数据挖掘中简单高效的变量约简技术

金仁浩，
北京物资学院信息学院

数据挖掘主要内容：

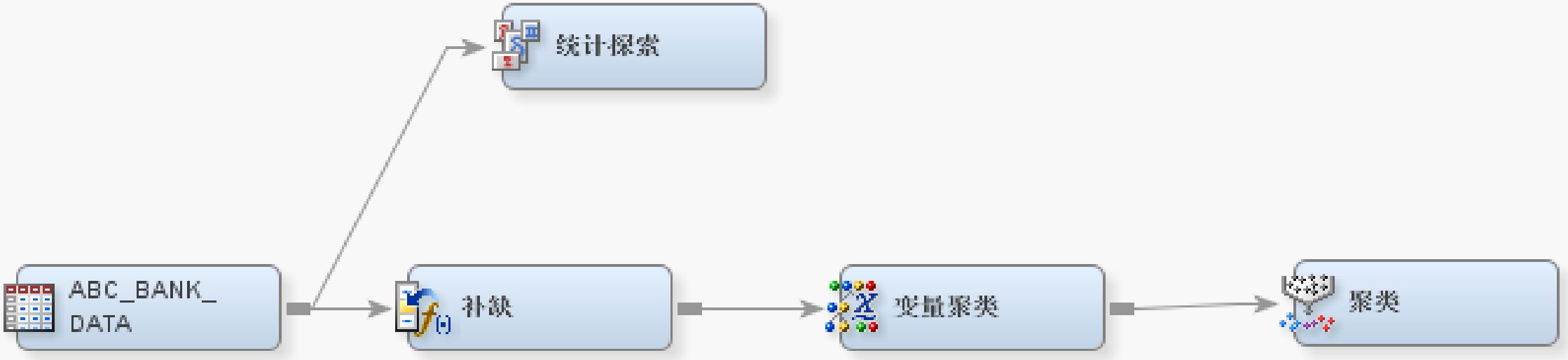
一. 关联规则和链接分析；

二. 基于聚类分析的各种应用；

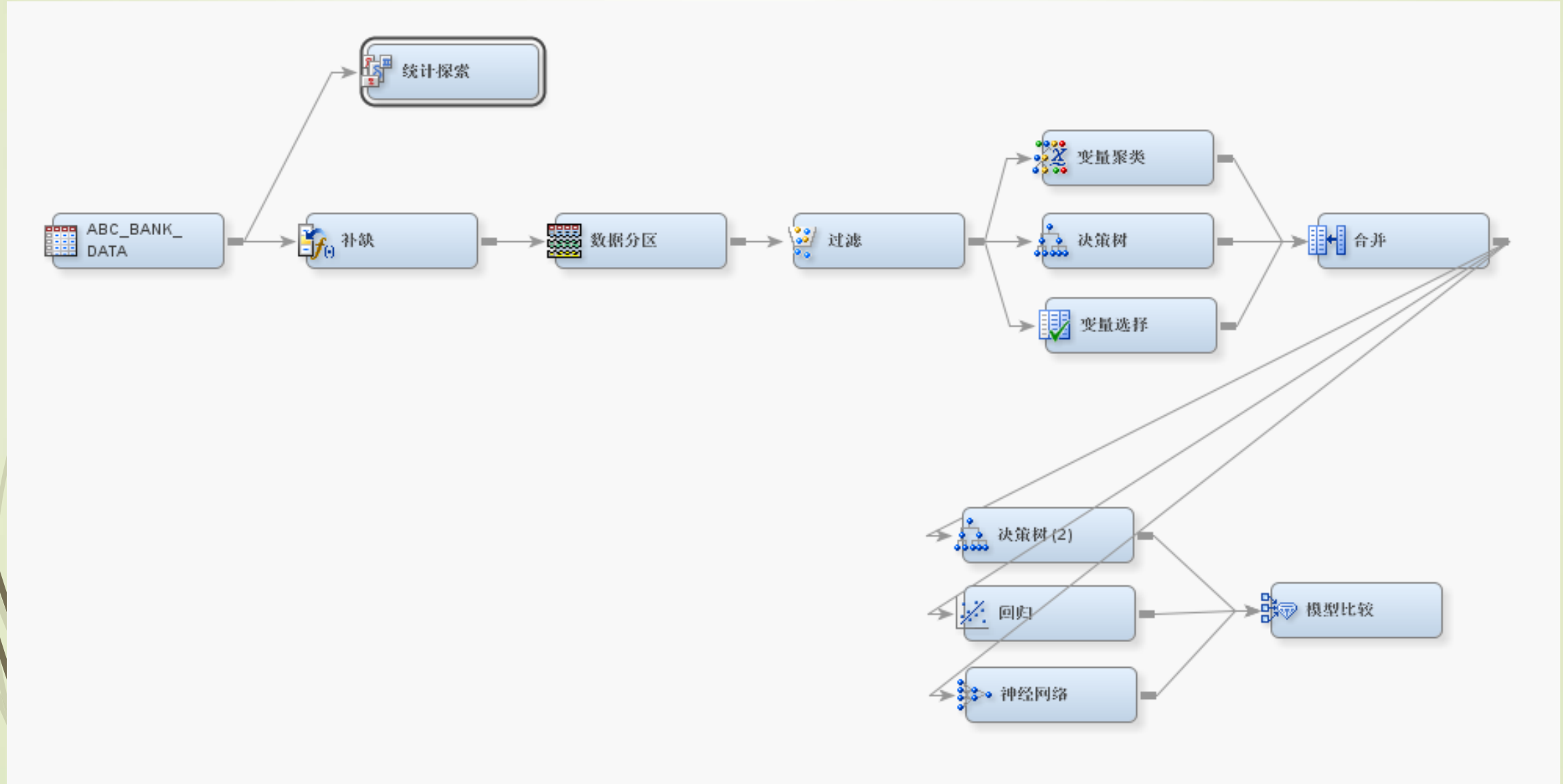
三. 预测模型---决策树，回归，神经网络；

四. 统计各个分支的其他应用.

一般数据聚类步骤：



一般预测模型的建模步骤：



报告内容：

一. 变量太多会导致的问题；

二. 基于主成分的变量聚类技术；

三. 经典统计学的变量选择技术；

四. 基于决策树技术的变量选择.

一. 变量太多会导致的问题

大数据的最重要的特点之一就是数据维度高(变量数多)。在有的行业(银行业)用于建模的变量数可能达到2000个甚至更多。

变量太多即是福音又是灾难：

1. 大量的变量通常意味着大量可用的描述性信息，大量的可用来构建更好的模型的信息。

2. 由于各种原因变量太多也会带来一系列的问题。

变量太多导致的问题：

1. 存在过多变量的数据集,往往会有稀疏性问题. 稀疏性意味着,众多的客户签名中,大多数取值为0或其他一些特定的值.

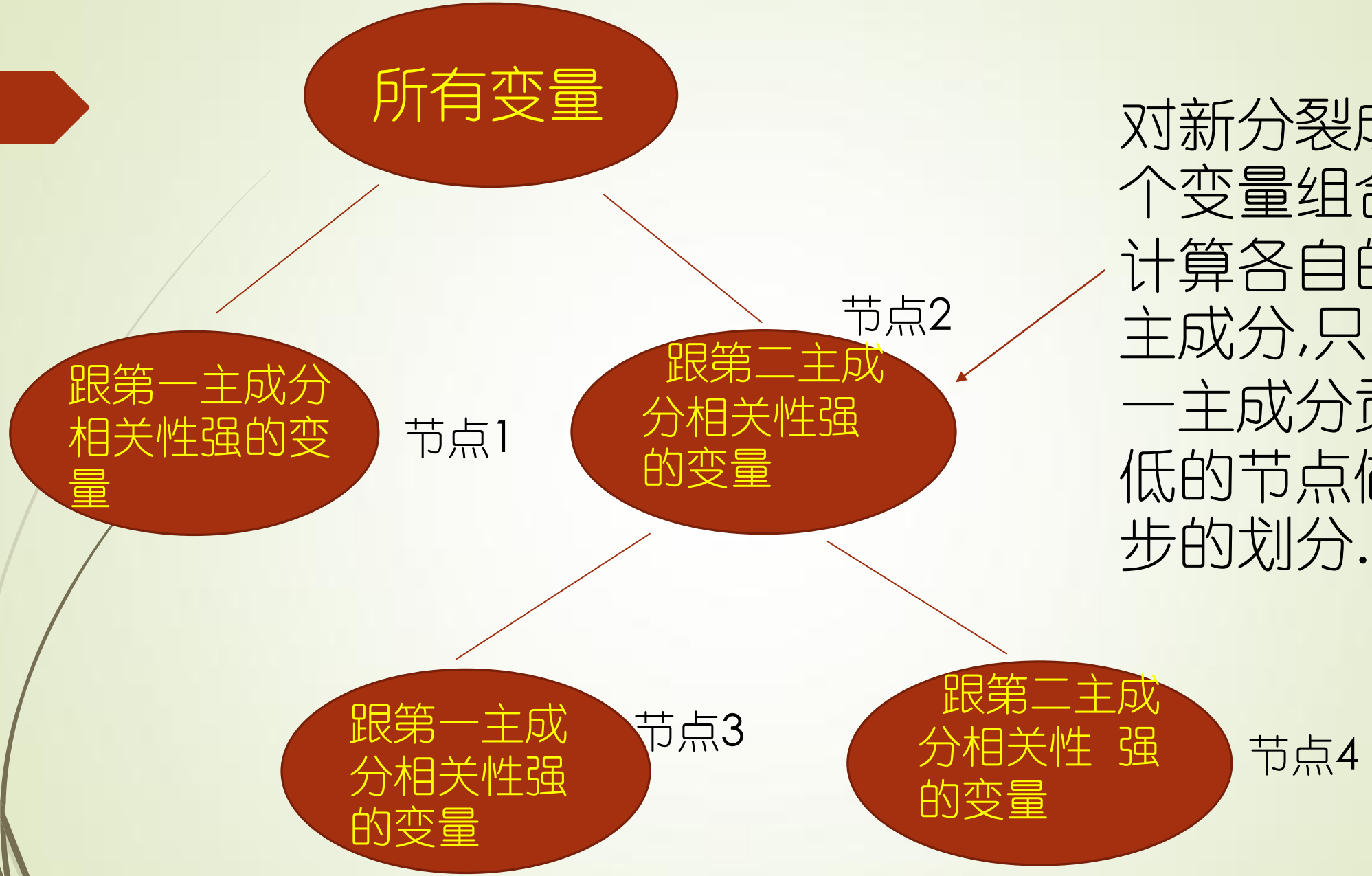
2. 输入变量之间彼此相关性,会导致聚类算法中某些特征的权重过大; 众多变量彼此相关会造成决策树的结果难以理解;也会导致回归模型难以发现真正有显著性影响的变量;

3. 变量太多也会带来相关数据挖掘模型过拟合,以及算法收敛时间较长等问题.

二. 基于主成分的变量聚类技术

- 变量聚类技术跟我们通常所说的聚类分析不是同一个概念. 变量聚类技术是对输入变量的聚类, 而一般意义上的聚类分析是对观测值(客户签名)的聚类.

- 变量聚类技术是无指导的数据挖掘技术, 在计算过程中不涉及到目标变量. 它首先根据变量之间的相关性对变量集进行分群, 然后对每个小分群寻找一个代表. 这个代表可以是小分群中的某一个变量, 也可以是小分群中变量的线性组合(主成分).



对新分裂成的两个变量组合,分别计算各自的第一主成分,只选择第一主成分贡献率低的节点做下一步的划分.

变量聚类技术算法：

- 1. 把数据集中所有的输入变量作为一个初始的大群.
- 2. 由上一群中的所有输入变量产生出第一主成分 C_1 和第二主成分 C_2. 主成分是一群中所有变量的线性组合.
- 3. 在满足相关群分裂标准条件下,根据两主成分 C_1 和 C_2 把大变量群分成 **2** 个小群:

If $R^2(X_i, C_1) > R^2(X_i, C_2)$, then $X_i \epsilon \{C_1\}$ else $X_i \epsilon \{C_2\}$, where R is correlation coefficient.

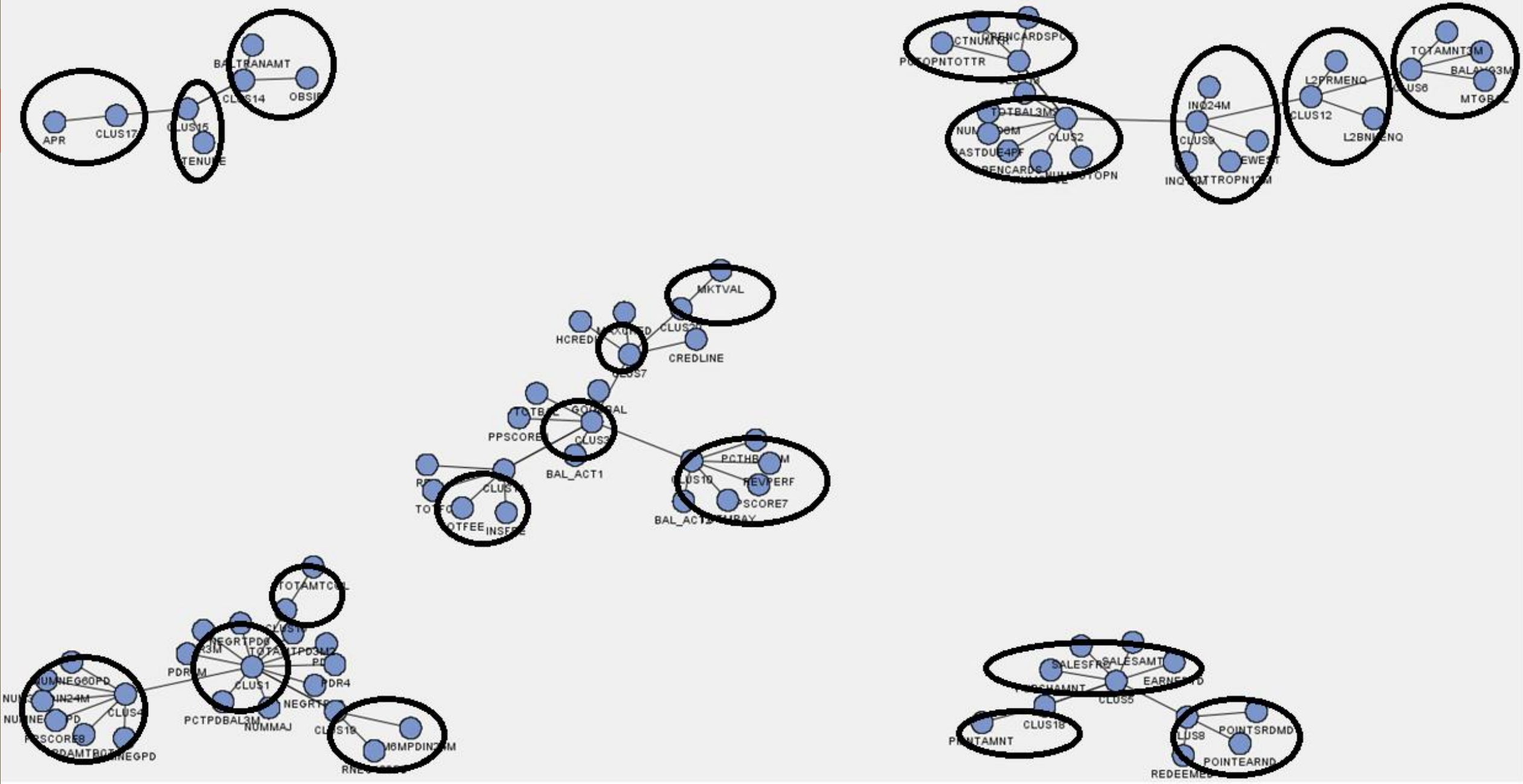
- 4. 根据相关条件,只选出其中**1**个小群做进一步划分。划分的过程重复**2-3**步.
- 5. 当满足相关终止条件,停止群得划分。

实例应用：

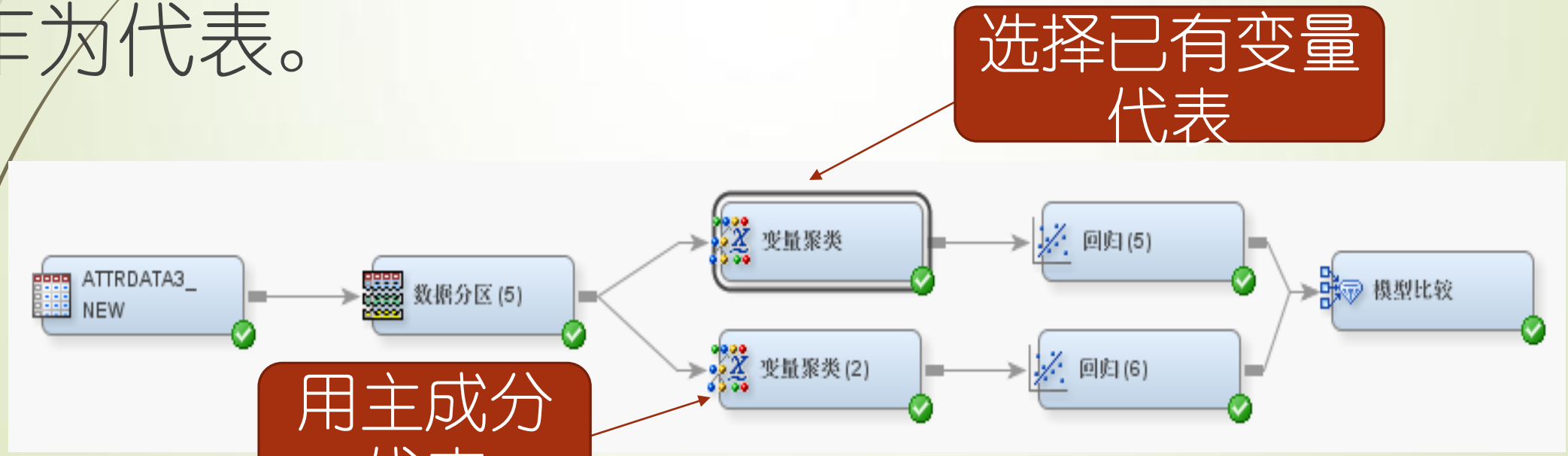
- 以一个银行客户数据为例,里面有15000个客户观测值,85个列变量(1个是目标变量,65区间型输入变量,19个分类变量). 如果只对其中65个区间型输入变量使用变量聚类技术,聚出20个子群,如前图所示. 变量缩减的效果非常明显。

Tips:

- 在对观测值进行聚类前,如果输入变量太多,可以用变量聚类技术先对输入变量进行缩减.往往会取得比较好得结果,尤其是当很多输入变量都彼此相关时.


APR
CLUS17
CLUS15
CLUS14
MKTVAL
CREDLINE
CLUS7
CLUS3
BAL_ACT1
CLUS10
REVTPERF
CLUS1
PDR4
CLUS4
CLUS18
CLUS8
POINTSRDMD
POINTEARND

在对每一子群选择变量代表时,这个代表可以是子群的**主成分**，也可以是子群中的与子群主成分**相关性最高的变量**。在模型评估阶段，如果两种选法相差不大，优先考虑选择子群中的某一变量作为代表。



Tips

1.在构建关联规则时,为了提高规则的支持度和置信度,常会使用大类产品, i.e. 把各种口味和品牌的冰淇淋聚为冰淇淋.关联规则里的变量合并是基于变量属性的,跟我们这里的变量聚类不是一回事.

2.在用聚类方法进行客户特征分析时,如变量太多也可以仅仅只在几个重要的变量上进行聚类.聚完类后再分析各类中其他未参加聚类的变量的特征.思想:有共性的客户在重要变量上常常也会相似.

3. K-均值聚类与层次化聚类互补效果会很好：K-均值聚类扩展性好，其本质是围绕质心的大圆，会使得聚得类较大。层次化聚类扩展性差，但结果比较有意义。两者结合，可以首先使用K-均值建立很多小类(譬如100个)，然后对这些群集(小类)再使用层次性聚类以建立群集的层次化结构。最后构成的较大群集可以是任意形状的。

三. 经典统计学的变量选择技术;

- **当目标变量是连续型变量：**

Step 1：分别单独计算每个输入变量与目标变量的简单相关系数 R, 首先剔除 $R^2 < 0.005$的那些输入变量;

Step 2：对Step 1中选出的变量，用向前回归法选择变量。

- **实例：** 在一个研究存款利率提高对银行顾客存款额度变化的例子中, 有277个输入变量. Step 1, 筛选出92个输入变量进入 Step 2; 而 Step 2 只选出5个输入变量。

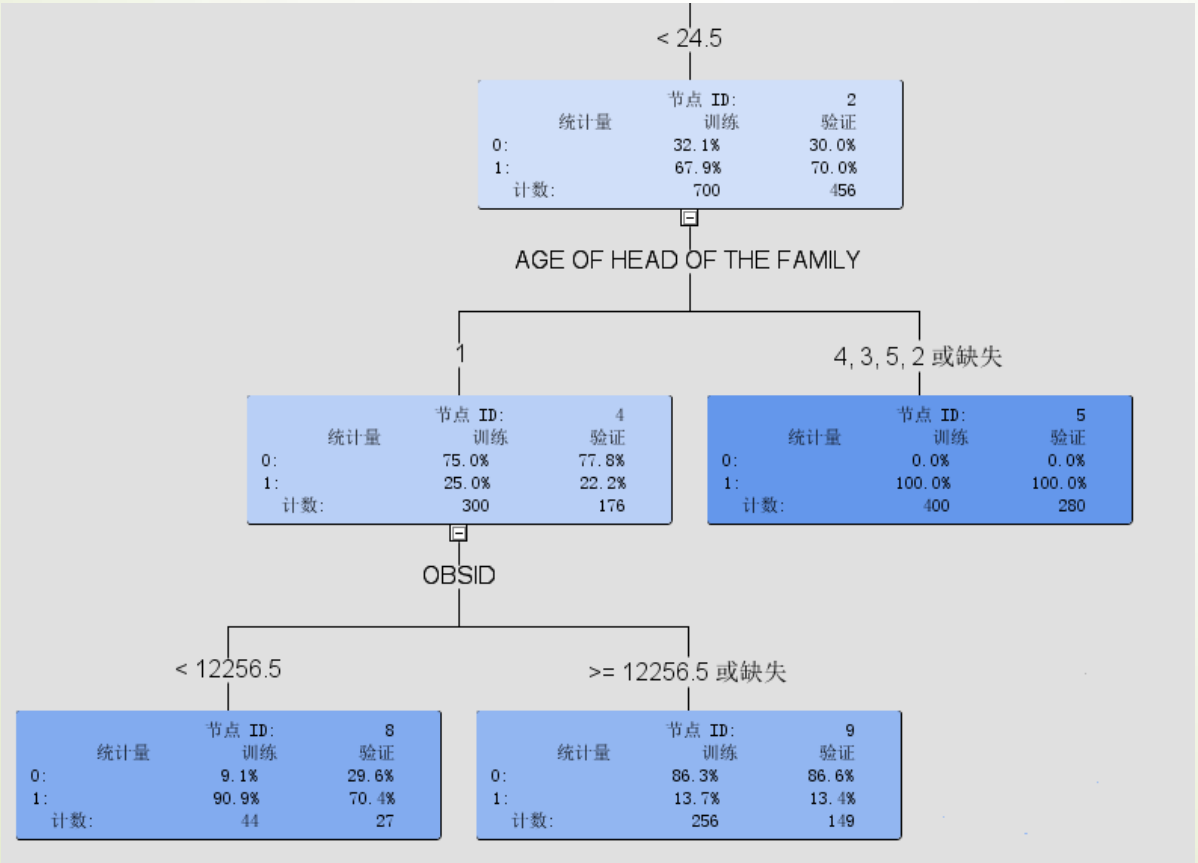
- 目标变量是二元目标变量(0或1)：

仍按照上面的两步法来选择变量,由于此时目标变量是二元的, Step 1里一般换为卡方检验, 且在Step 2里需要改为logistic 回归.

实例： 在研究影响爱尔兰牛群结核病感染的案例中, 二元目标变量是牛群中有没有发现有感染结核病的牛,有51个输入变量. Step 1通过Spearman's相关系数, 筛选出15个输入变量进入Step 2; 而 Step 2 只选出4个输入变量.

四. 基于决策树技术的变量选择

- 在有目标变量的时候,决策树技术可以根据目标变量的取值把观测值集合分成很多小组(叶结点). 对决策树节点划分起作用的变量可以作为变量选择的结果,进入到下一步的模型建立中.



分类目标变量和连续目标变量情形都可以用决策树模型来进来变量选择。

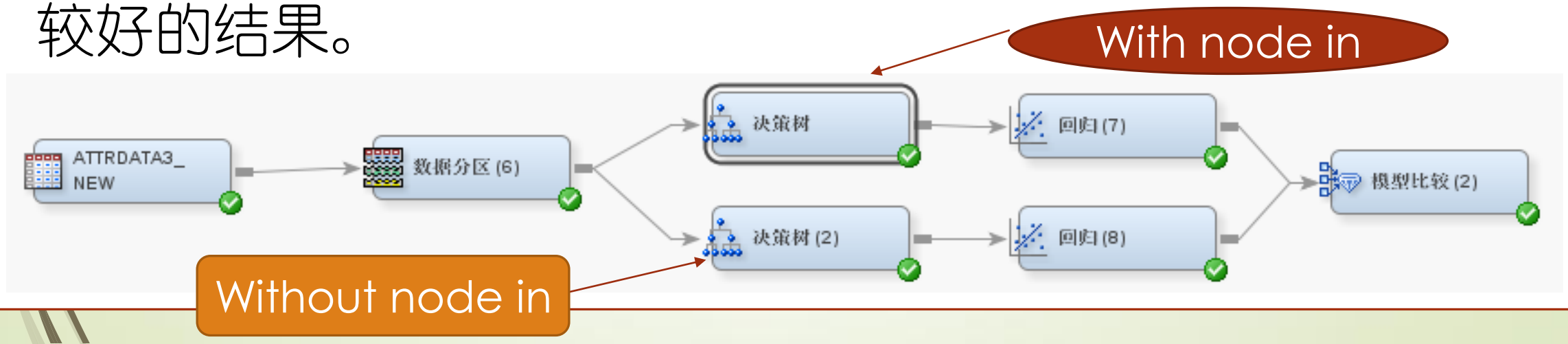
Tips：

用来估计连续型目标变量的决策树一般称为回归树。一般不适用回归树这样的离散模型来估计连续值，因为回归树的取值就是有限的叶节点值。但是回归树模型是可以用来选择变量的.

Tips:

- 1. 当变量太多时，可以先通过经典统计的技术先删除一些不相关变量，然后再利用决策树技术做进一步的变量筛选。

- 2. 在建立决策树的时，对每个观测值附带会产生一个新的分类变量 (叶节点变量,该变量值表示观测值被分配到哪个叶节点)。把叶节点变量带入到下一步的模型建立，往往会取得比较好的结果。



总结

- 1.方法是死的,人是活得,恰当使用方法常会带来很好的效果;
- 2. 在聚类分析的时候，为了避免输入变量之间相关性, 会给聚类结果带来不好的影响, 可以采用基于主成分技术的变量聚类技术和只在关键变量上聚类来解决. 当然也可以只对部分变量采用变量聚类技术;
- 3. 在有目标变量建模时,可以通过变量聚类技术,基于经典统计的变量选择法和决策树技术方法进行变量约简. 为了防止某种选择方法误剔除了重要变量,可以降低剔变量的系数阀值,以及把几种方法选出的变量放在一起交由下一步的预测模型来做进一步的选择 (一般神经网络模型没法做变量选择).


THANK YOU FOR YOUR ATTENTION
FINALLY OVER!
quickmeme.com

金仁浩，应用数理统计博士

兴趣：空间统计学，应用数据分析，数据挖掘算法的灵活运用

手机/微信：18612490817

邮箱：18612490817@163.com