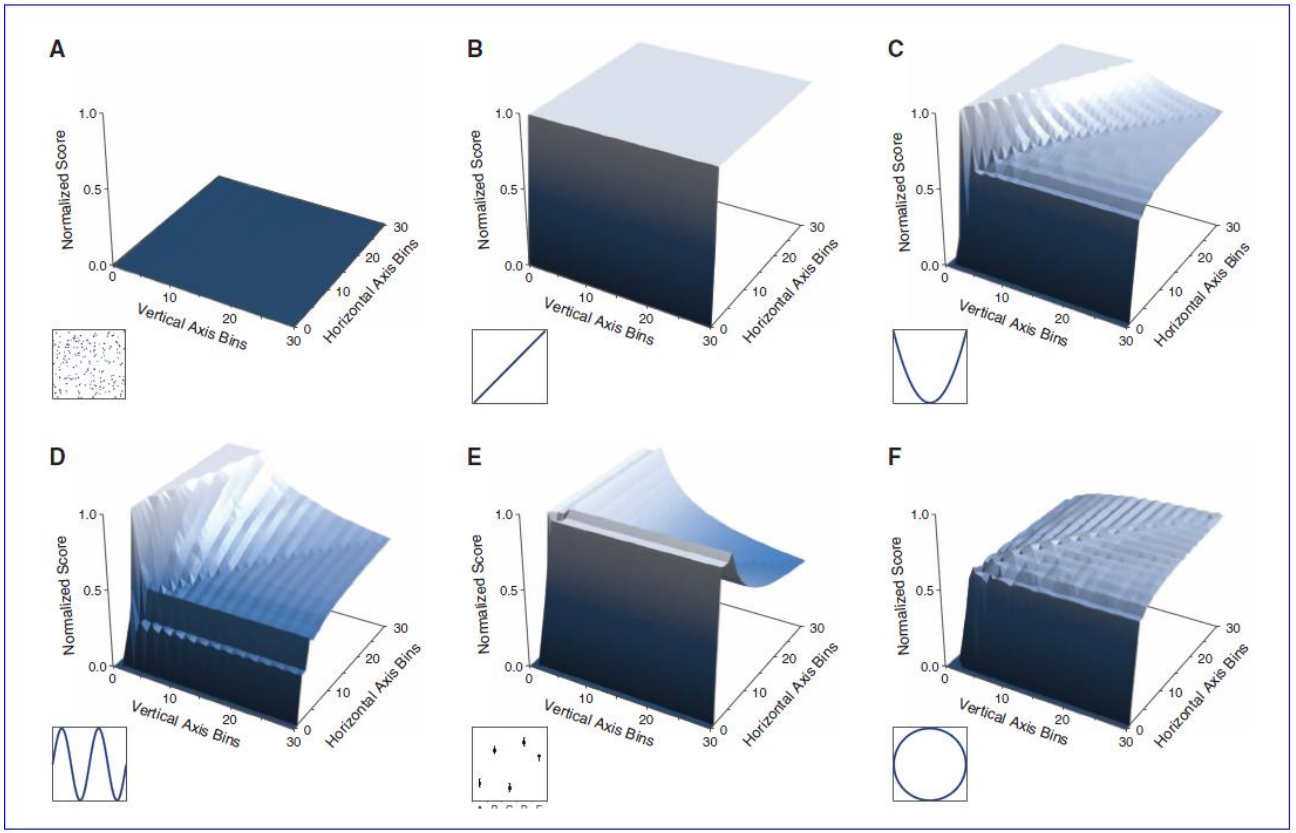
DATAGURU 炼数成金

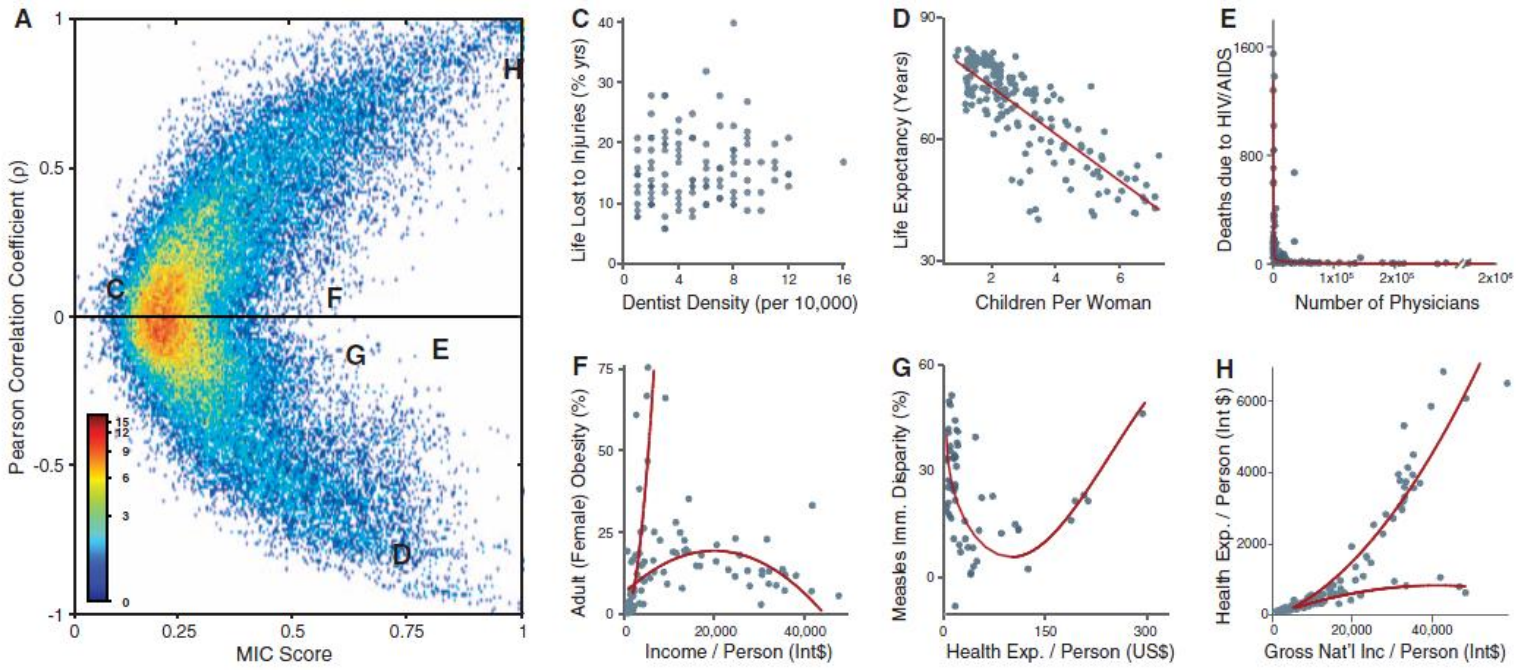


数据分析与R语言 第6周

2012.6.12

传统回归模型的困难

- 为什么一定是线性？或某种非线性模型？
- 过分依赖于分析者的经验
- 对于非连续的离散数据难以处理



2012.6.12

网格方法

- 《Science》上的文章《Detecting Novel Associations in Large Data Sets》
- 方法概要：用网格判断数据的集中程度，集中程度意味着是否有关联关系
- 方法具有一般性，即无论数据是怎样分布的，不限于特定的关联函数类型，此判断方法都是有效
- 方法具有等效性，计算的熵值和噪音的程度有关，跟关联的类型无关
- MIC：the Maximal Information Coefficient
- MINE：Maximal Information-based Nonparametric Exploration

MIC值计算

传统能看出来的,MIC都能看出来
MIC能看出来,传统不一定

G也不能太细,细到了每个网格一个点,就没意义了

- 坐标平面被划分为(x,y)网格G（未必等宽），其中$xy<n^{0.6}$
- 在G上可以诱导出“自然概率密度函数”p(x,y)，任何一个方格（box）内的概率密度函数值为这个方格所 包含的样本点数量占全体样本点的比例
- 计算网格划分G下的 **mutual information值** I_G

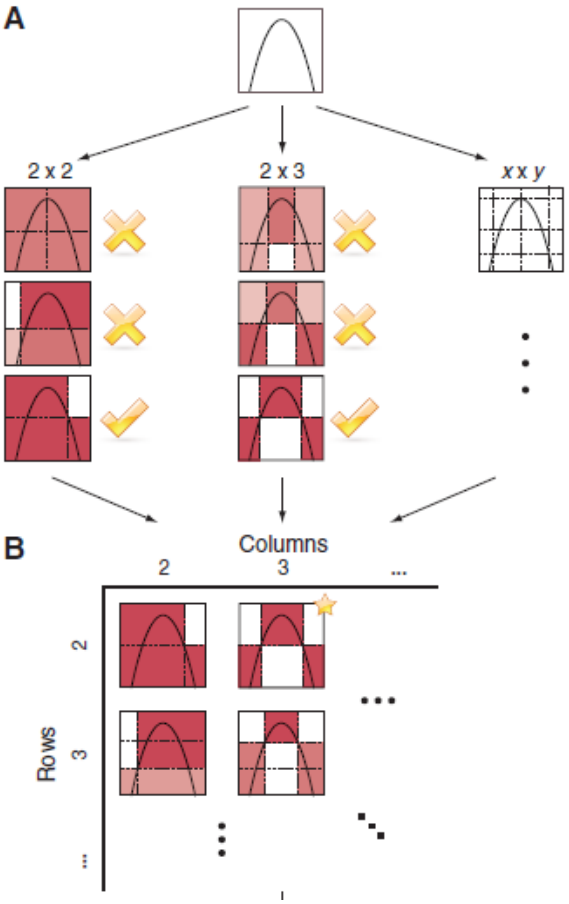
交互信息系数

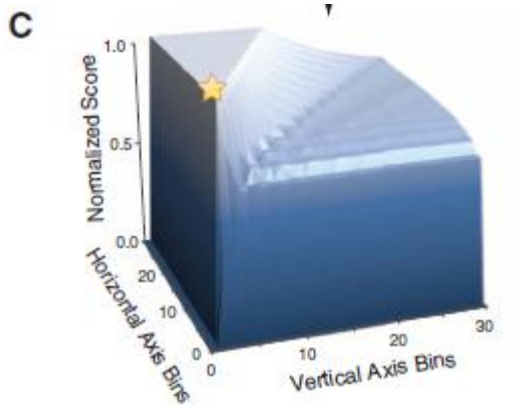
每一种网格划分我都算出来一个I_G

$$I(X;Y) = \int_Y \int_X p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right) dx\, dy$$

MIC值计算

- 构造**特征矩阵**{m_{xy}}，矩阵的元素m_{xy}=max{I_G}/log min{x,y}。max取遍所有可能的(x,y)网格G
- MIC=max {m_{xy}}。Max取遍所有可能的(x,y)对





MIC值计算

- Mxy的计算是个难点，数据科学家构造了一个近似的逼近算法以提高效率

http://www.sciencemag.org/content/suppl/2011/12/14/334.6062.1518.DC1

在作者的网站上，可以下载MINE计算MIC的程序（Java和R）以及测试用数据集

http://www.exploredata.net/Downloads

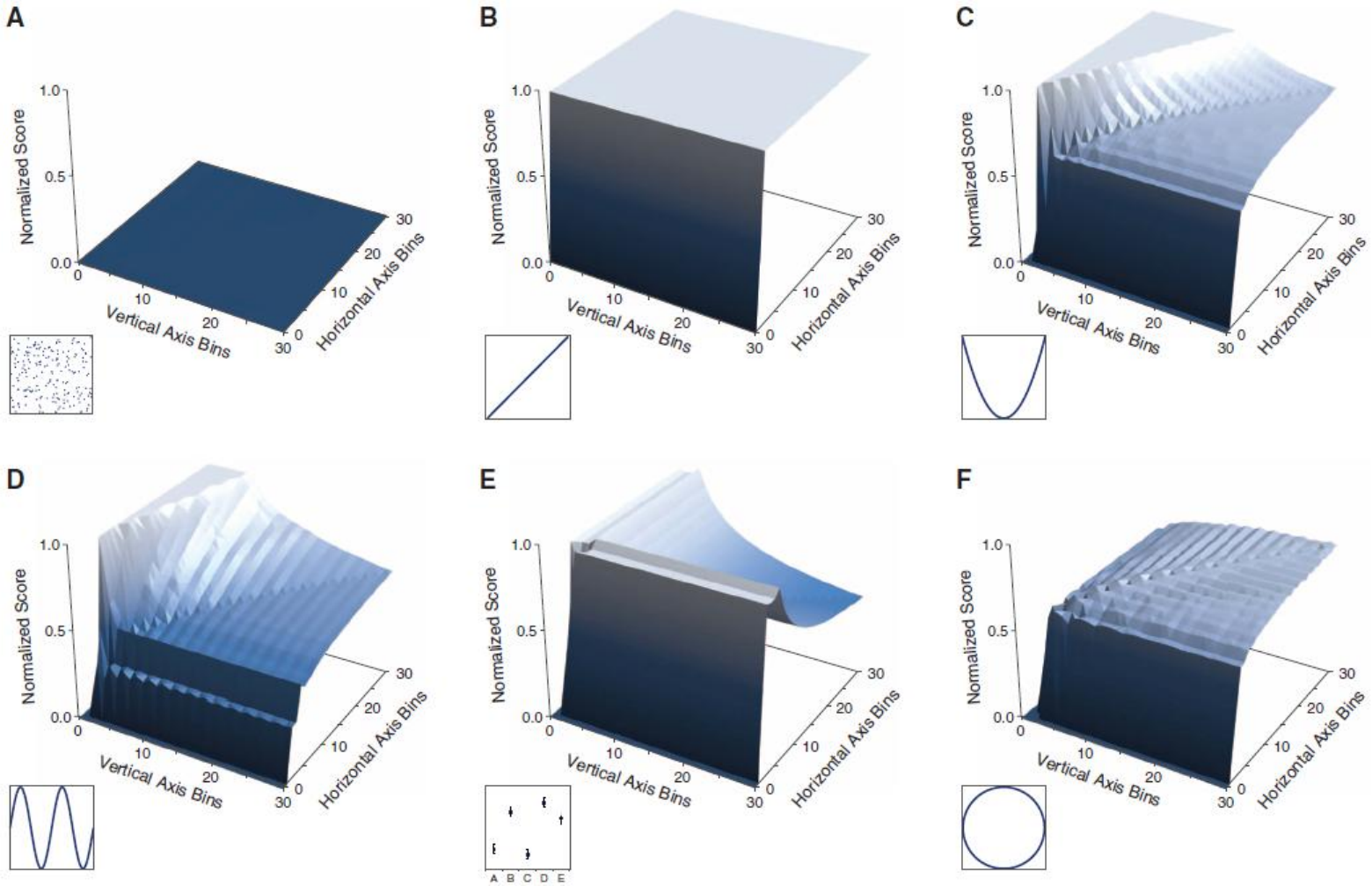
实验：WHO数据集，垒球数据集...

MIC的性质

- 如果变量对x,y存在函数关系，则当样本数增加时，MIC必然趋向于1
- 如果变量对x,y可以由参数方程c(t)=[x(t),y(t)]所表达的曲线描画，则当样本数增加时，MIC必然趋于1
- 如果变量对x,y在统计意义下互相独立，则当样本数增加时，MIC趋于0

MIC观察

只要有规律,都会趋向于1, 说明了某种关联关系

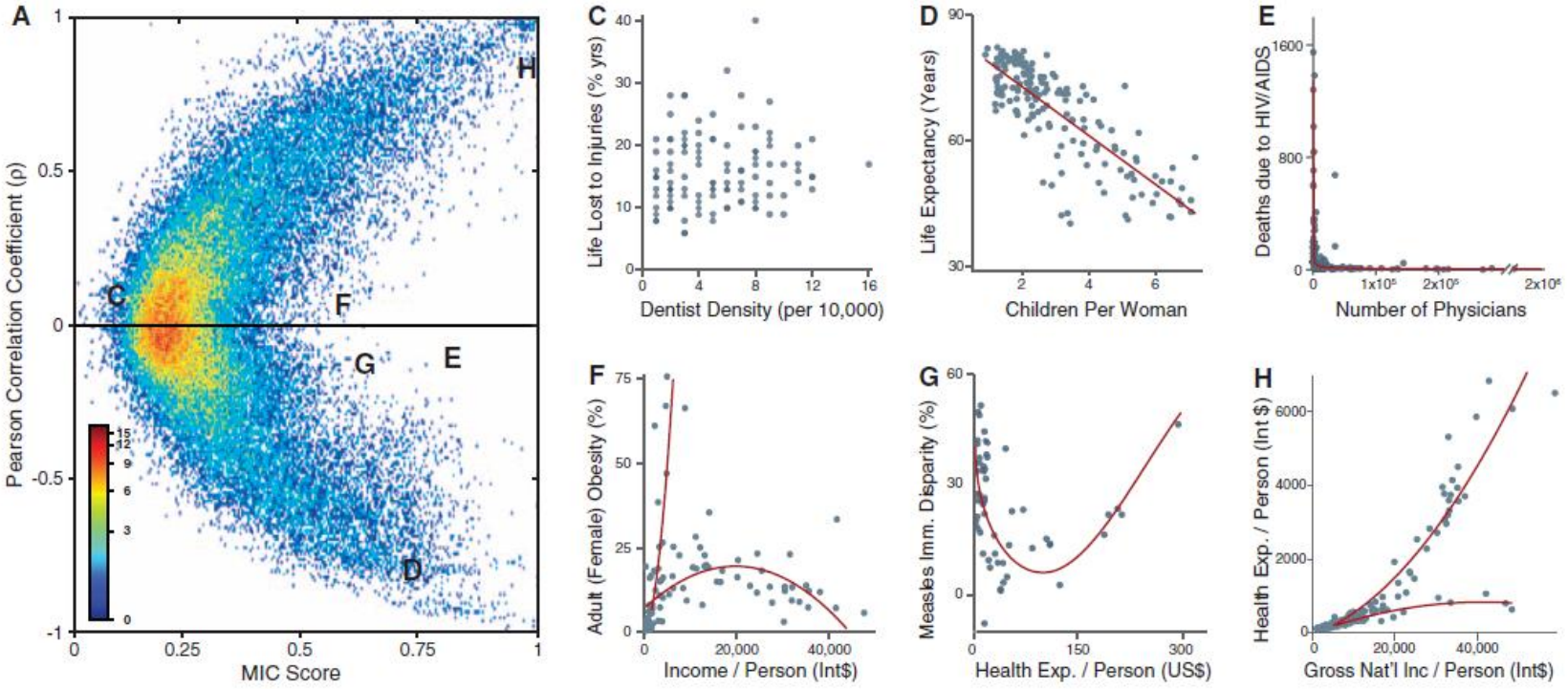


2012.6.12

样本点都集中在几个点上面

MIC与线性回归模型对比

纵轴是相关系数,上面是1,下面是-1
横轴是MIC值



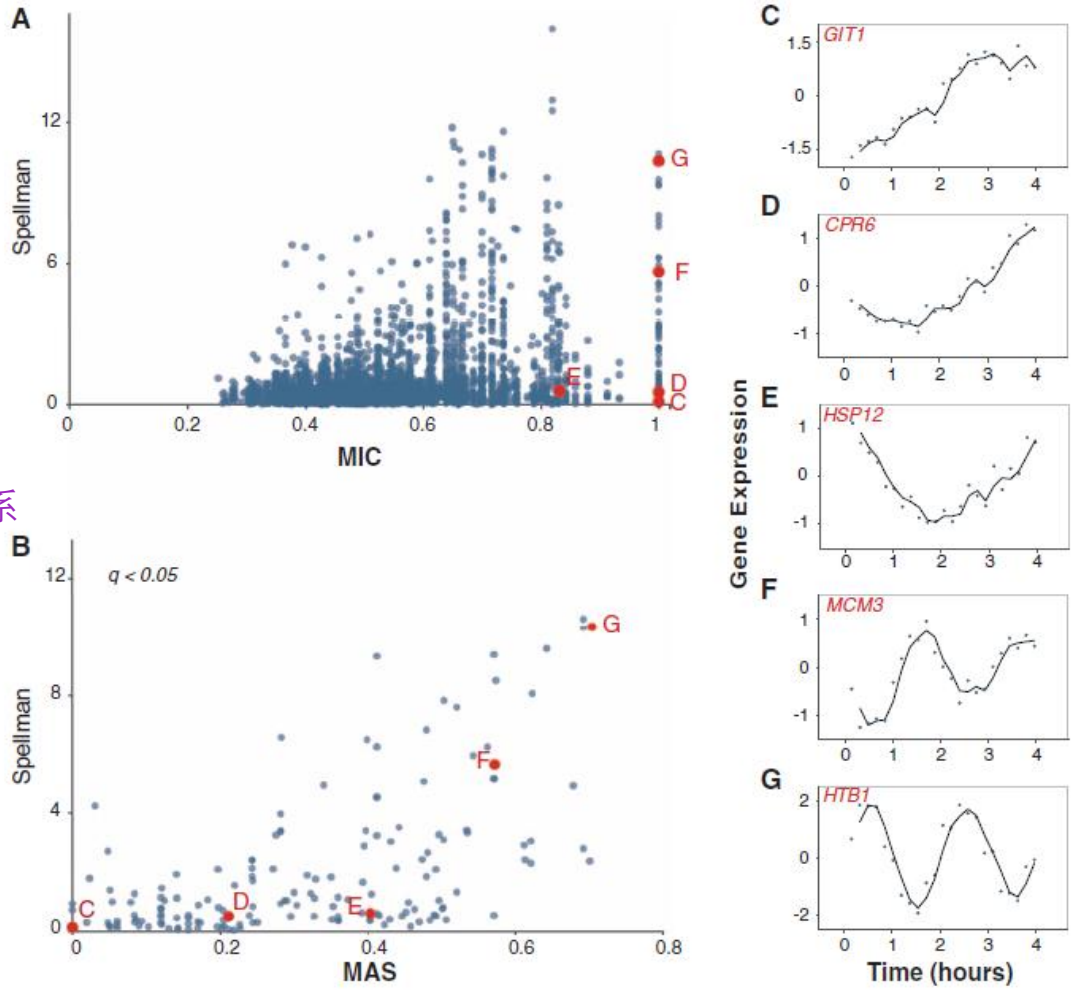
FGE三幅图,相关系数分析认为相关性很低,但是MIC值认为很高,说了MIC法相对于传统方法的优越性

2012.6.12

对基因数据集spellman的探索

- 数据集包含6223组基因数据
- MINE对关联关系的辨认力明显强于以往的方法，例如双方都发现了HTB1，但MINE方法挖出了过去未被发现的HSP12

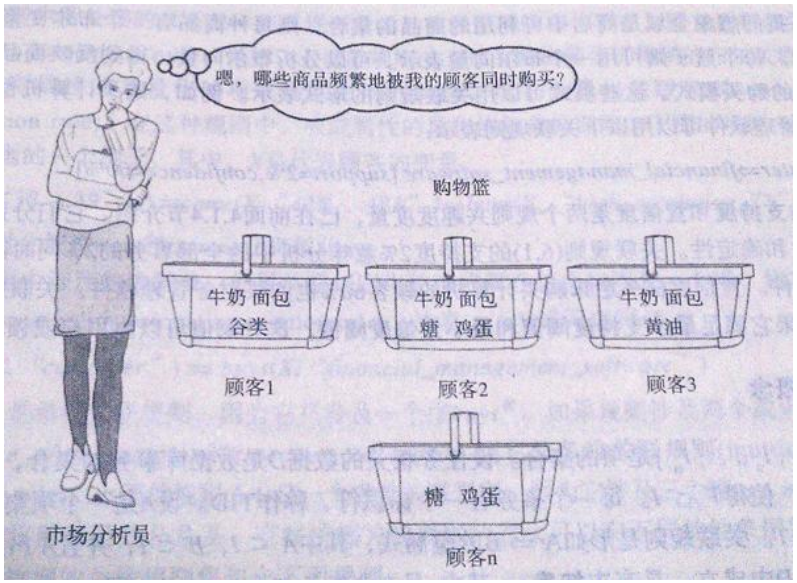
用这种方法发现了很多新的基因相关关系



2012.6.12

数据挖掘：关联规则挖掘

- 例子：购物篮分析



比如超市常常故意把会一起买的东西分开放得很远,可以让人们买的时候经过更多货架,提升销量

名词

- 挖掘数据集：购物篮数据
- 挖掘目标：关联规则
- 关联规则：牛奶=>鸡蛋【支持度=2%，置信度=60%】
- 支持度：分析中的全部事务的2%同时购买了牛奶和鸡蛋
- 置信度：购买了牛奶的筒子有60%也购买了鸡蛋
- 最小支持度阈值和最小置信度阈值：由挖掘者或领域专家设定 一次证明不是偶尔发生的,而是经常发生

名词

- 项集：项（商品）的集合
- k-项集：k个项组成的项集
- 频繁项集：满足最小支持度的项集，频繁k-项集一般记为L_k
- 强关联规则：满足最小支持度阈值和最小置信度阈值的规则

关联规则挖掘路线图

- 两步过程：找出所有频繁项集；由频繁项集产生强关联规则
- 算法：Apriori　　以最小支持度2作为门槛
- 例子

共九个购物篮,T100,T200,一直到T900

TID	项ID的列表
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

I1,I2,代表各种商品

图6-2　AllElectronics某分店的事务数据

Apriori算法的工作过程

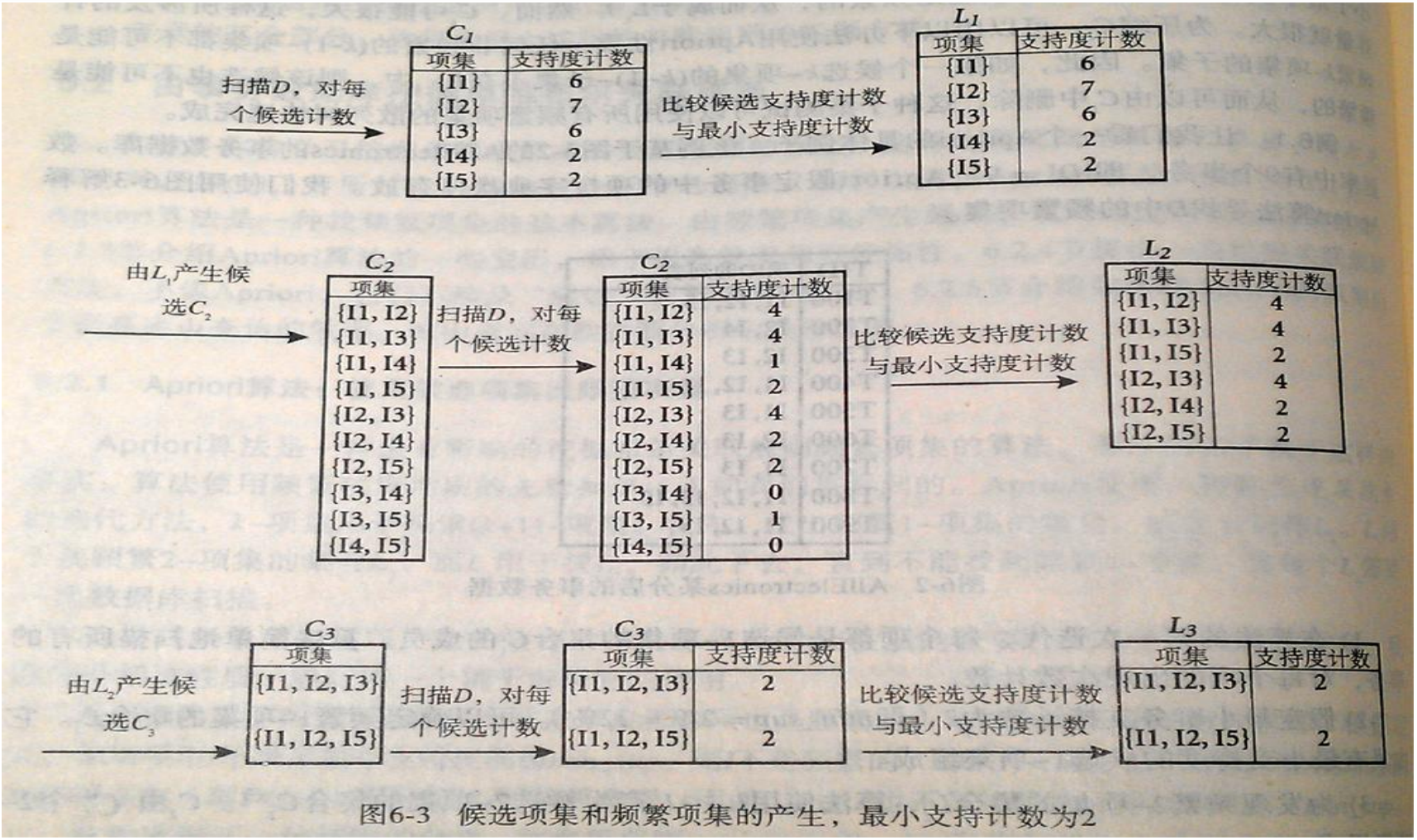


图6-3 候选项集和频繁项集的产生，最小支持计数为2

2012.6.12

步骤说明

- 扫描D，对每个候选项计数，生成候选1-项集C1
- 定义最小支持度阈值为2，从C1生成频繁1-项集L1
- 通过L1xL1生成候选2-项集C2
- 扫描D，对C2里每个项计数，生成频繁2-项集L2
- 计算L3xL3，利用apriori性质：频繁项集的子集必然是频繁的，我们可以删去一部分项，从而得到C3，由C3再经过支持度计数生成L3
- 可见Apriori算法可以分成 **连接，剪枝** 两个步骤不断循环重复

由频繁项集提取关联规则

- 例子：我们计算出频繁项集{I1,I2,I5}，能提取哪些规则？

I1^I2=>I5，由于{I1,I2,I5}出现了2次，{I1,I2}出现了4次，故置信度为2/4=50%

类似可以算出

$$
\begin{aligned}
I1 \wedge I2 &\Rightarrow I5, && confidence = 2/4 = 50\% \\
I1 \wedge I5 &\Rightarrow I2, && confidence = 2/2 = 100\% \\
I2 \wedge I5 &\Rightarrow I1, && confidence = 2/2 = 100\% \\
I1 &\Rightarrow I2 \wedge I5, && confidence = 2/6 = 33\% \\
I2 &\Rightarrow I1 \wedge I5, && confidence = 2/7 = 29\% \\
I5 &\Rightarrow I1 \wedge I2, && confidence = 2/2 = 100\%
\end{aligned}
$$

用 R 进行购物篮分析

- 安装arules包并加载
- 内置Groceries数据集

library(arules) #加载arules程序包

data(Groceries) #调用数据文件

Inspect(Groceries) #观看数据集里的数据

```
      specialty bar}
9823 {yogurt,
      long life bakery product}
9824 {pork,
      frozen vegetables,
      pastry}
9825 {ice cream,
      long life bakery product,
      specialty chocolate,
      specialty bar}
9826 {chicken,
      hamburger meat,
      citrus fruit,
```

用 R 进行购物篮分析

- 求频繁项集

frequentsets=eclat(Groceries,parameter=list(support=0.05,maxlen=10))

```
parameter specification:
 tidLists support minlen maxlen            target   ext
    FALSE    0.05      1     10 frequent itemsets FALSE

algorithmic control:
 sparse sort verbose
      7   -2    TRUE

eclat - find frequent item sets with the eclat algorithm
version 2.6 (2004.08.16)          (c) 2002-2004   Christian Borgelt
create itemset ...
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [28 item(s)] done [0.00s].
creating sparse bit matrix ... [28 row(s), 9835 column(s)] done [0.00s].
writing  ... [31 set(s)] done [0.02s].
Creating S4 object  ... done [0.00s].
```

用 R 进行购物篮分析

- 观看频繁项集

inspect(frequentsets[1:10])

inspect(sort(frequentsets,by="support")[1:10])　#根据支持度对求得的频繁项集排序并察看

```
    items                  support
1   {whole milk}        0.25551601
2   {other vegetables}  0.19349263
3   {rolls/buns}        0.18393493
4   {soda}              0.17437722
5   {yogurt}            0.13950178
6   {bottled water}     0.11052364
7   {root vegetables}   0.10899847
8   {tropical fruit}    0.10493137
9   {shopping bags}     0.09852567
10  {sausage}           0.09395018
```

用 R 进行购物篮分析

- 利用apriori函数提取关联规则

rules=apriori(Groceries,parameter=list(support=0.01,confidence=0.5))

```
> rules=apriori(Groceries,parameter=list(support=0.01,confidence=0.5))

parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE    0.01      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 ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [88 item(s)] done [0.00s].
creating transaction tree ... done [0.02s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [15 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

用 R 进行购物篮分析

- 列出关联规则

summary(rules)　#察看求得的关联规则之摘要

inspect(rules)

```
> inspect(rules)
   lhs                          rhs                    support confidence     lift
1  {curd,
    yogurt}                  => {whole milk}       0.01006609  0.5823529 2.279125
2  {other vegetables,
    butter}                  => {whole milk}       0.01148958  0.5736041 2.244885
3  {other vegetables,
    domestic eggs}           => {whole milk}       0.01230300  0.5525114 2.162336
4  {yogurt,
    whipped/sour cream}      => {whole milk}       0.01087951  0.5245098 2.052747
5  {other vegetables,
    whipped/sour cream}      => {whole milk}       0.01464159  0.5070423 1.984385
6  {pip fruit,
    other vegetables}        => {whole milk}       0.01352313  0.5175097 2.025351
7  {citrus fruit,
    root vegetables}         => {other vegetables} 0.01037112  0.5862069 3.029608
8  {tropical fruit,
    root vegetables}         => {other vegetables} 0.01230300  0.5845411 3.020999
9  {tropical fruit,
```

用 R 进行购物篮分析

- 按需要筛选关联规则

```
x=subset(rules,subset=rhs%in%"whole milk"&lift>=1.2)    #求所需要的关联规则子集
inspect(sort(x,by="support")[1:5])    #根据支持度对求得的关联规则子集排序并察看
```

其中 lift=P(L,R)/(P(L)P(R)) 是一个类似相关系数的指标。lift=1时表示L和R独立。这个数越大，越表明L和R存在在一个购物篮中不是偶然现象。

购物篮分析的应用

- 超市里的货架摆设设计
- 电子商务网站的交叉推荐销售

2012.6.12

购物篮分析的应用

货到付款的客户注意： 如果有客户想选择货到付款的，拍下商品后，请一定要与旺旺客服联系。如果拍下来，不与旺旺客服联系，在48小时内我们会关闭交易。

浏览了该宝贝的会员还浏览了

754之顶级经典 全新 华硕 K8V-mx 带集成显卡 带AGP

￥116.00

478 865PE 豪华大板 GA-8IPE1000-G 集声和

￥108.00

英特尔915G D915GAG 全集 128显卡 千兆网卡

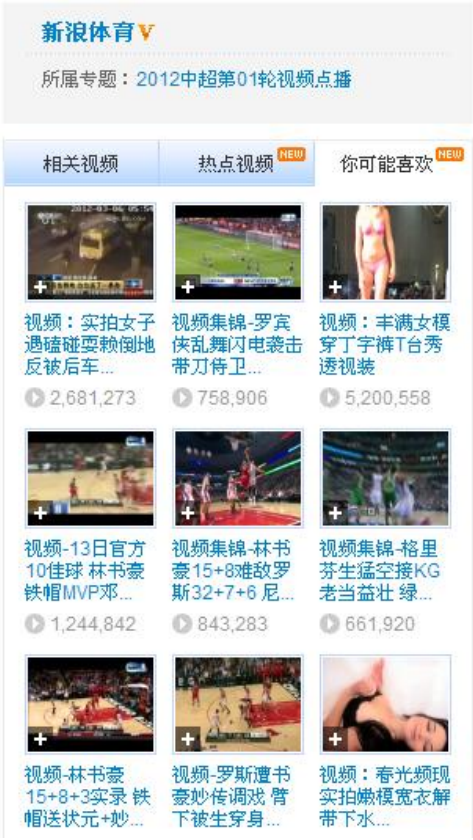
￥85.00

特价75元 技嘉845GV GA-8I845GV 板载声显网卡 支

￥100.00

购物篮分析的应用

- 网站或节目的阅读/收听推荐



2012.6.12

Thanks

FAQ时间