# 第三阶段 从数据分析到数据挖掘

# SPSS Modeler-[数据分析与挖掘3]

## 网校相关视频

第二章 SPSS Modeler模型篇

## 明确本次课知识点，明确重点难点

### 【知识点目标】

* SPSS Modeler 数据挖掘--关联：寻找事务之间的联系
* SPSS Modeler 数据挖掘--聚类：对数据进行划分

## 本次课程任务讲解

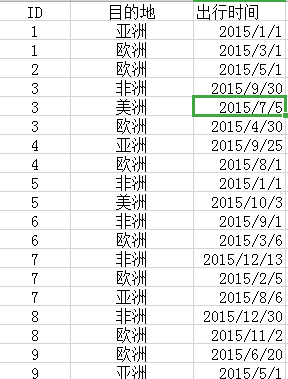
### 【知识点1】SPSS Modeler 数据挖掘--关联

#### 关联分析的基本概念

关联分析又称关联挖掘，就是在交易数据、关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构。常用于购物篮分析、旅游线路分析、社交媒体分析、推荐系统等

事务表（行穷表）

每行记录一条具体信息/完整事务（数据表）



真值表

每行记录一个完整事务，但把每一个项目都单独作为哑变量，取值为0/1(或T/F),表示是否有过该行为



事务表和真值表的转换

通过数据透视/交叉表形式，可以实现从事务表到真值表的转化

逆透视可以从真值表转化为事务表

1. 关联规则的有效性指标

前项：先购买/进行的事务

后项：后购买/进行的事务

支持度

X-Y的支持度为S=N(X,Y)/N,即同时出现X,Y的事务数/总事务数

置信度

X-Y的置信度为C=N(X,Y)/N(X),即同时出现X,Y的事务数/出现X的事务数（前项）

同一个组合，若前后项不同，得到的置信度可能不一样

为生成合适的关联规则，通常需同时设置最小支持度和最小置信度

提升度

提升度=规则置信度/后项支持度，反映相比与总体，后项受前项的影响程度，>1为正向，<1为负向

部署能力

部署能力=前项支持度-规则支持度，反映了有多少已经购买前件但还没有购买后件的客户比例

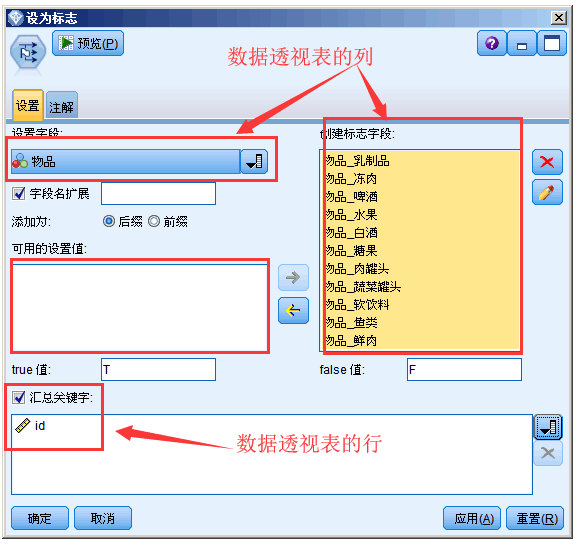
1. Apriori算法实践

算法原理

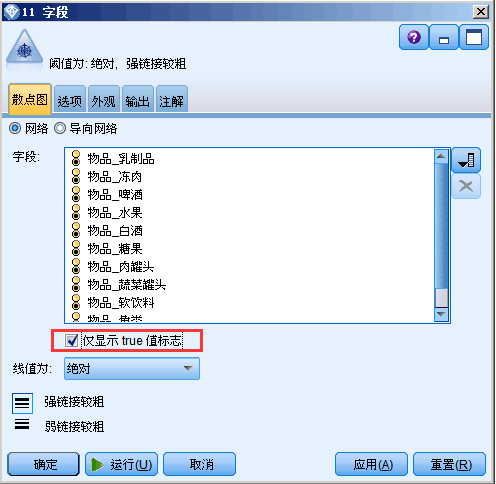
根据最小支持度和置信度生成频繁集

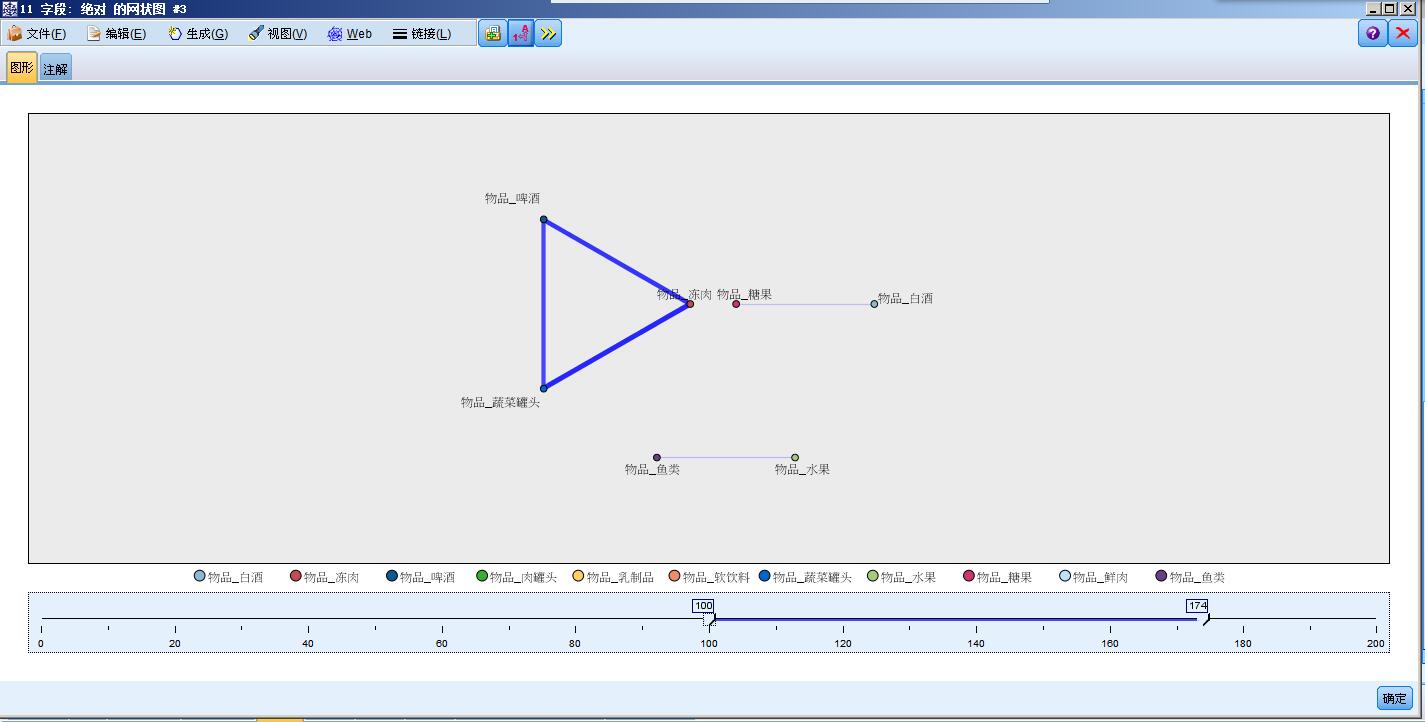
操作步骤

将事务表转为真值表（设为标志节点）

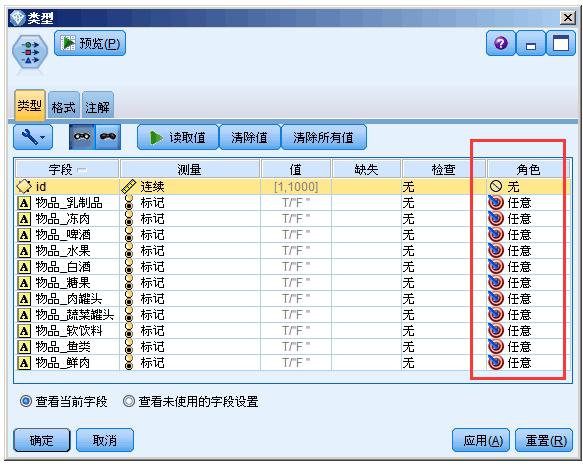


通过网络图观察关联关系

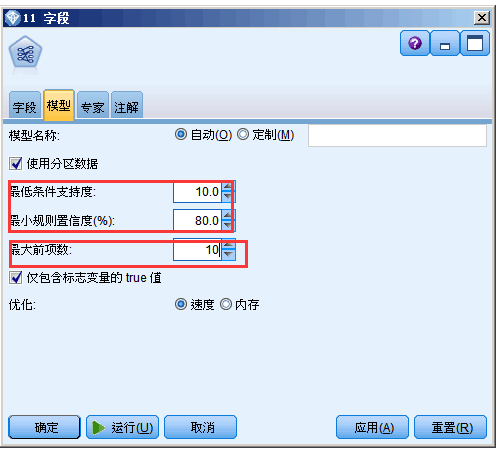




设置角色，注意在Apriori算法中，所有的项（前、后项）既是输入也是目标



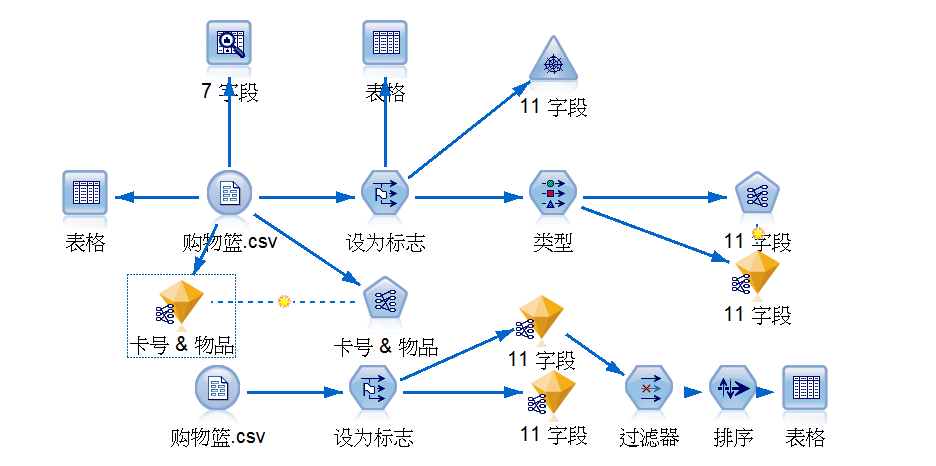
模型参数设置



结果解读



部署



### 【知识点2】SPSS Modeler 数据挖掘--聚类

#### 聚类算法回顾

分组/分簇

距离

1. 基本思路
2. Means：

指定聚类数目K

初始化K个聚类的中心（随机）

计算距离，将样本划分到不同群

重新计算聚类中心（群所有样本的均值）

判定是否达到停止条件

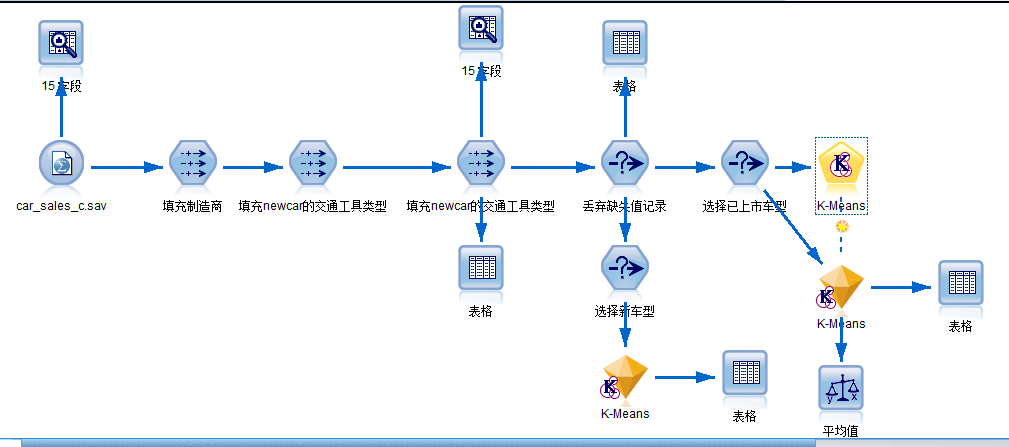
K-Means的注意问题：

针对连续变量

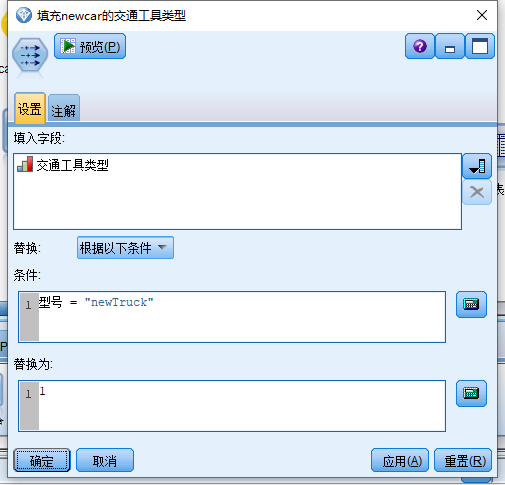
连续变量标准化（归一化，0~1）

分类变量转换为连续变量（哑变量）标准化

1. 实例



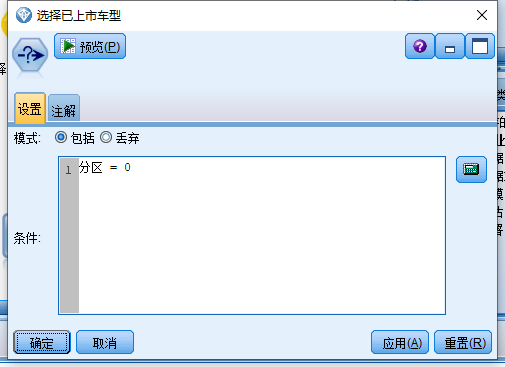
1、整理数据——填充制造商、交通工具类型



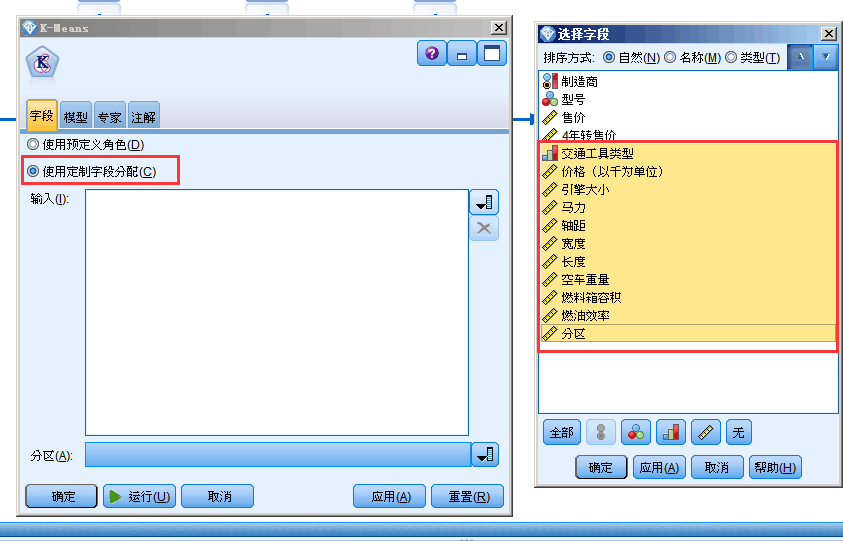
2、丢弃（除了售价和四年转售价）有缺失值的记录



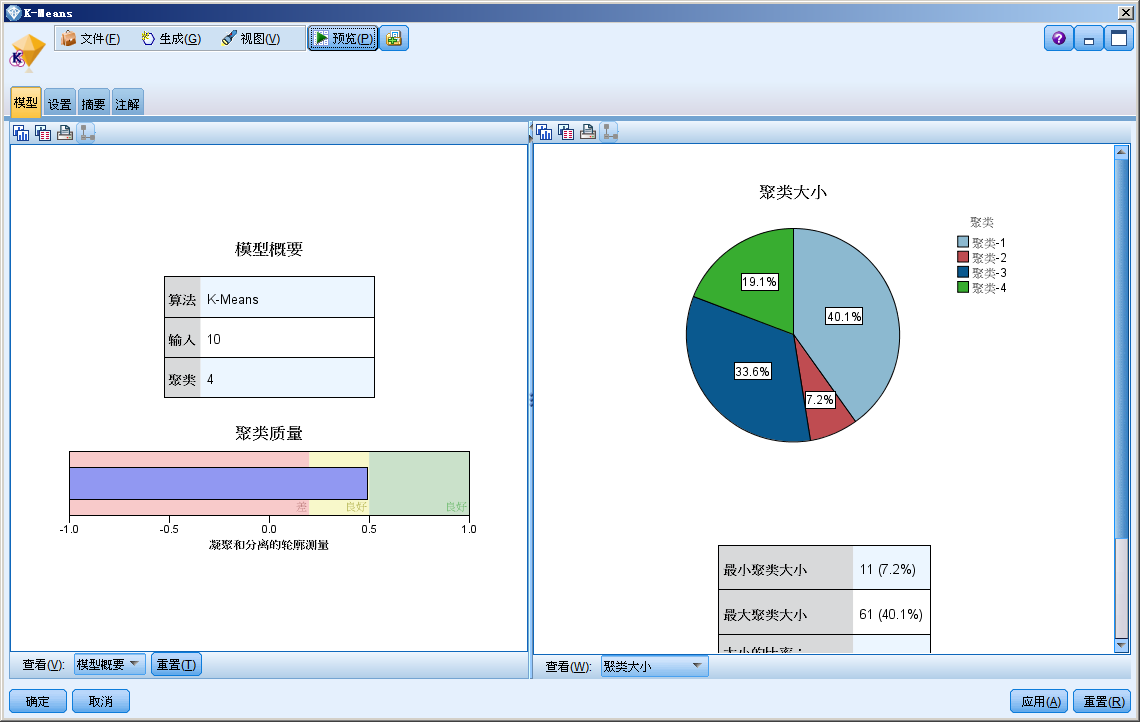
3、选择已上市的车型进行聚类分析

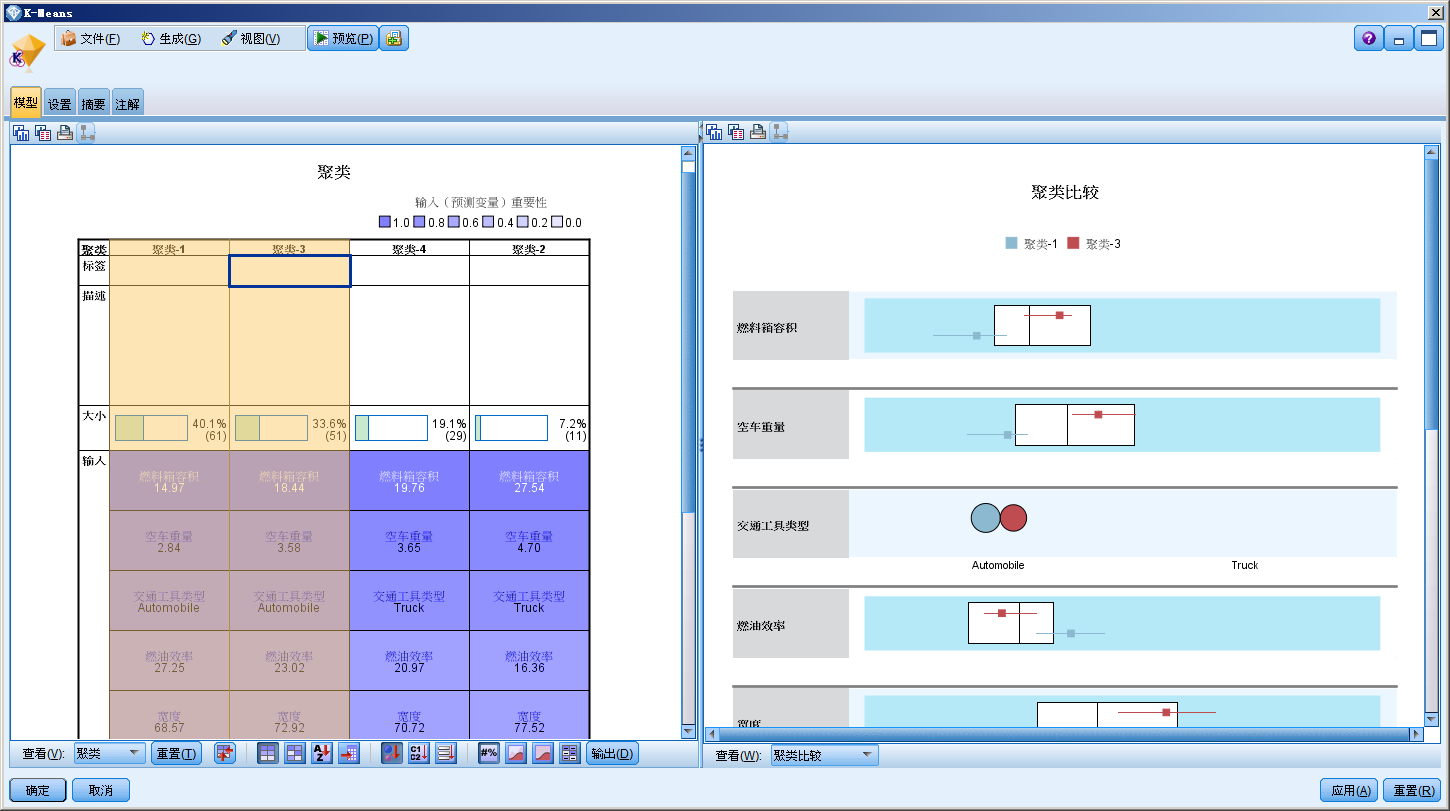


4、模型设置



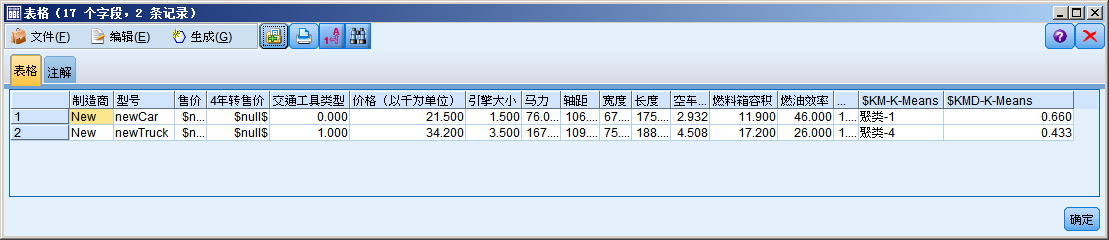
结果解读





两两比较，聚类3比聚类1的值大，聚类3属于高级轿车，聚类1属于低端轿车，聚类2比聚类4的值大，聚类2属于高级卡车，聚类4属于低端卡车

预测（产品定位）



## 自主学习作业讲解

无

## 课程总结

* SPSS Modeler 数据挖掘--关联
* 关联分析的基本概念
* 事务表
* 真值表
* 关联分析的有效性指标
* 前项
* 后项
* 支持度
* 置信度
* Apriori算法实践
* 原理
* 操作步骤
* 结果解读
* 部署
* SPSS Modeler 数据挖掘--聚类
* 聚类算法回顾
* 聚类算法基本思路
* K-MEANS基本思路
* K-MEANS注意事项
* 聚类算法实际操作
* 模型设置
* 结果解读
* 预测分析

## 下次自主学习任务布置

#### 观看预习视频

#### 课后作业