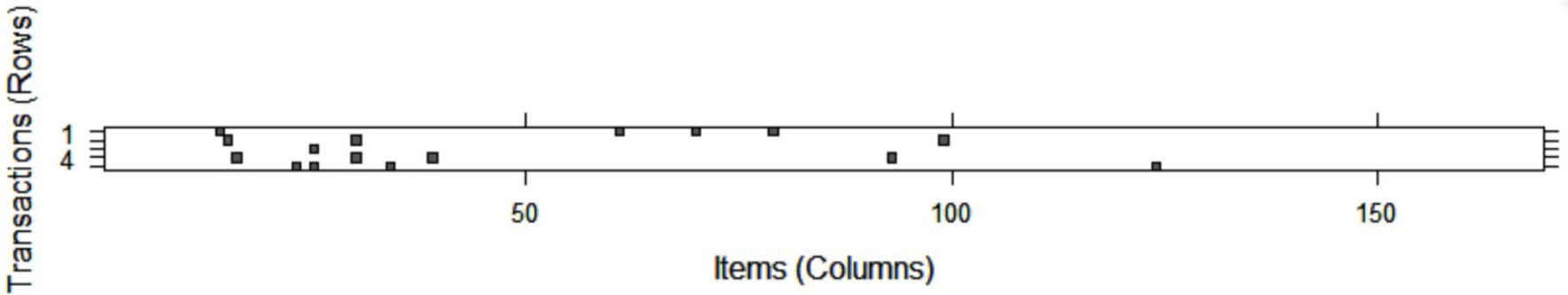
可视化商品的支持度——商品的频率图，topN = 10表示支持度排在前10的商品。

```
> itemFrequencyPlot(groceries, topN=10, horiz=T)
```



可视化交易数据——绘制稀疏矩阵

```
image(Groceries[1:5]) # 生成一个5行169列的矩阵，矩阵中填充有黑色的单元表示在此次交易（行）中，该商品（列）被购买了
```



利用apriori()函数，可以进行规制挖掘。

默认设置 support = 0.1 ， confidence = 0.8

```
grocery_rules <- apriori(data=Groceries,parameter=list(support =,confidence =,minlen =))
```

设置支持度和置信度参数来产生合理数量的关联规则时，可能需要进行大量的试验与误差评估。

案例分析—训练模型

使用R语言进行规则挖掘，参数设置support = 0.006 ， confidence = 0.25

```
> groceryrules <- apriori(groceries, parameter = list(support =
+                                                     0.006, confidence = 0.25, minlen = 2))

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
       0.25    0.1    1 none FALSE            TRUE   0.006      2     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 ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [109 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.01s].
writing ... [463 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

使用summary（）函数查看规则规汇总信息

```
> summary(groceryrules)
set of 463 rules

rule length distribution (lhs + rhs):sizes
  2   3   4
150 297  16

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.711   3.000   4.000

summary of quality measures:
     support            confidence          lift
 Min.   :0.006101   Min.   :0.2500   Min.   :0.9932
 1st Qu.:0.007117   1st Qu.:0.2971   1st Qu.:1.6229
 Median :0.008744   Median :0.3554   Median :1.9332
 Mean   :0.011539   Mean   :0.3786   Mean   :2.0351
 3rd Qu.:0.012303   3rd Qu.:0.4495   3rd Qu.:2.3565
 Max.   :0.074835   Max.   :0.6600   Max.   :3.9565

mining info:
      data ntransactions support confidence
 groceries          9835   0.006       0.25
```

使用inspect（）查看具体的规则

```
> inspect(groceryrules[1:5])
  lhs                    rhs                    support     confidence lift
1 {potted plants}     => {whole milk}           0.006914082 0.4000000  1.565460
2 {pasta}             => {whole milk}           0.006100661 0.4054054  1.586614
3 {herbs}             => {root vegetables}      0.007015760 0.4312500  3.956477
4 {herbs}             => {other vegetables}     0.007727504 0.4750000  2.454874
5 {herbs}             => {whole milk}           0.007727504 0.4750000  1.858983
```

使用sort（）对关联规则集合排序

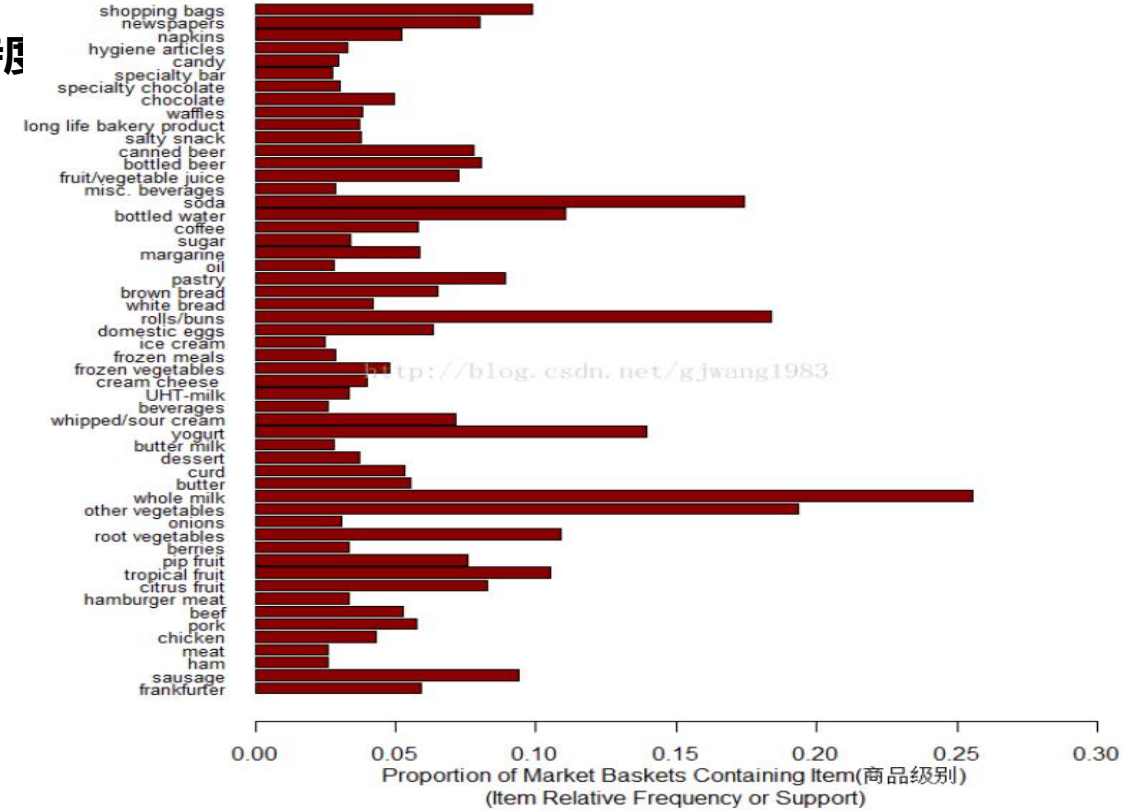
```
> ordered_groceryrules <- sort(groceryrules, by="lift")
> inspect(ordered_groceryrules[1:5])
  lhs                     rhs                      support confidence     lift
1 {herbs}              => {root vegetables}     0.007015760  0.4312500 3.956477
2 {berries}            => {whipped/sour cream}  0.009049314  0.2721713 3.796886
3 {other vegetables,
   tropical fruit,
   whole milk}         => {root vegetables}     0.007015760  0.4107143 3.768074
4 {beef,
   other vegetables}   => {root vegetables}     0.007930859  0.4020619 3.688692
5 {other vegetables,
   tropical fruit}     => {pip fruit}           0.009456024  0.2634561 3.482649
```

使用itemInfo（）进行查看summary（）函数结果中level1和level2字段的详细信息

```
> print(levels(itemInfo(Groceries)[["level1"]]))
 [1] "canned food"         "detergent"           "drinks"              "fresh products"      "fruit and vegetables" "meet and sausage"    "non-food"
 [8] "perfumery"           "processed food"      "snacks and candies"
> print(levels(itemInfo(Groceries)[["level2"]]))
 [1] "baby food"                 "bags"                      "bakery improver"           "bathroom cleaner"
 [5] "beef"                      "beer"                      "bread and backed goods"    "candy"
 [9] "canned fish"               "canned fruit/vegetables"   "cheese"                    "chewing gum"
[13] "chocolate"                 "cleaner"                   "coffee"                    "condiments"
[17] "cosmetics"                 "dairy produce"             "delicatessen"              "dental care"
[21] "detergent/softener"        "eggs"                      "fish"                      "frozen foods"
[25] "fruit"                     "games/books/hobby"         "garden"                    "hair care"
[29] "hard drinks"               "health food"               "jam/sweet spreads"         "long-life bakery products"
[33] "meat spreads"              "non-alc. drinks"           "non-food house keeping products" "non-food kitchen"
[37] "packaged fruit/vegetables" "perfumery"                 "personal hygiene"          "pet food/care"
[41] "pork"                      "poultry"                   "pudding powder"            "sausage"
[45] "seasonal products"         "shelf-stable dairy"        "snacks"                    "soap"
[49] "soups/sauces"              "staple foods"              "sweetener"                 "tea/cocoa drinks"
[53] "vegetables"                "vinegar/oils"              "wine"
```

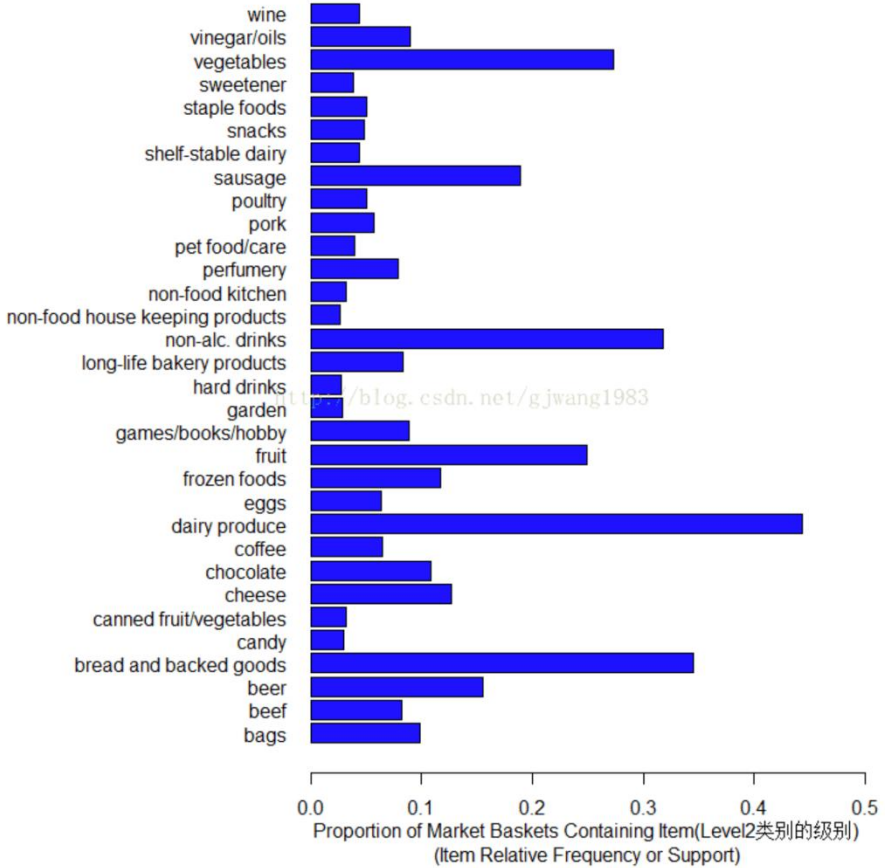
案例分析—模型评估（进阶）

aggregate前的商品（item）支持

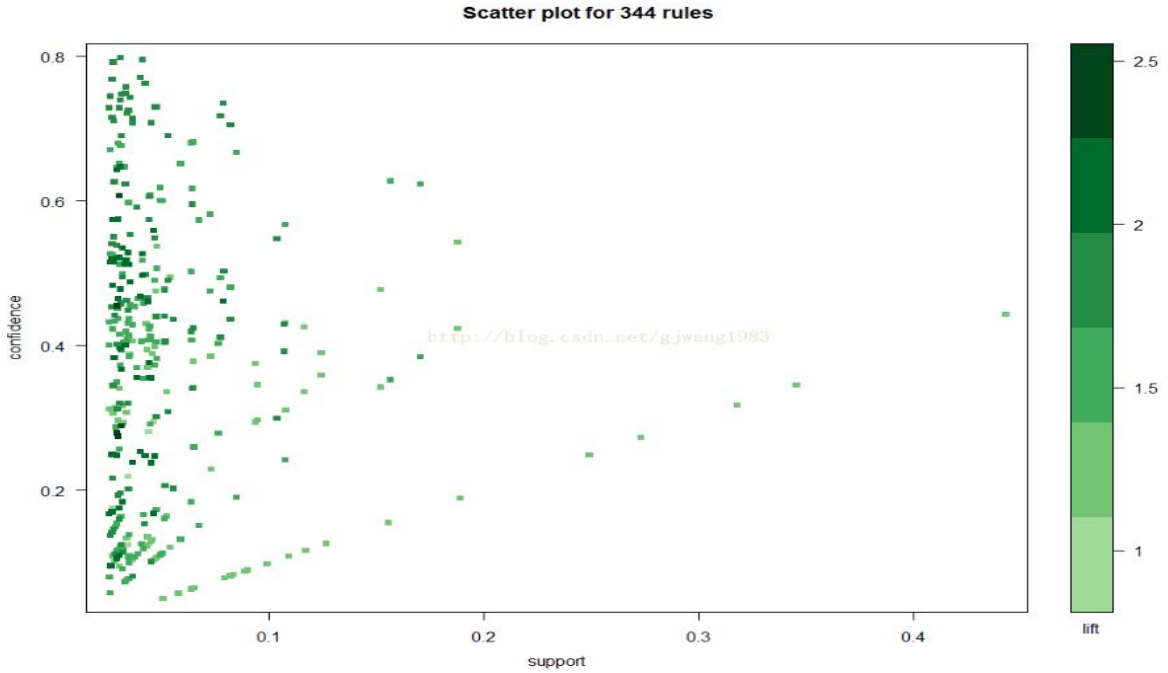


案例分析—模型评估（进阶）

aggregate后**类别（category）支持度图**

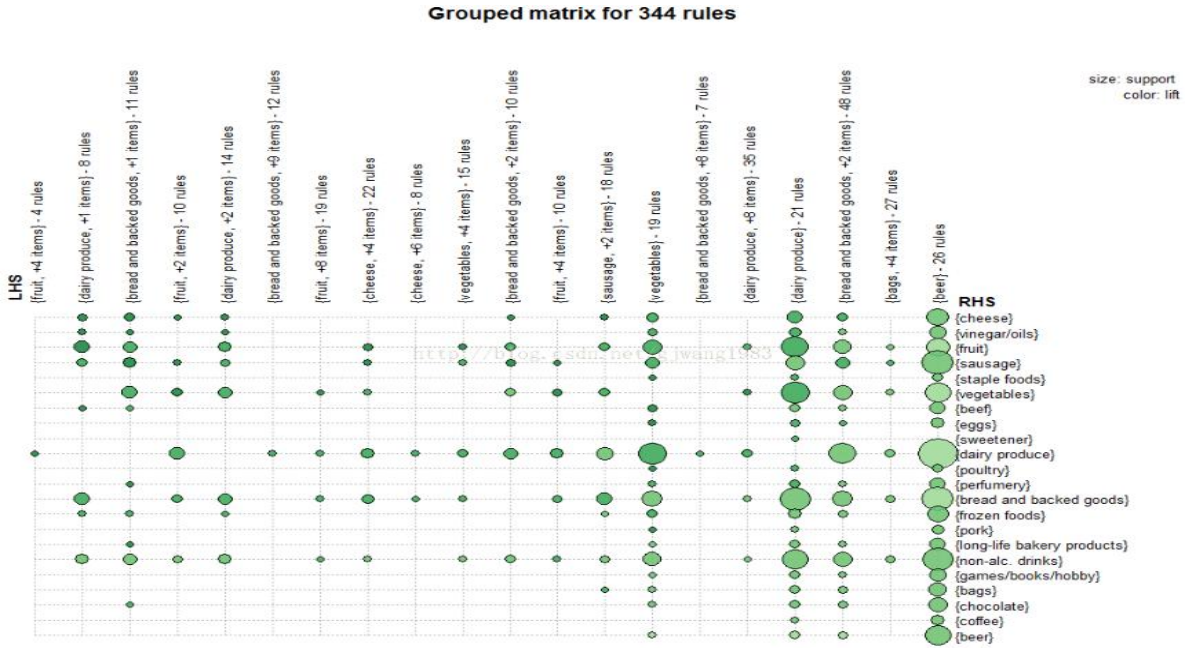


这幅散点图表示了规则的分布图，横轴支持度，纵轴置信度，颜色深浅为提升度大部分规则的support在0.1以内，Confidence在0-0.8内。



案例分析—模型评估

最后，给出grouped图以有向网状图的形式展示关联规则，图中横坐标为规则前项，纵坐标为规则后项，圆圈表示关联规则，圆圈大小表示支持度大小；颜色深浅代表规则提升度的高低。



谢 谢

Thank you for your attention!