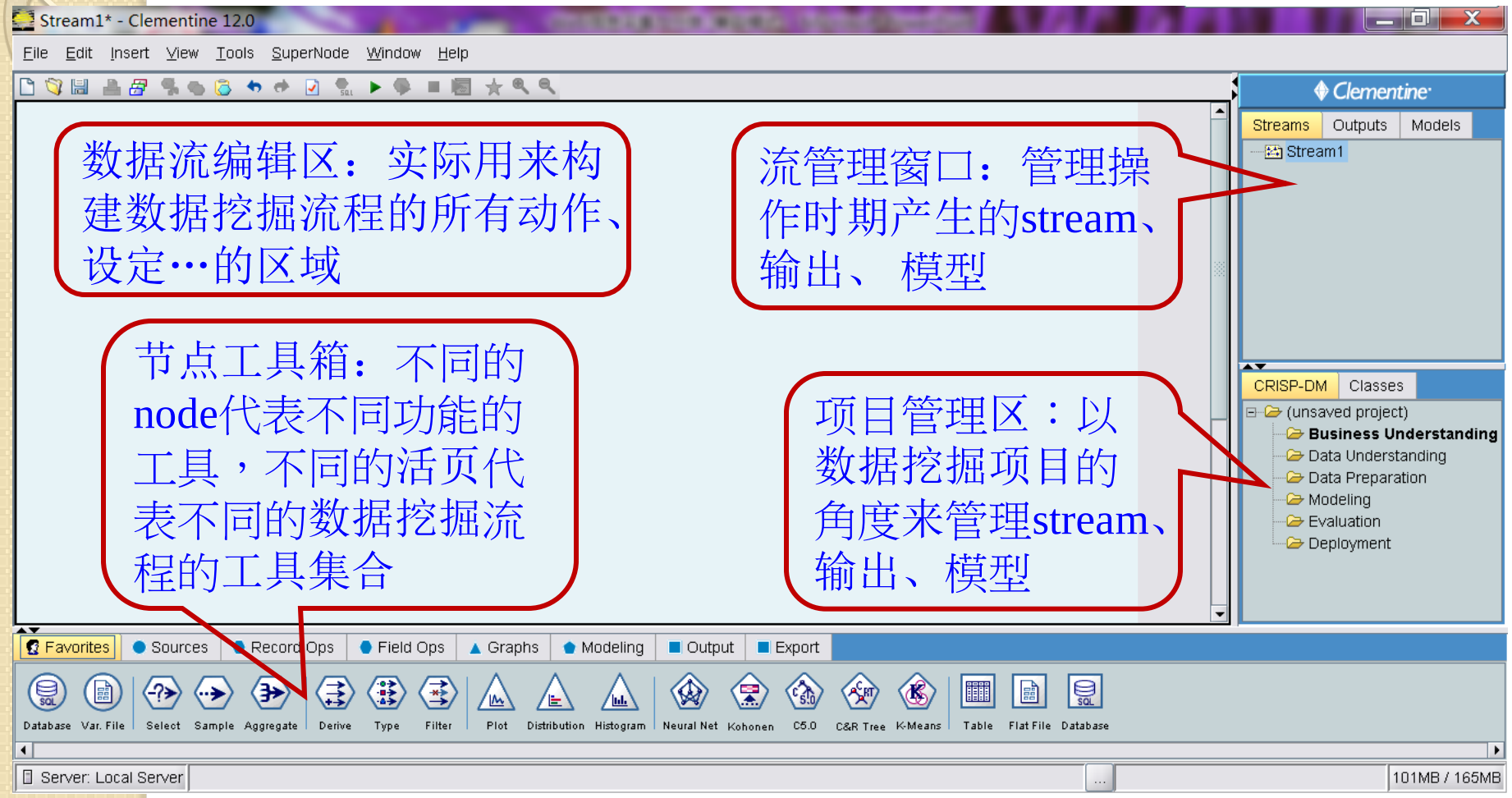
Artificial Intelligence
——工具与应用

Yanghui Rao
Assistant Prof., Ph.D
School of Data and Computer Science,
Sun Yat-sen University
raoyangh@mail.sysu.edu.cn

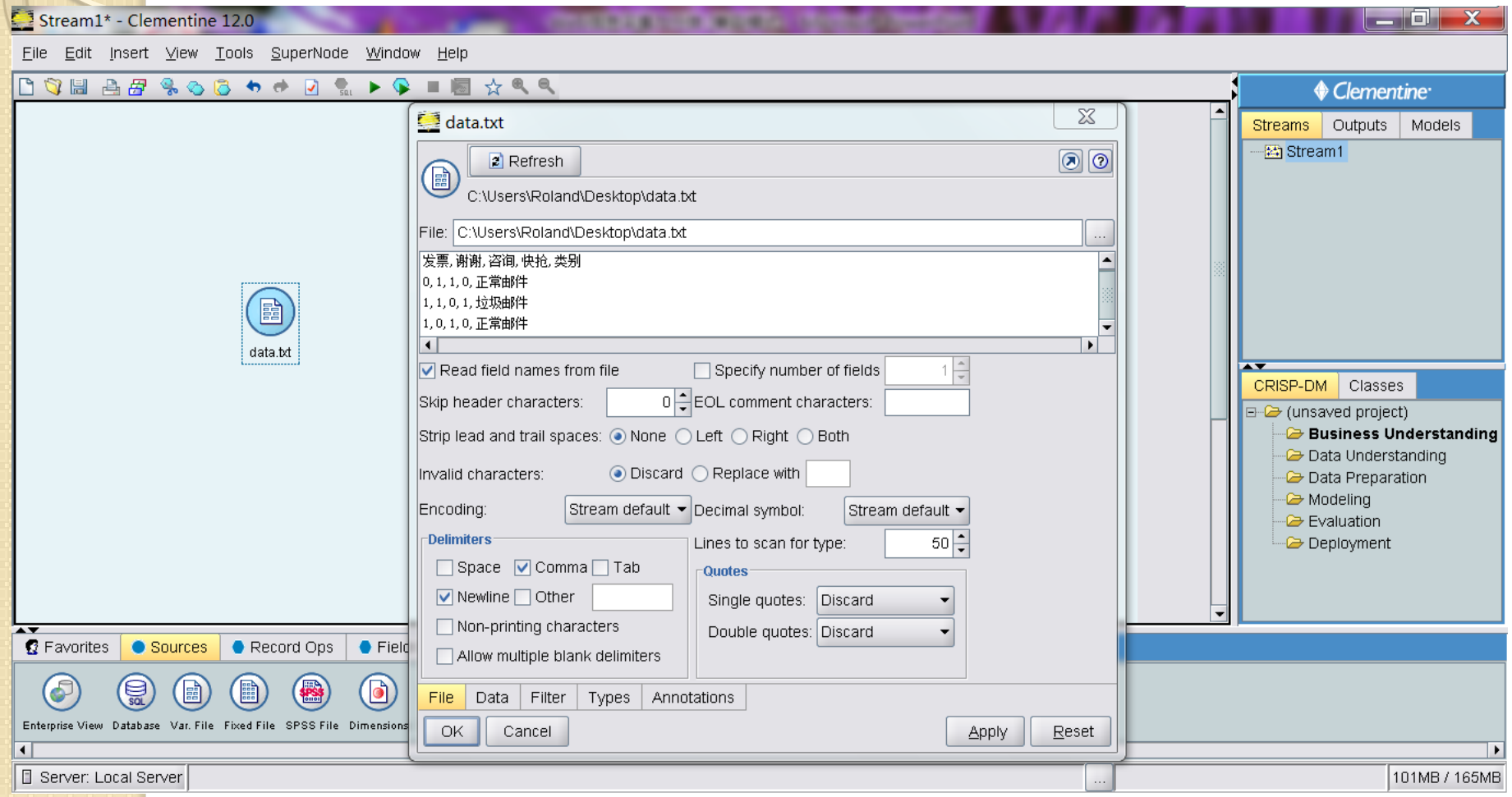
SPSS Clementine

- Clementine作为一个受欢迎的数据挖掘平台，充分利用了计算机系统的运算能力和图形展示能力，快速有效的实现决策树分类或回归、关联规则挖掘、聚类等模型。
- Clementine操作步骤为：导入数据集—>数据集预处理—>建模—>评估模型。
- https://en.wikipedia.org/wiki/SPSS_Modeler

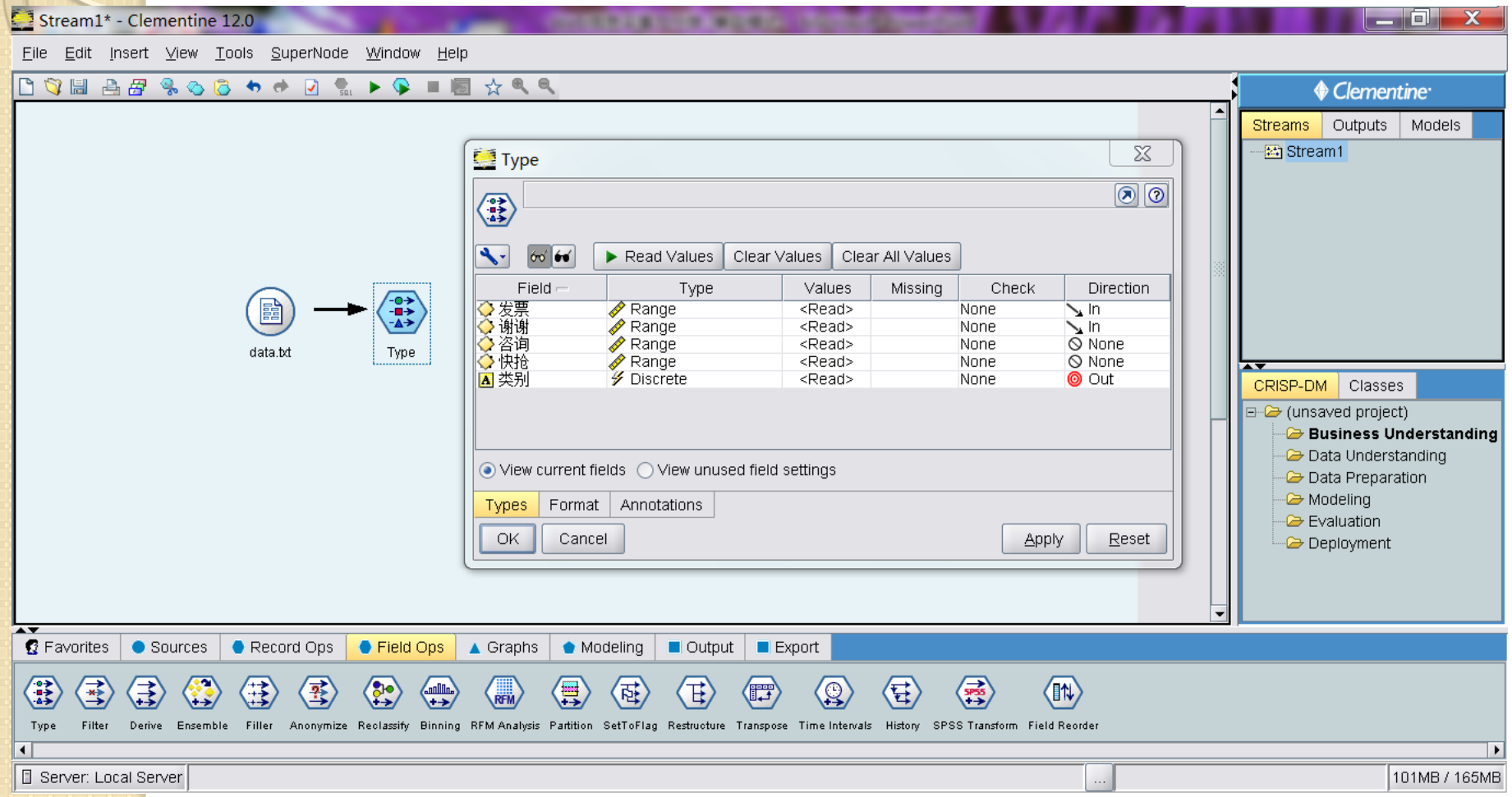
SPSS Clementine



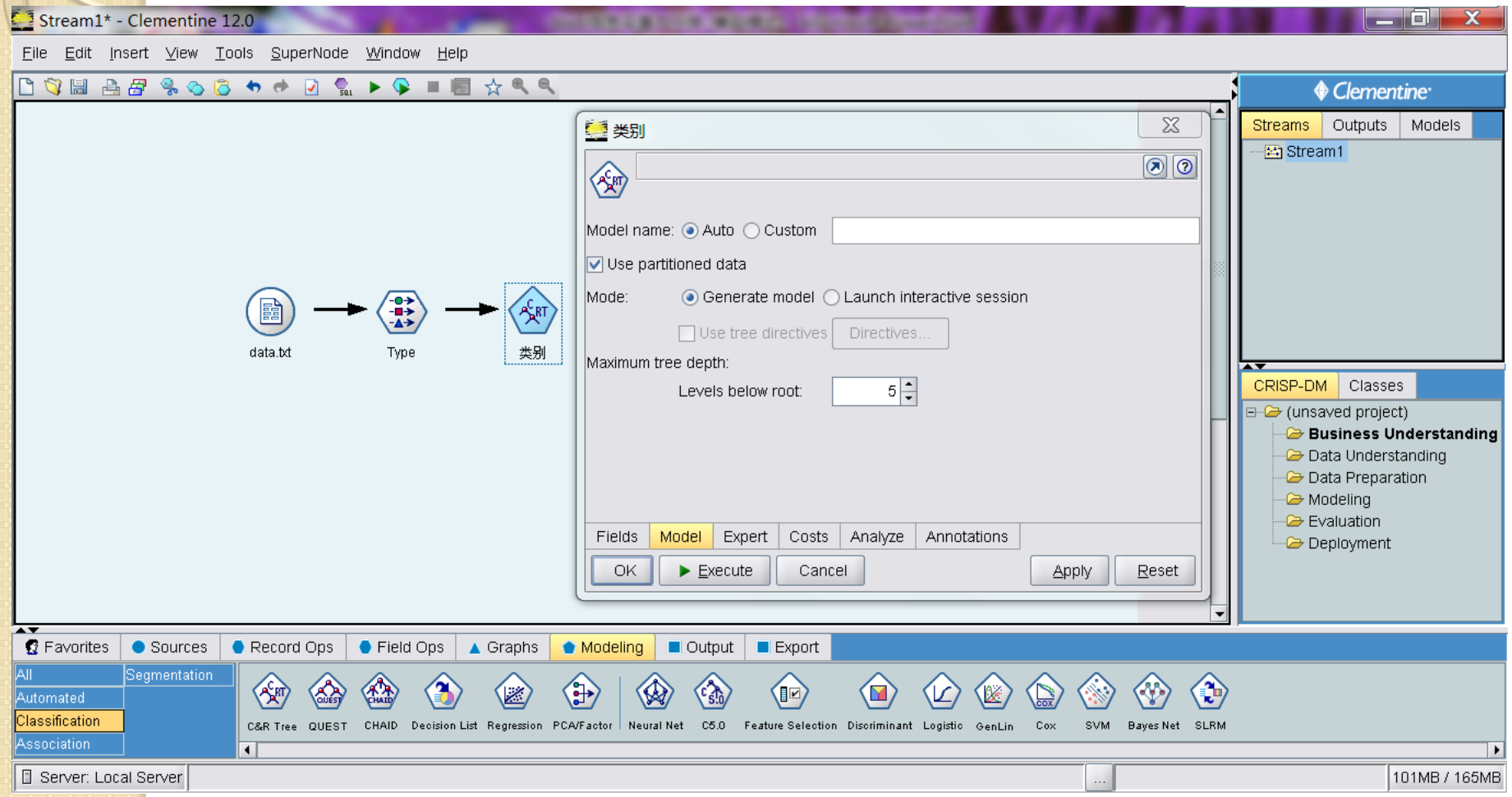
Step 1: Input Data



Step 2: Set Features



Step 3: Select Models



Example: Decision Trees

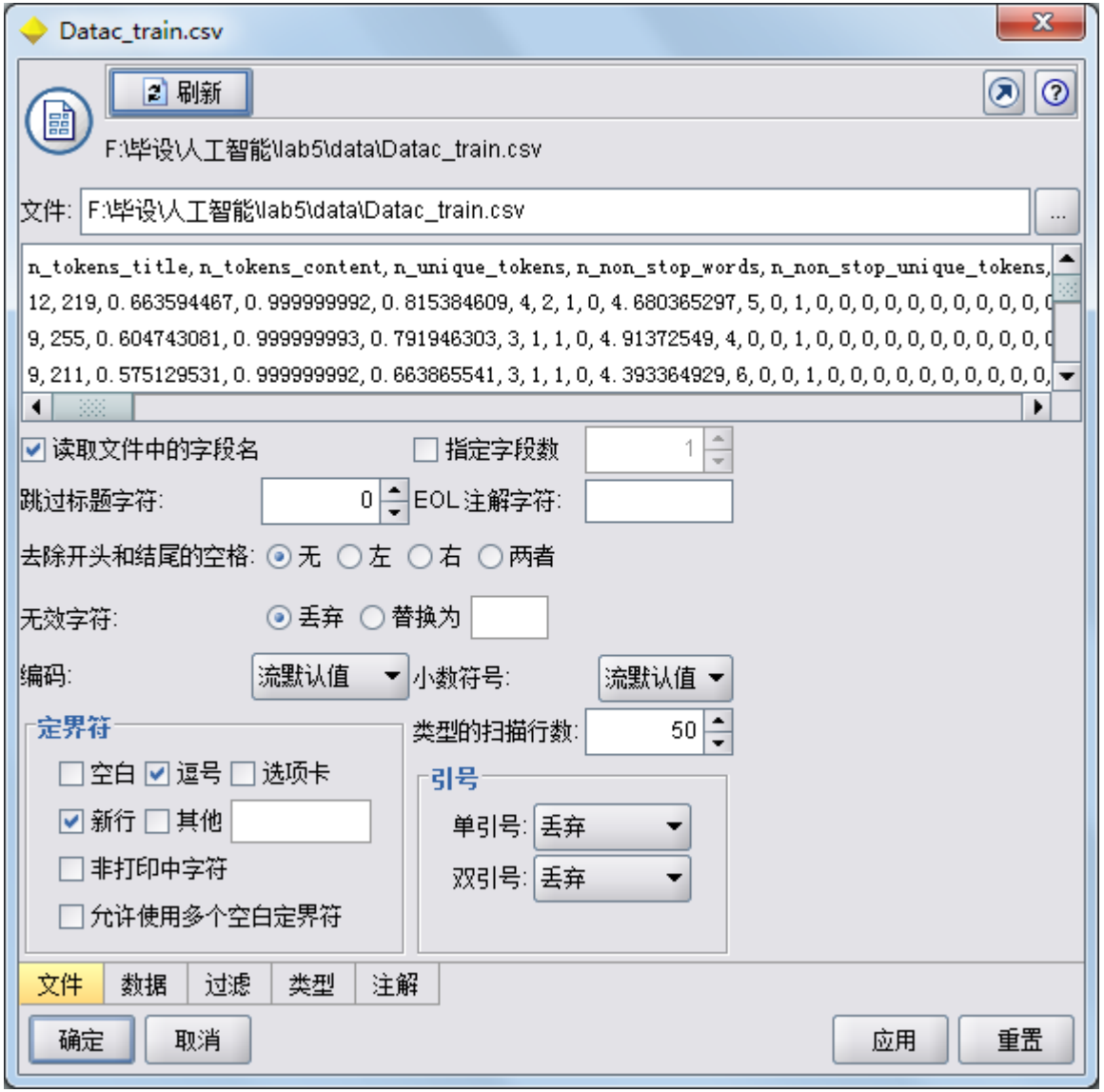
- 新闻分享
  - 给定一篇新闻的58个属性（特征），比如新闻标题中词的个数（n_tokens_title）等等，预测这篇新闻在社交网络中被分享的次数（shares）。被分享的次数越多，表示这篇新闻越"热门/受欢迎"。
  - shares："1"代表popular（即该篇新闻为"热门/受欢迎"的），"0"代表unpopular。

global_ra	rate_pos	rate_nega	avg_posi	min_posi	max_posi	avg_nega	min_nega	max_nega	title_su	title_se	abs_title	abs_title	shares
0.016667	0.827957	0.172043	0.402039	0.1	1	-0.22448	-0.5	-0.05	0	0	0.5	0	0
0.015167	0.846939	0.153061	0.42772	0.1	1	-0.24278	-0.5	-0.05	1	0.5	0.5	0.5	0
0.020619	0.6	0.4	0.566667	0.4	0.8	-0.125	-0.125	-0.125	0.125	0	0.375	0	1
0.030303	0.5625	0.4375	0.298413	0.1	0.5	-0.2381	-0.5	-0.1	0	0	0.5	0	0
0.020833	0.648649	0.351351	0.40448	0.1	1	-0.41506	-1	-0.1	0	0	0.5	0	1
0.010695	0.714286	0.285714	0.435	0.2	0.7	-0.2625	-0.4	-0.125	0	0	0.5	0	1
0.029197	0.636364	0.363636	0.37551	0.2	0.7	-0.31042	-0.6	-0.05	1	-1	0.5	1	0
0.052632	0.347826	0.652174	0.4575	0.16	1	-0.33789	-0.7	-0.1	1	-1	0.5	1	1
0.011583	0.571429	0.428571	0.249091	0.136364	0.5	-0.13869	-0.1875	-0.05	0.75	0.55	0.25	0.55	0

Step 1

- 输入训练集

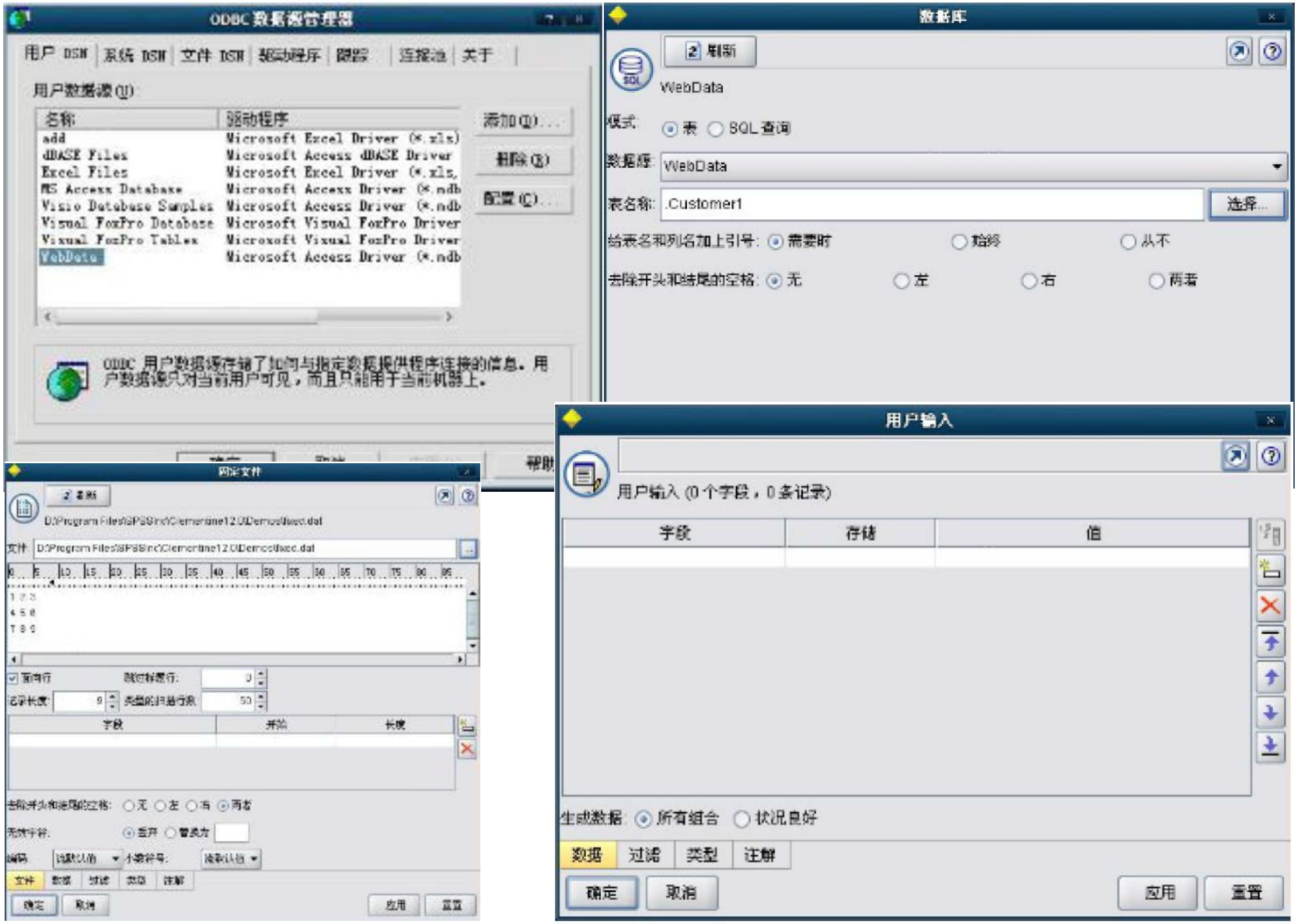

可变文件


Datac_train.csv
刷新
F:\毕设\人工智能\lab5\data\Datac_train.csv
文件: F:\毕设\人工智能\lab5\data\Datac_train.csv
n_tokens_title, n_tokens_content, n_unique_tokens, n_non_stop_words, n_non_stop_unique_tokens,
12, 219, 0.663594467, 0.999999992, 0.815384609, 4, 2, 1, 0, 4.680365297, 5, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
9, 255, 0.604743081, 0.999999993, 0.791946303, 3, 1, 1, 0, 4.91372549, 4, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
9, 211, 0.575129531, 0.999999992, 0.663865541, 3, 1, 1, 0, 4.393364929, 6, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
读取文件中的字段名
指定字段数
1
跳过标题字符:
0
EOL 注解字符:
去除开头和结尾的空格:
无
左
右
两者
无效字符:
丢弃
替换为
编码:
流默认值
小数符号:
流默认值
定界符
空白
逗号
选项卡
新行
其他
非打印中字符
允许使用多个空白定界符
类型的扫描行数:
50
引号
单引号:
丢弃
双引号:
丢弃
文件
数据
过滤
类型
注解
确定
取消
应用
重置

源节点说明

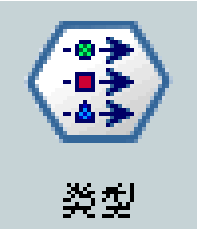
- 可变文件：用于导入逗号等分隔了的ASCII数据。
- 固定文件：用于导入固定字段（字段未被分隔，但是始于相同的位置，并有固定长度）的ASCII数据。
- Excel：用于导入Excel电子表格。
- SPSS文件：用于导入SPSS文件。
- SAS文件：用于导入SAS格式的文件。
- 数据库：用于通过ODBC导入数据。
- 用户输入：用于代替已存在的来源节点，也可通过在已存在节点上点击鼠标右键的方式使用该节点。

源节点说明


ODBC 数据源管理器
用户 DSN
系统 DSN
文件 DSN
驱动程序
跟踪
连接池
关于
用户数据源(U):
名称
驱动程序
add
Microsoft Excel Driver (*.xls)
dBASE Files
Microsoft Access dBASE Driver
Excel Files
Microsoft Excel Driver (*.xls,
MS Access Database
Microsoft Access Driver (*.mdb
Visio Database Samples
Microsoft Access Driver (*.mdb
Visual FoxPro Database
Microsoft Visual FoxPro Driver
Visual FoxPro Tables
Microsoft Visual FoxPro Driver
WebData
Microsoft Access Driver (*.mdb
添加(D)...
删除(R)
配置(C)...
ODBC 用户数据源存储了如何与指定数据提供程序连接的信息。用户数据源只对当前用户可见，而且只能用于当前机器上。
帮助
数据库
刷新
WebData
模式:
表
SQL 查询
数据源:
WebData
表名称:
.Customer1
选择...
给表名和列名加上引号:
需要时
始终
从不
去除开头和结尾的空格:
无
左
右
两者
固定文件
刷新
文件:
字段
开始
长度
确定
取消
应用
重置
用户输入
用户输入 (0 个字段，0 条记录)
字段
存储
值
生成数据:
所有组合
状况良好
数据
过滤
类型
注解
确定
取消
应用
重置

Step 2

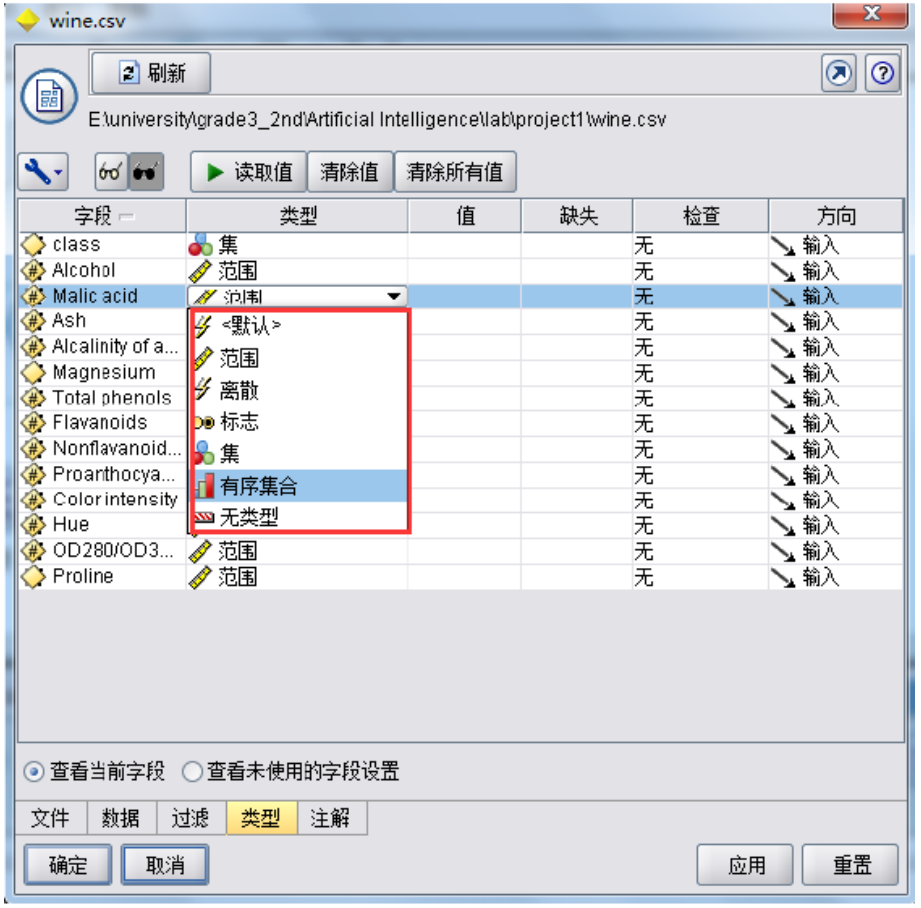
- 属性设置



字段	类型	值	缺失	检查	方向
n_tokens_title	范围	[2,19]		无	输入
n_tokens_c...	范围	[0,8474]		无	输入
n_unique_to...	范围	[0.0,0.999...		无	输入
n_non_stop...	范围	[0.0,1.0]		无	输入
n_non_stop...	范围	[0.0,0.999...		无	输入
num_hrefs	范围	[0,187]		无	输入
num_self_hr...	范围	[0,74]		无	输入
num_imgs	范围	[0,128]		无	输入
num_videos	范围	[0,91]		无	输入
average_tok...	范围	[0.0,7.695...		无	输入

属性类型

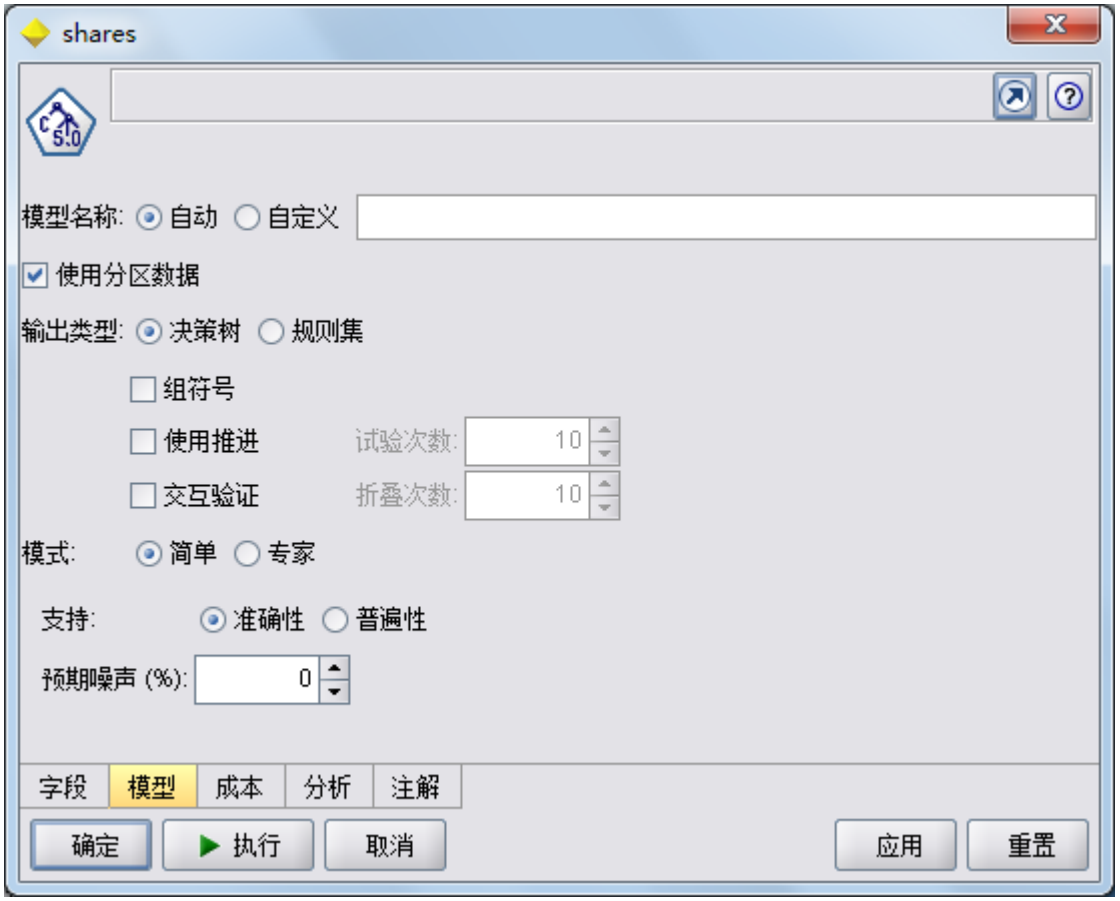
- 范围型：如年龄
- 离散型：如职业
- 标志型：如性别
- 集合型：如日期
- 有序集型：如学历
- 缺省型：明确的变量类型
- 无类型：不属于上述类型



Step 3

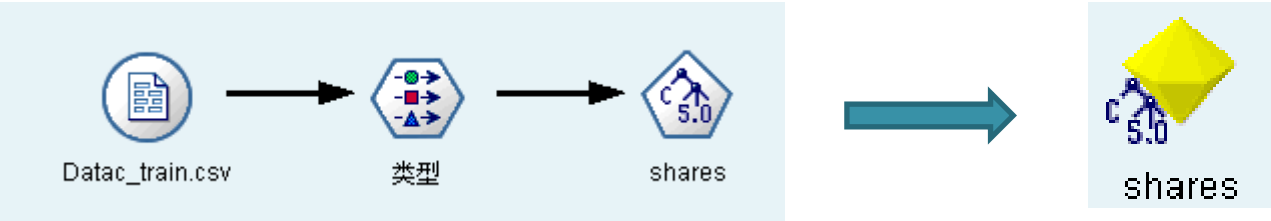
- 选择模型

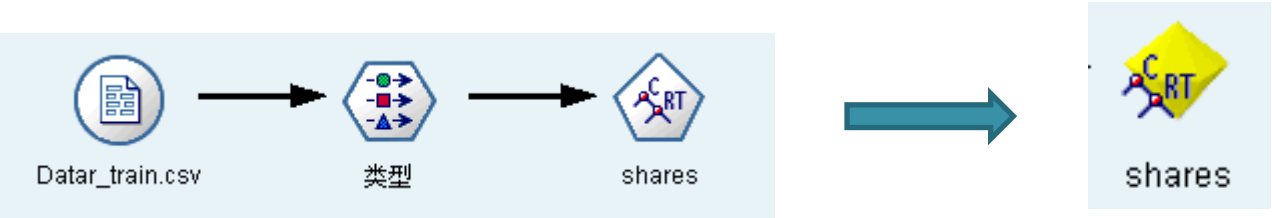
C5.0是C4.5应用于大数据集上的分类算法，主要在执行效率和内存使用方面进行了改进



Step 4

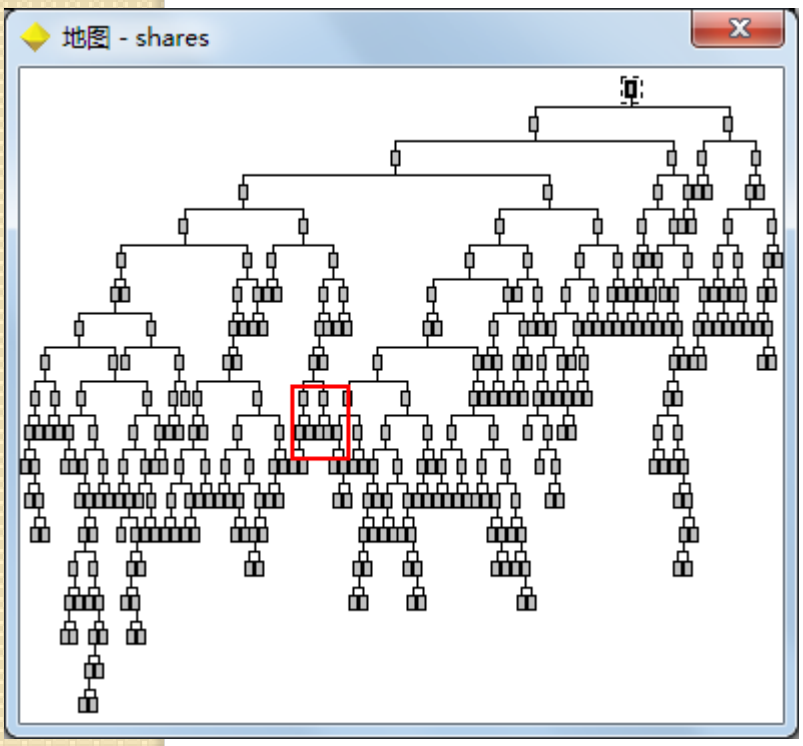
- 训练模型

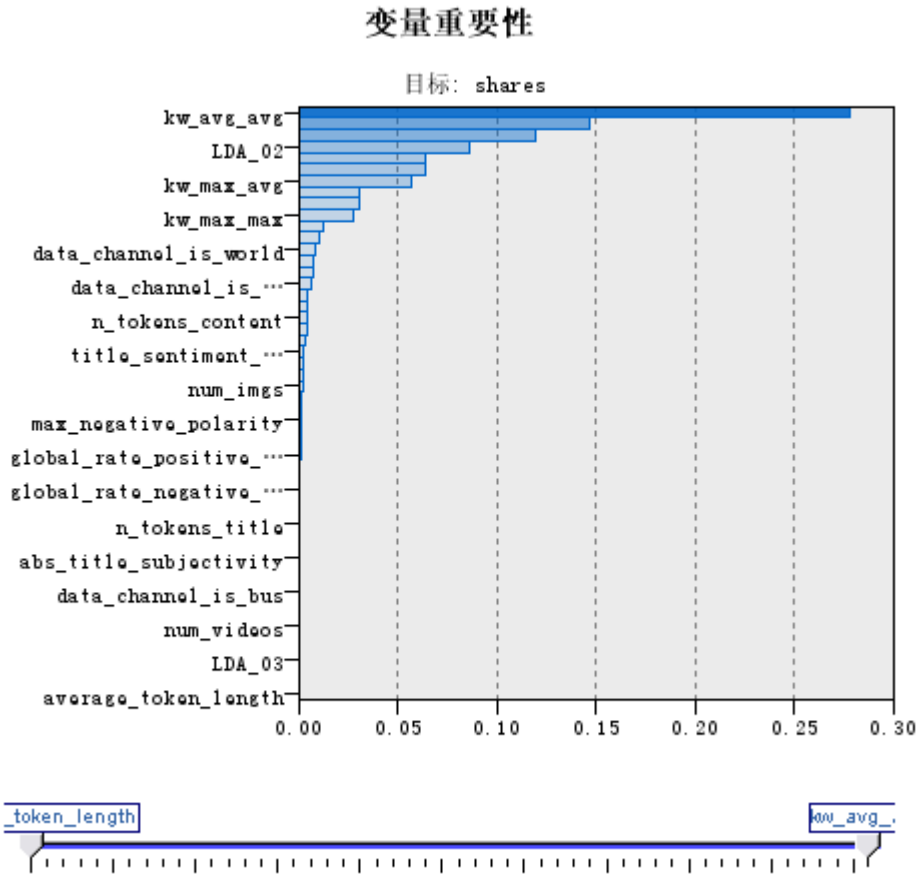

Datac_train.csv
类型
shares
shares


Datar_train.csv
类型
shares
shares

Step 5

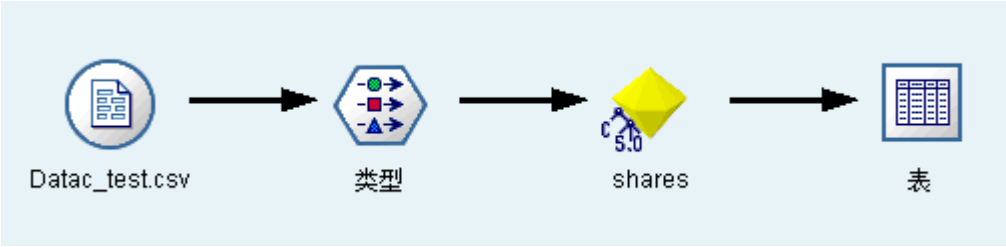
- 模型训练后的结果

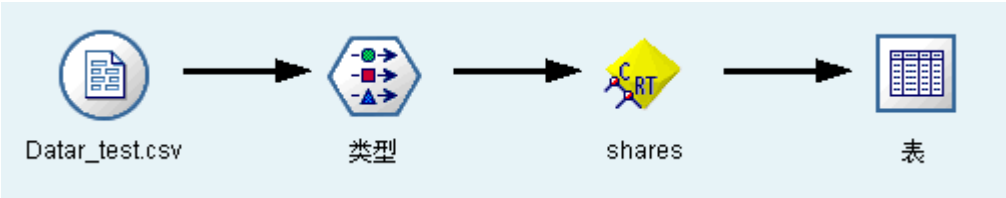




Step 6

- 将训练好的模型用于预测
  - 输入测试集





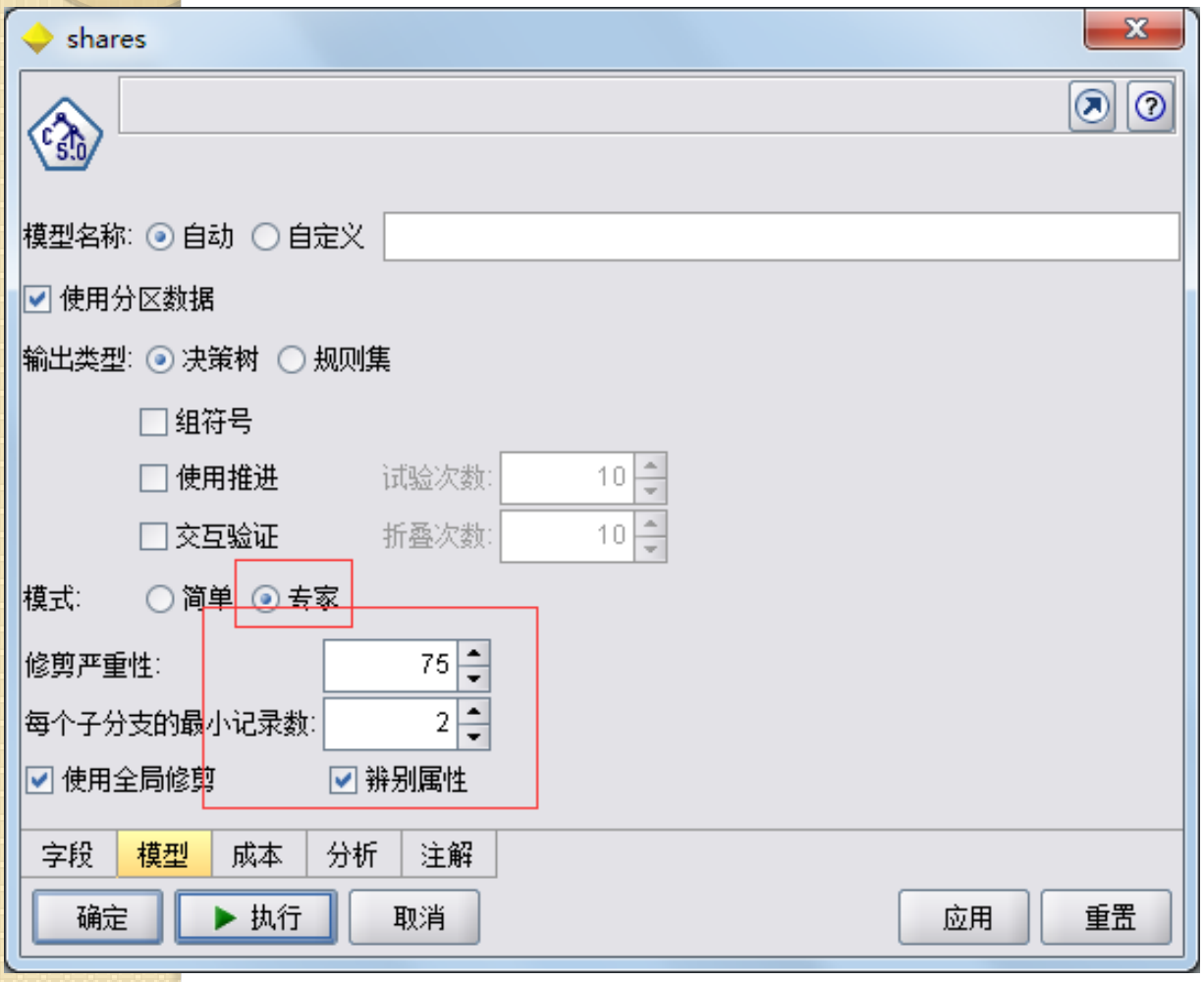
Step 7

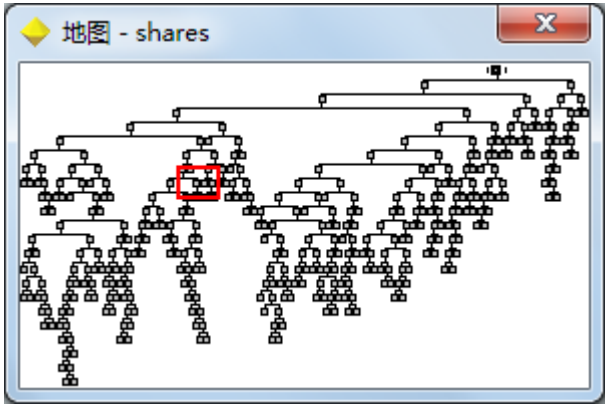
- 预测结果

y	shares	$C-shares	$CC-shares
0	0	0	0.711
0	0	0	0.711
0	0	0	0.860
0	0	0	0.660
0	0	0	0.833
0	1	1	0.541
0	1	1	0.525
0	0	1	0.630
0	0	0	0.685
0	0	1	0.630
0	1	0	0.633
0	1	0	0.732
0	1	1	0.742
0	0	0	0.773
7	0	0	0.577
0	1	1	0.635
0	0	0	0.833
0	1	1	0.555
0	1	1	0.635
0	0	0	0.833
0	1	1	0.635
0	0	1	0.742
0	0	0	0.833
0	0	0	0.714
0	0	0	0.711
0	1	1	0.635
0	0	1	0.630
7	1	0	0.733
6	0	1	0.630
0	1	1	0.880
0	1	1	0.742
0	1	1	0.635
0	0	0	0.525
0	1	0	0.733

Step 8

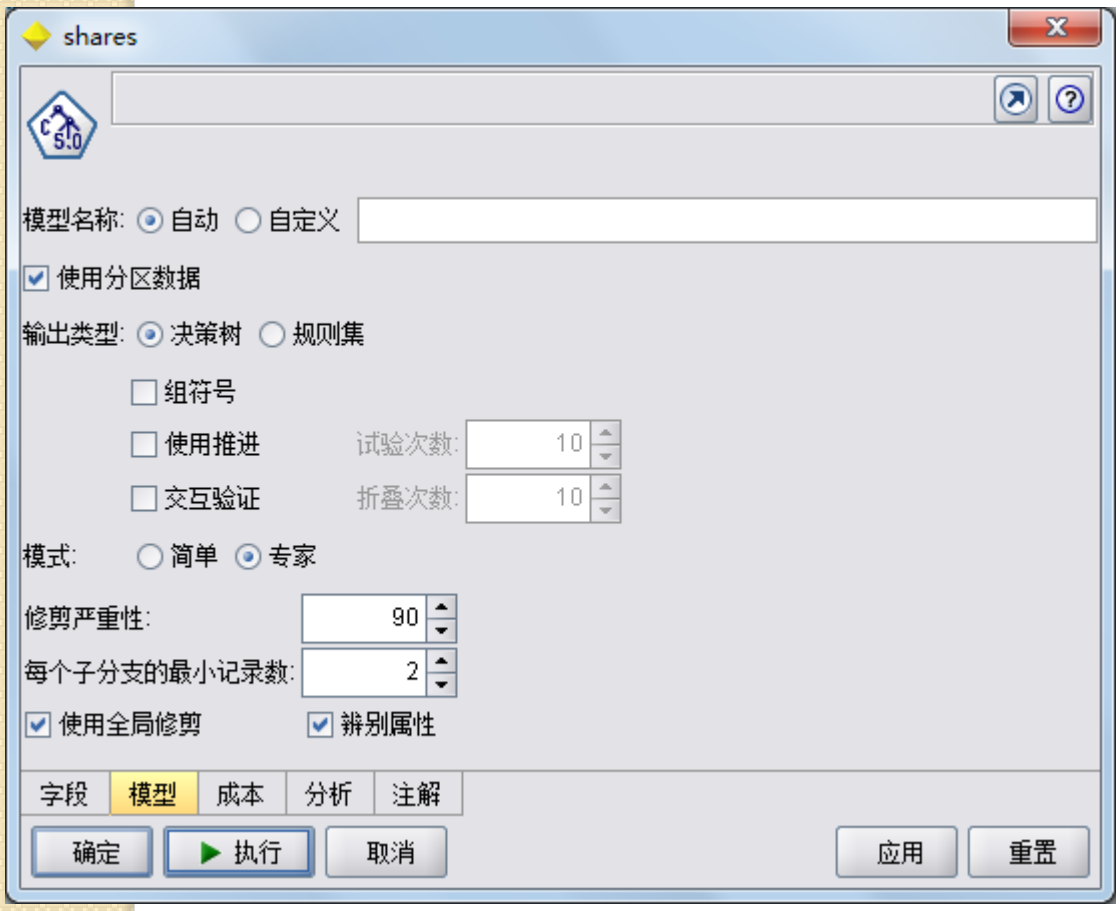
- 剪枝

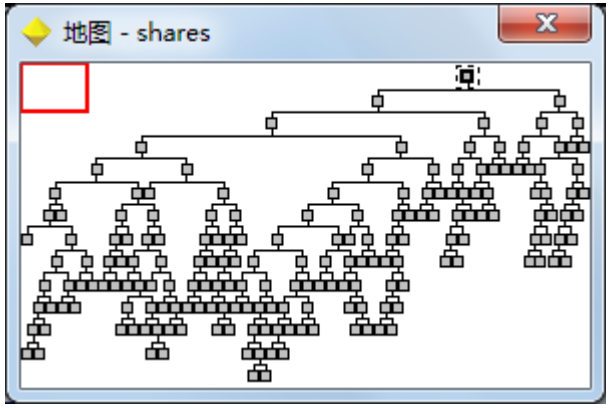

shares
模型名称: 自动 自定义
使用分区数据
输出类型: 决策树 规则集
组符号
使用推进
试验次数: 10
交互验证
折叠次数: 10
模式: 简单 专家
修剪严重性: 75
每个子分支的最小记录数: 2
使用全局修剪
辨别属性
字段 模型 成本 分析 注解
确定 执行 取消 应用 重置


地图 - shares

Step 9

- 剪枝参数设置





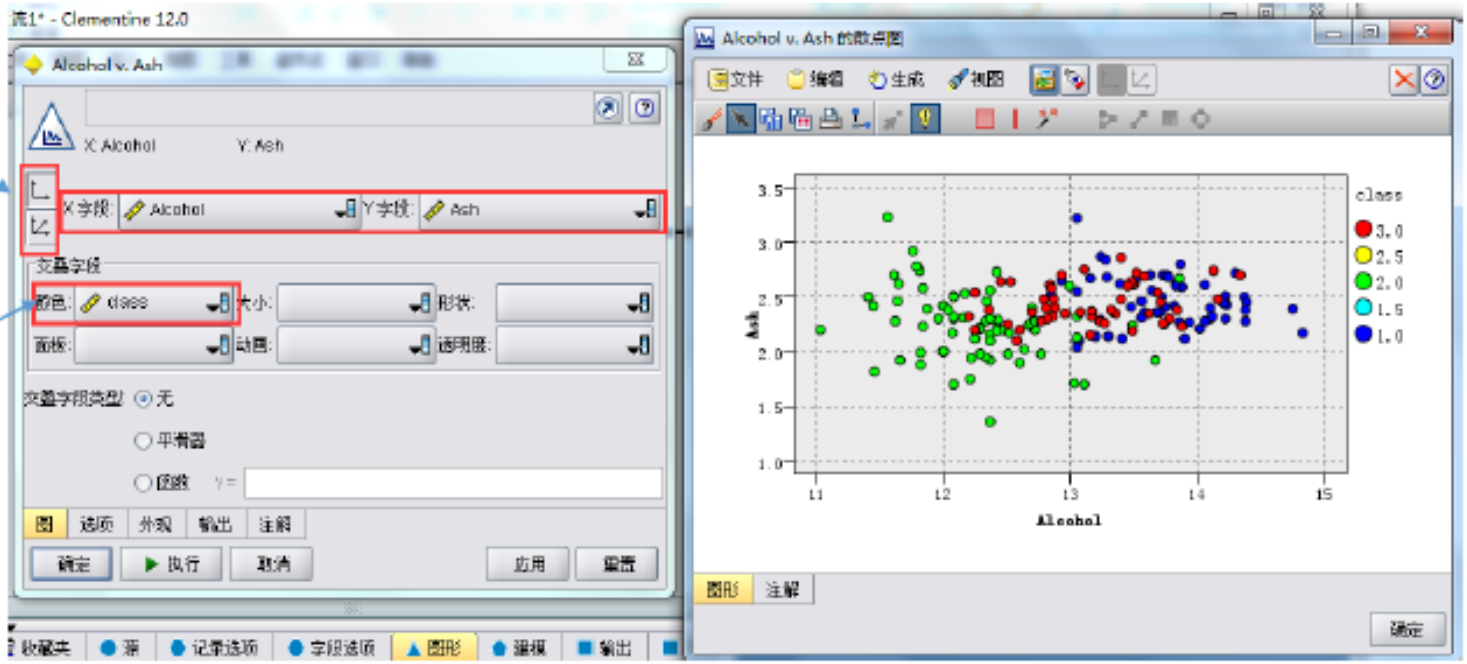
Graphics

- 用于可视化数据集、检查新导出字段的分布和关联，方便对数据进行探索性分析。
- 图形包含以下节点：

  - 图（散点图）、分布图（条形图）、直方图（柱形图）、集合、多重散点图、网络图、时间散点图、评估图

选择2D散点图或3D散点图

观察这一变量的取值情况



Example: Apriori

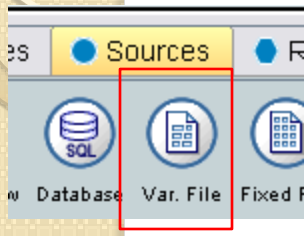
- 超市售货
  - 某超市共有20种商品：意大利面、牛奶、水、饼干、咖啡、奶油蛋糕、酸奶、速冻蔬菜、金枪鱼、啤酒、番茄汁、可乐、大米、果汁、咸饼干、油、速冻鱼、冰淇淋、奶酪、罐头肉
  - 46243位顾客的交易数据。对于每一次交易数据，若购买了特定的商品，则相应的列为1；否则为0

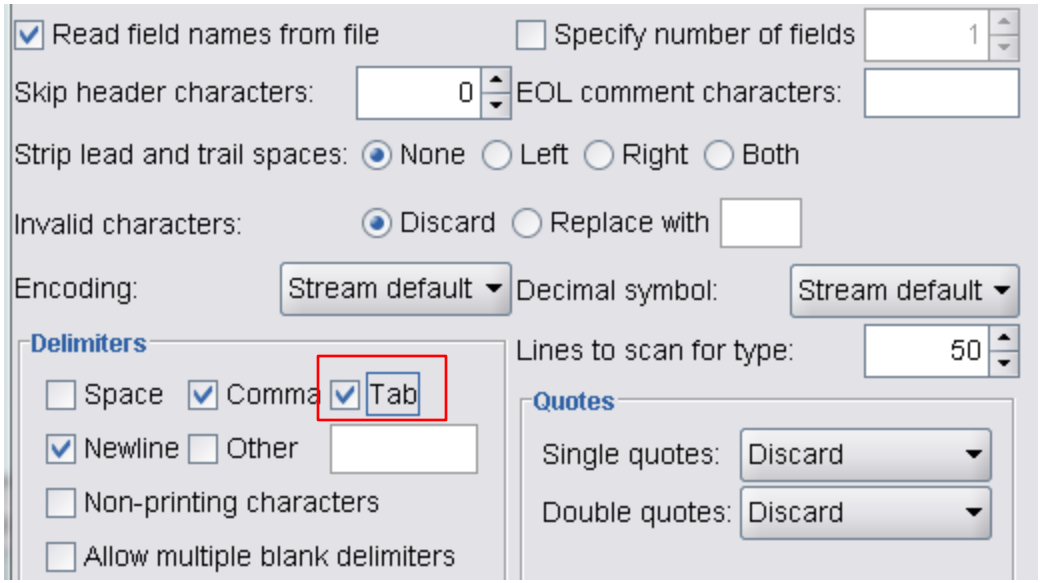
- 关联规则挖掘能够发现频繁被同时购买的商品组合，从而能够为超市商品的排架、促销等提供参考

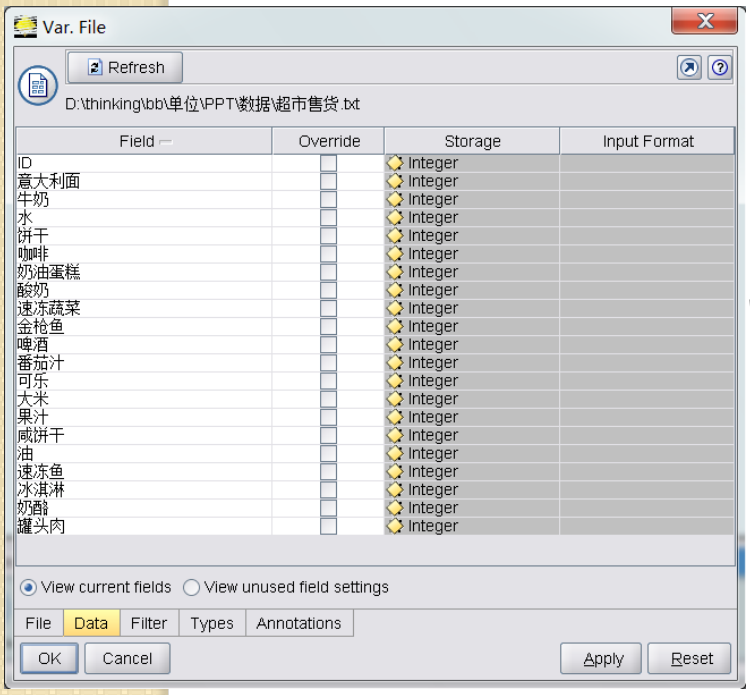
Input Data

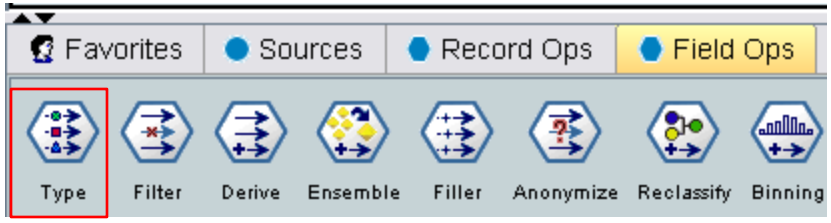
ID	意大利面			牛奶		水	饼干		咖啡		奶油蛋糕			酸奶		速冻蔬菜			金枪鱼	
1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
6	1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
7	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
12	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	1	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0
14	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Step 1

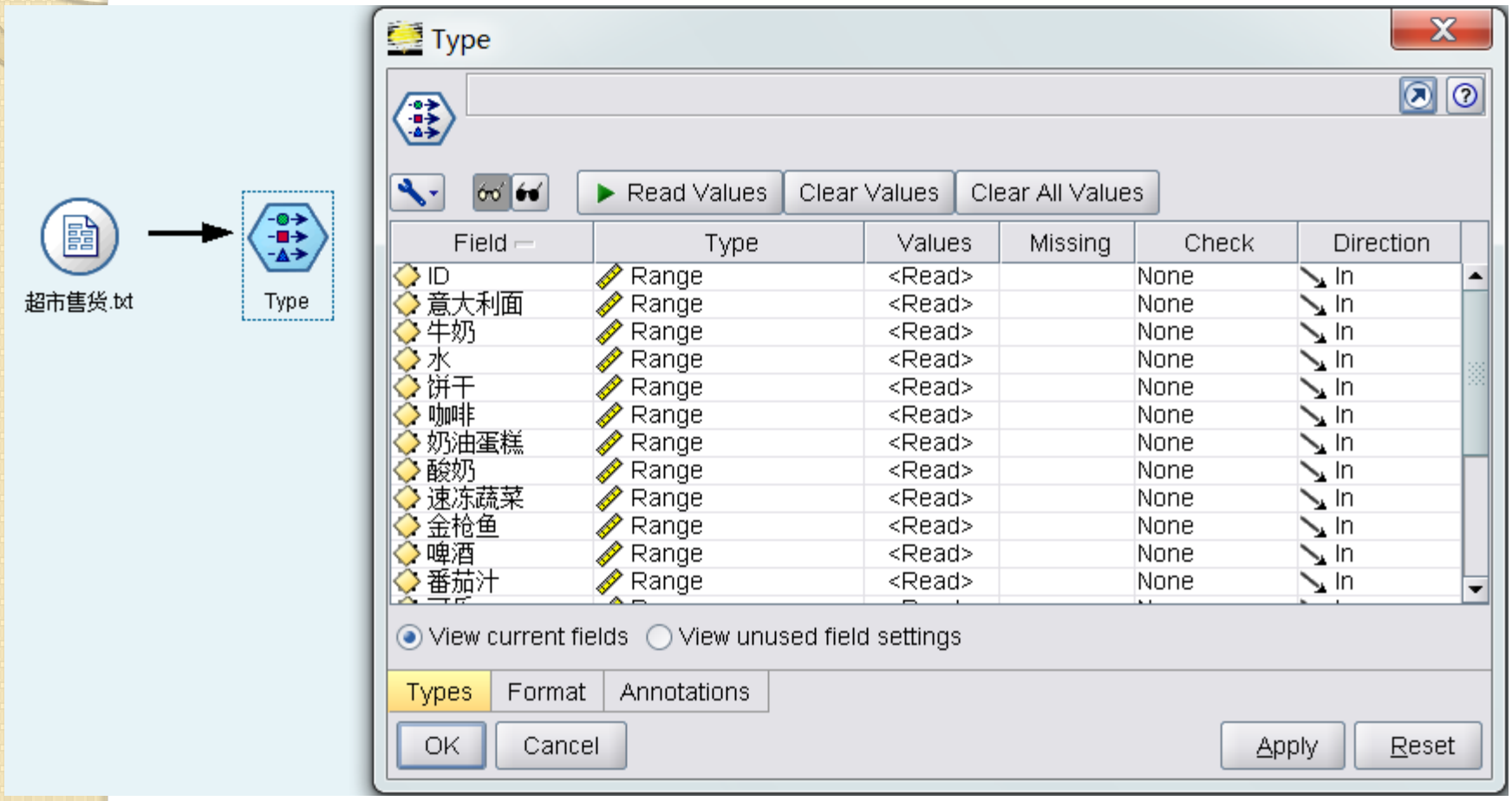

Sources
Database
Var. File
Fixed F


Read field names from file
Specify number of fields
1
Skip header characters:
0
EOL comment characters:
Strip lead and trail spaces:
None
Left
Right
Both
Invalid characters:
Discard
Replace with
Encoding:
Stream default
Decimal symbol:
Stream default
Delimiters
Space
Comma
Tab
Newline
Other
Non-printing characters
Allow multiple blank delimiters
Lines to scan for type:
50
Quotes
Single quotes:
Discard
Double quotes:
Discard


Var. File
Refresh
D:\thinking\bb\单位\PPT\数据\超市售货.txt
Field
Override
Storage
Input Format
ID
意大利面
牛奶
水
饼干
咖啡
奶油蛋糕
酸奶
速冻蔬菜
金枪鱼
啤酒
番茄汁
可乐
大米
果汁
咸饼干
油
速冻鱼
冰淇淋
奶酪
罐头肉
Integer
View current fields
View unused field settings
File
Data
Filter
Types
Annotations
OK
Cancel
Apply
Reset

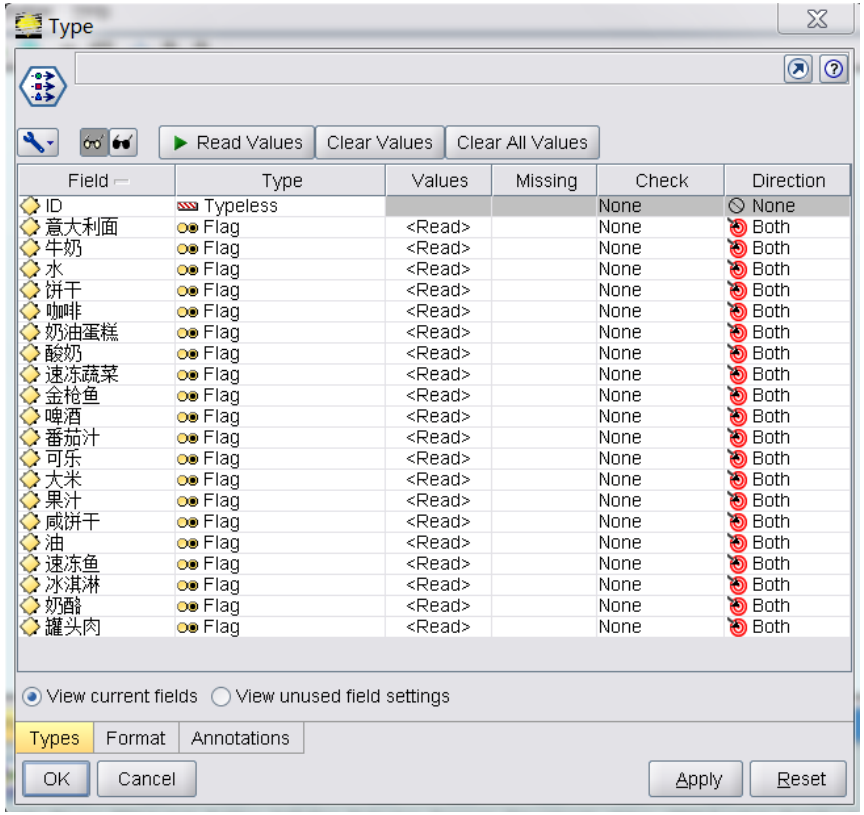

Favorites
Sources
Record Ops
Field Ops
Type
Filter
Derive
Ensemble
Filler
Anonymize
Reclassify
Binning

Step 2


超市售货.txt
Type
Type
Read Values
Clear Values
Clear All Values
Field
Type
Values
Missing
Check
Direction
ID Range <Read> None In
意大利面 Range <Read> None In
牛奶 Range <Read> None In
水 Range <Read> None In
饼干 Range <Read> None In
咖啡 Range <Read> None In
奶油蛋糕 Range <Read> None In
酸奶 Range <Read> None In
速冻蔬菜 Range <Read> None In
金枪鱼 Range <Read> None In
啤酒 Range <Read> None In
番茄汁 Range <Read> None In
View current fields
View unused field settings
Types
Format
Annotations
OK
Cancel
Apply
Reset

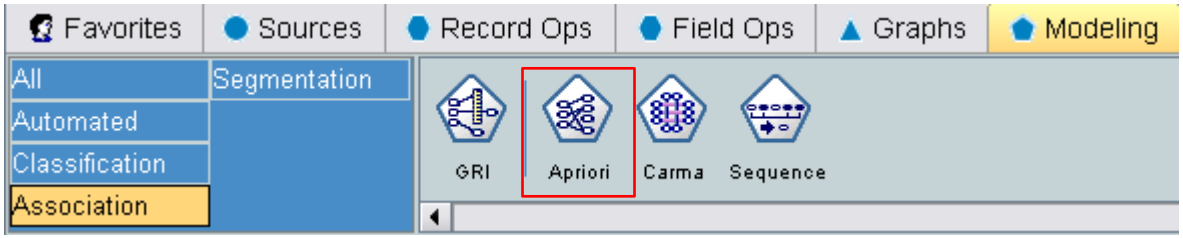
Step 3

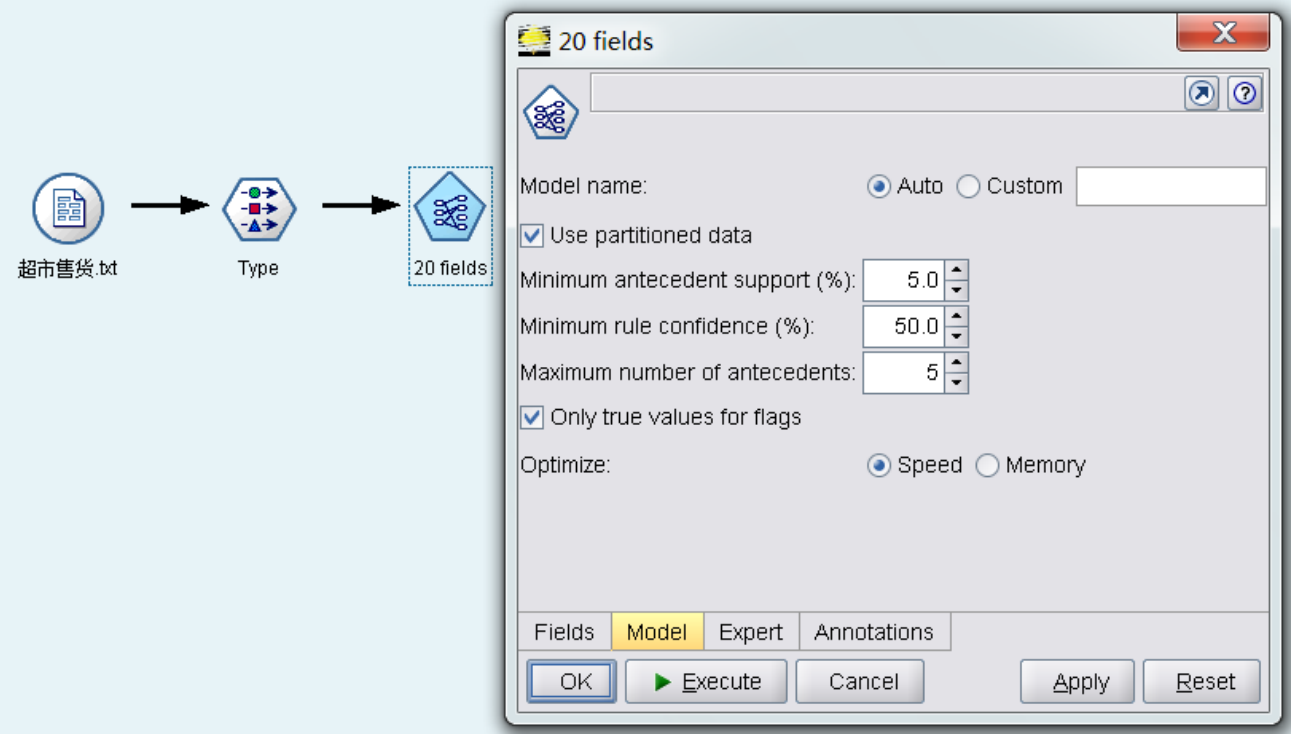
- 右键——“Select All”——“Set Type”----Flag, “Set Direction”----Both
- 将特征“ID”的Type单独设置为Typeless



Field	Type	Values	Missing	Check	Direction
ID	Typeless			None	None
意大利面	Flag	<Read>		None	Both
牛奶	Flag	<Read>		None	Both
水	Flag	<Read>		None	Both
饼干	Flag	<Read>		None	Both
咖啡	Flag	<Read>		None	Both
奶油蛋糕	Flag	<Read>		None	Both
酸奶	Flag	<Read>		None	Both
速冻蔬菜	Flag	<Read>		None	Both
金枪鱼	Flag	<Read>		None	Both
啤酒	Flag	<Read>		None	Both
番茄汁	Flag	<Read>		None	Both
可乐	Flag	<Read>		None	Both
大米	Flag	<Read>		None	Both
果汁	Flag	<Read>		None	Both
咸饼干	Flag	<Read>		None	Both
油	Flag	<Read>		None	Both
速冻鱼	Flag	<Read>		None	Both
冰淇淋	Flag	<Read>		None	Both
奶酪	Flag	<Read>		None	Both
罐头肉	Flag	<Read>		None	Both

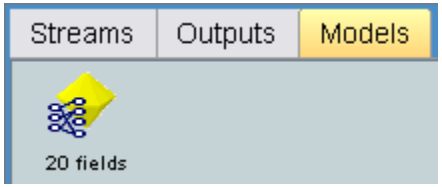
Step 4

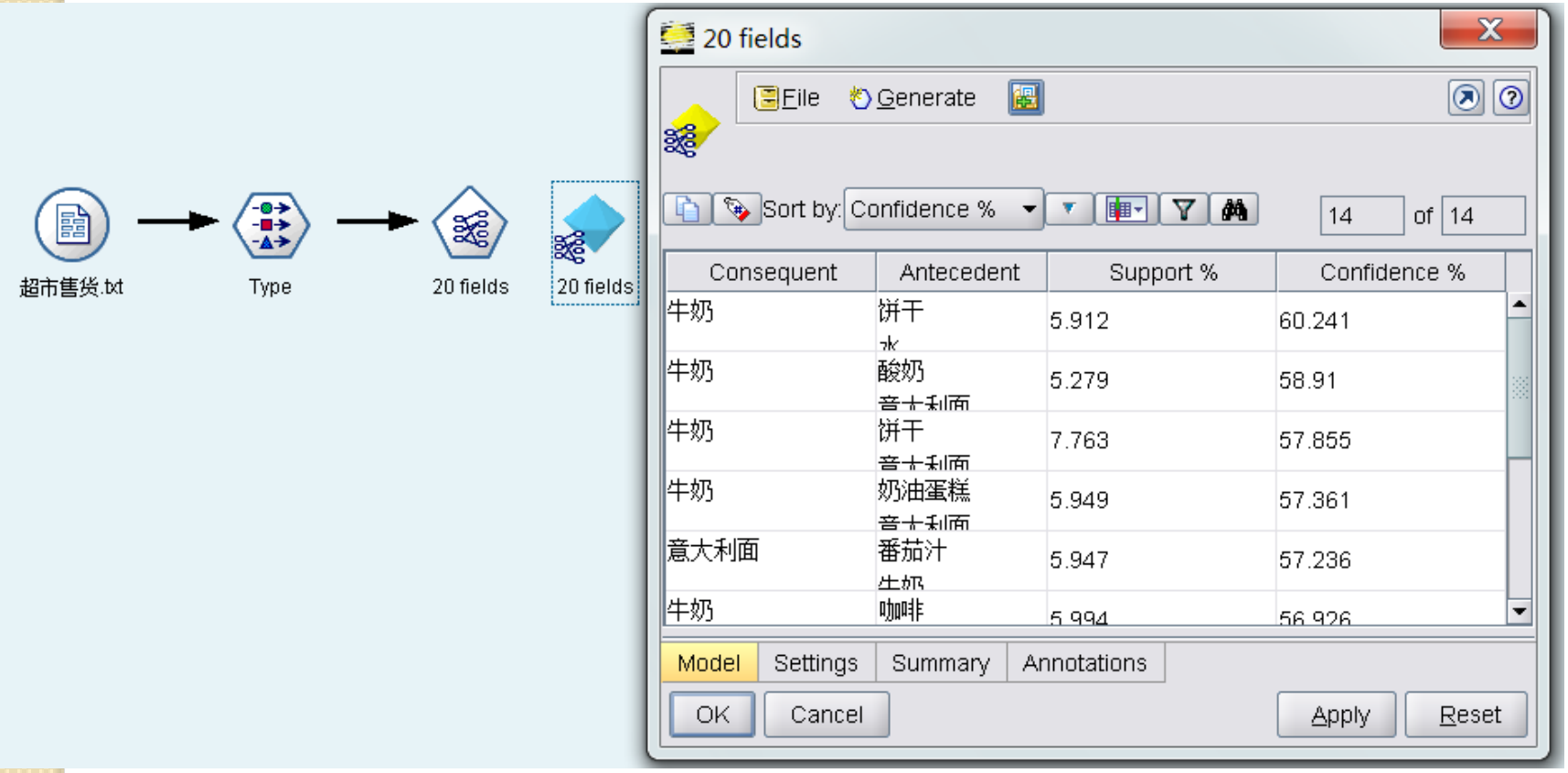

Favorites
Sources
Record Ops
Field Ops
Graphs
Modeling
All
Segmentation
Automated
Classification
Association
GRI
Apriori
Carma
Sequence


超市售货.txt
Type
20 fields
20 fields
Model name:
Auto
Custom
Use partitioned data
Minimum antecedent support (%):
5.0
Minimum rule confidence (%):
50.0
Maximum number of antecedents:
5
Only true values for flags
Optimize:
Speed
Memory
Fields
Model
Expert
Annotations
OK
Execute
Cancel
Apply
Reset

Step 5

- 点击“Execute”后得到结果





Weka

- WEKA is a machine learning tool
  - Tools for pre-processing data
  - Many popular machine learning algorithms
  - Visualization tools
  - Experiment management tools
- An open source package
  - Written in Java
  - Available under the GNU Public License
- http://www.cs.waikato.ac.nz/~ml/weka/

Weka

- WEKA has many algorithms useful for text mining
  - Classification: decision trees, Naïve Bayes, …
  - Clustering: k-means, EM, …
- Using WEKA for text mining
  - Feature selection (outside of WEKA) to reduce the size of the problem
- E. Frank. "Machine learning with WEKA." Department of Computer Science, University of Waikato, New Zealand.

Weka Input

- WEKA input must be in ARFF format

@relation weather

以下为属性定义:

@attribute outlook {sunny, overcast, rainy}

@attribute temperature real

@attribute humidity real

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}

The last attribute is the class to be predicted

Weka Input

- WEKA input must be in ARFF format

@data

sunny,85,85,FALSE,no

sunny,80,90,TRUE,no

overcast,83,86,FALSE,yes

rainy,70,96,FALSE,yes

rainy,68,80,FALSE,yes

rainy,65,70,TRUE,no

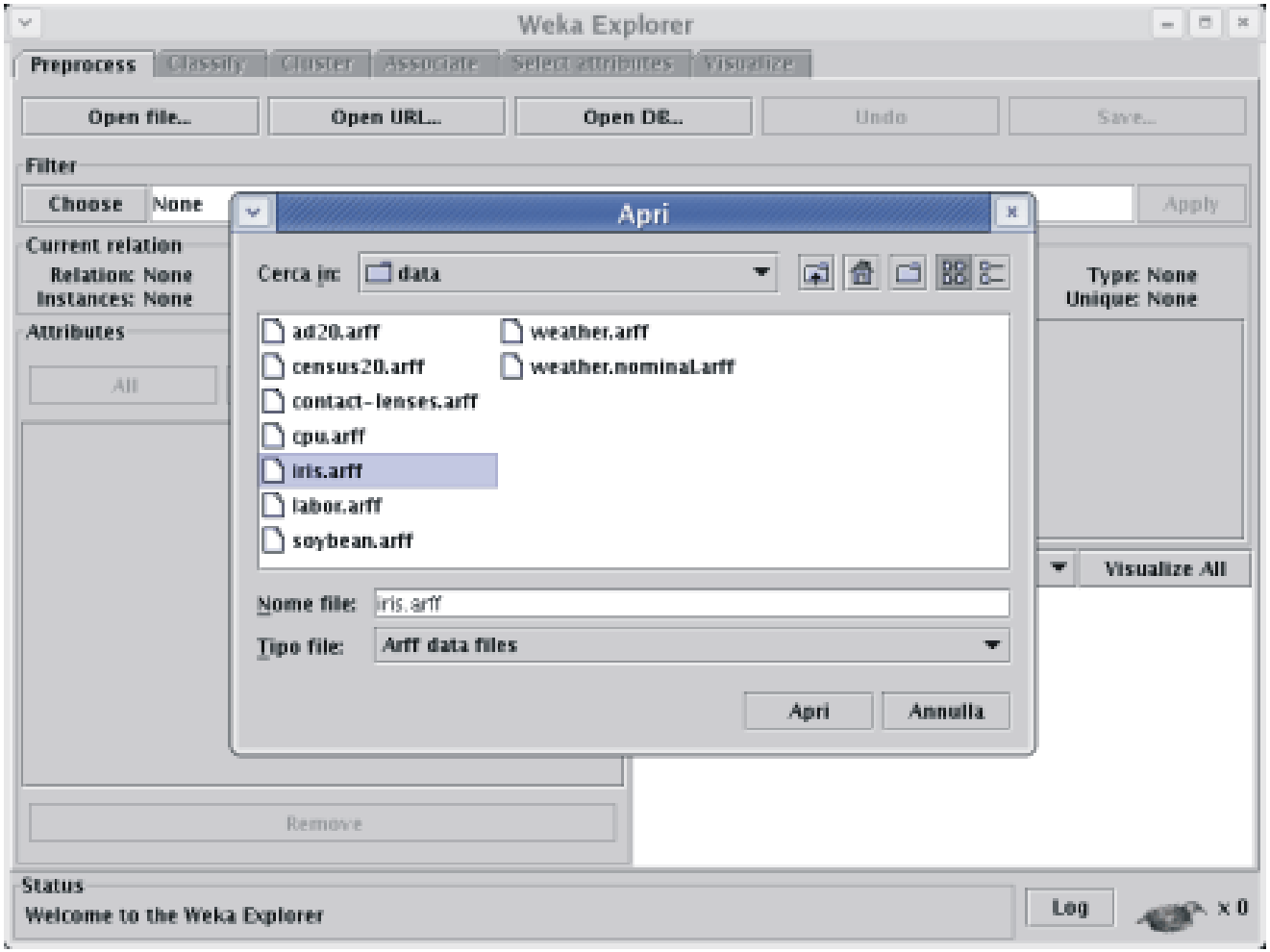
…

Weka Input

- 将CSV转换为ARFF最迅捷的办法是使用WEKA所带的命令行工具
  - 运行WEKA的主程序，在菜单中找到"Simple CLI"模块，它可提供命令行功能。在新窗口最下方输入框写入：java weka.core.converters.CSVLoader filename.csv > filename.arff
- 在WEKA 3.5中提供了一个"Arff Viewer"模块，可以用它打开一个CSV文件将进行浏览，然后另存为ARFF文件。或者进入"Exploer"模块，从上方的按钮中打开CSV文件然后另存为ARFF文件

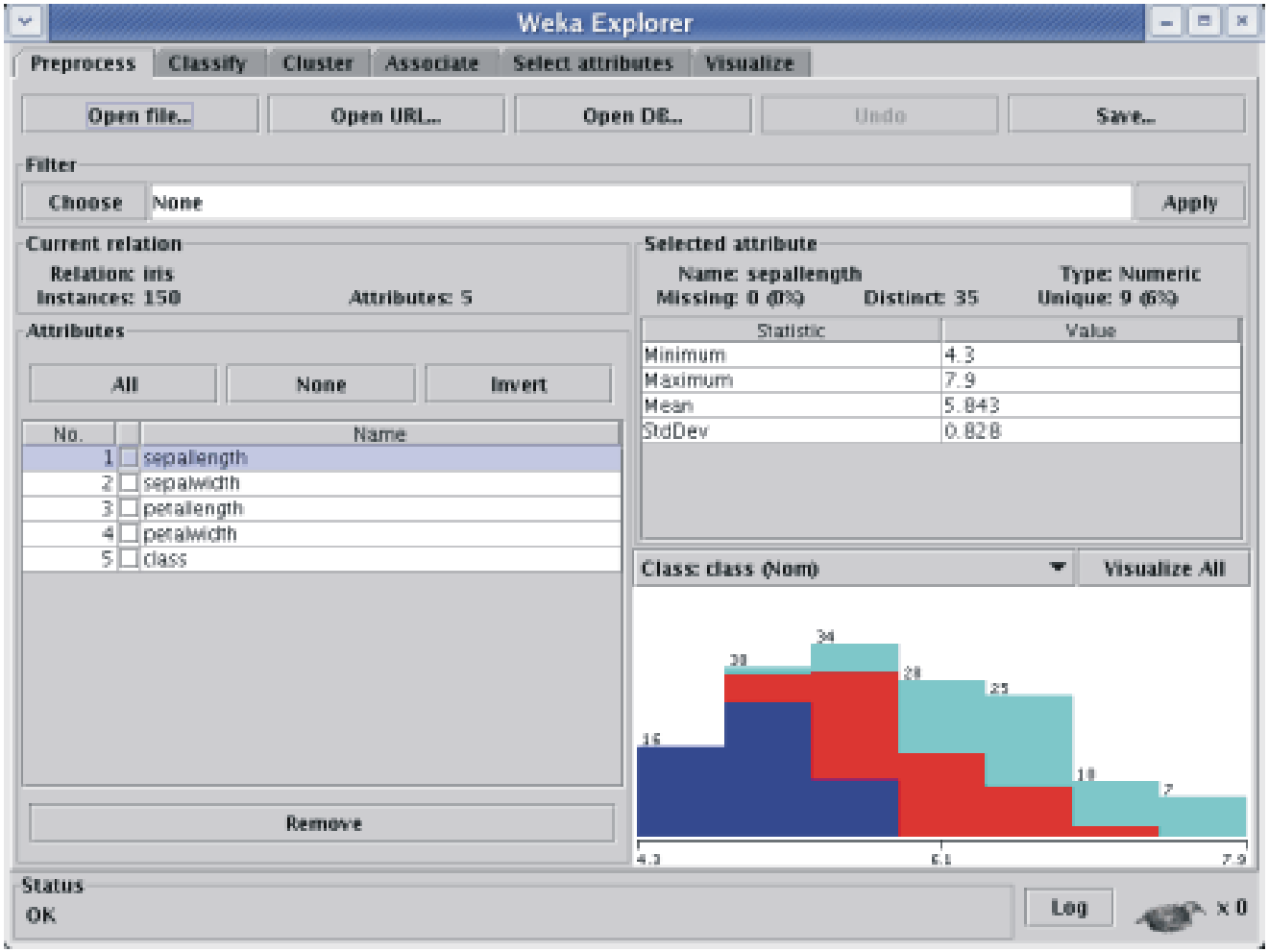
Example: Classification

- Input Dataset



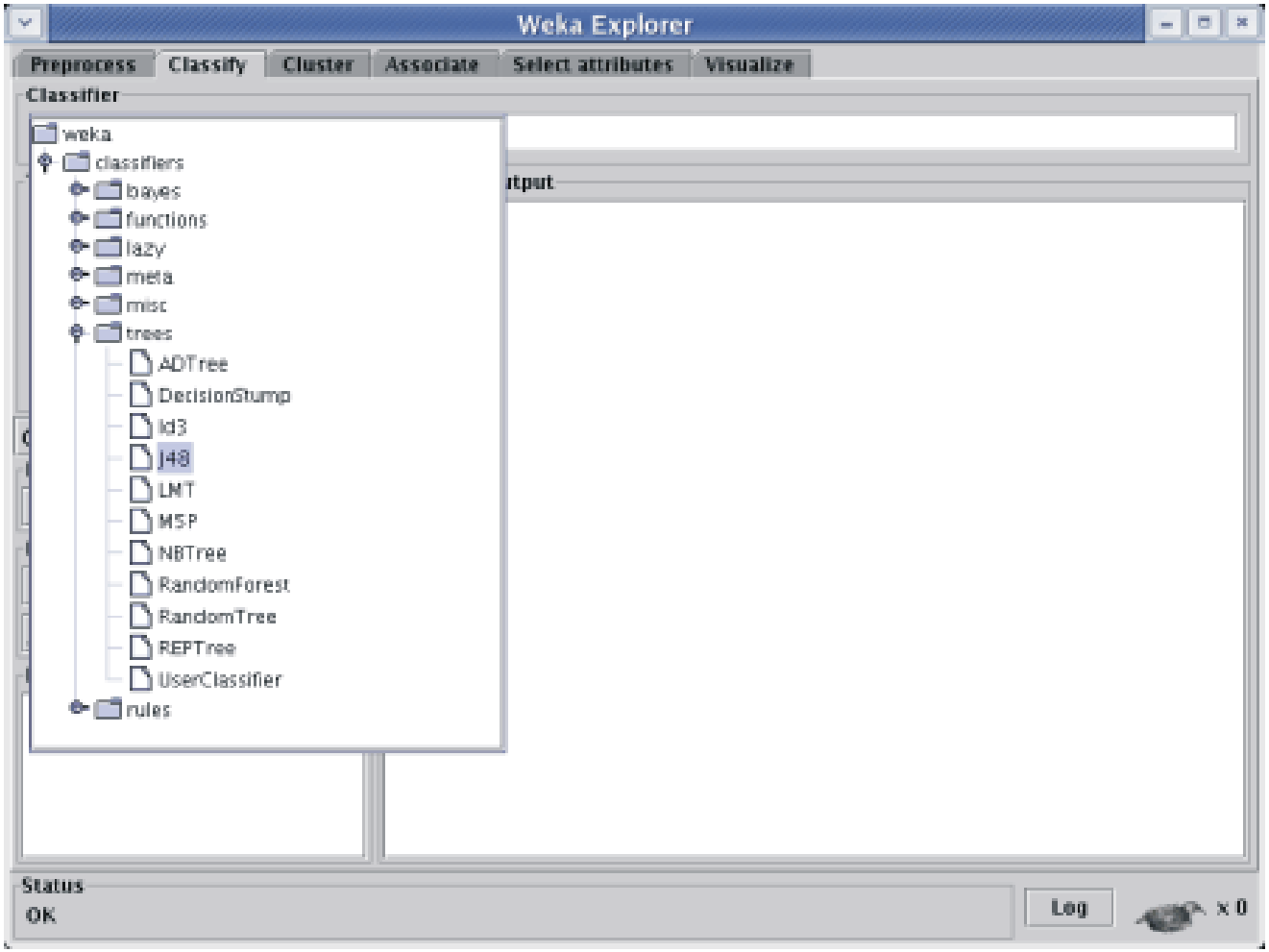
Example: Classification

- Dataset Statistics



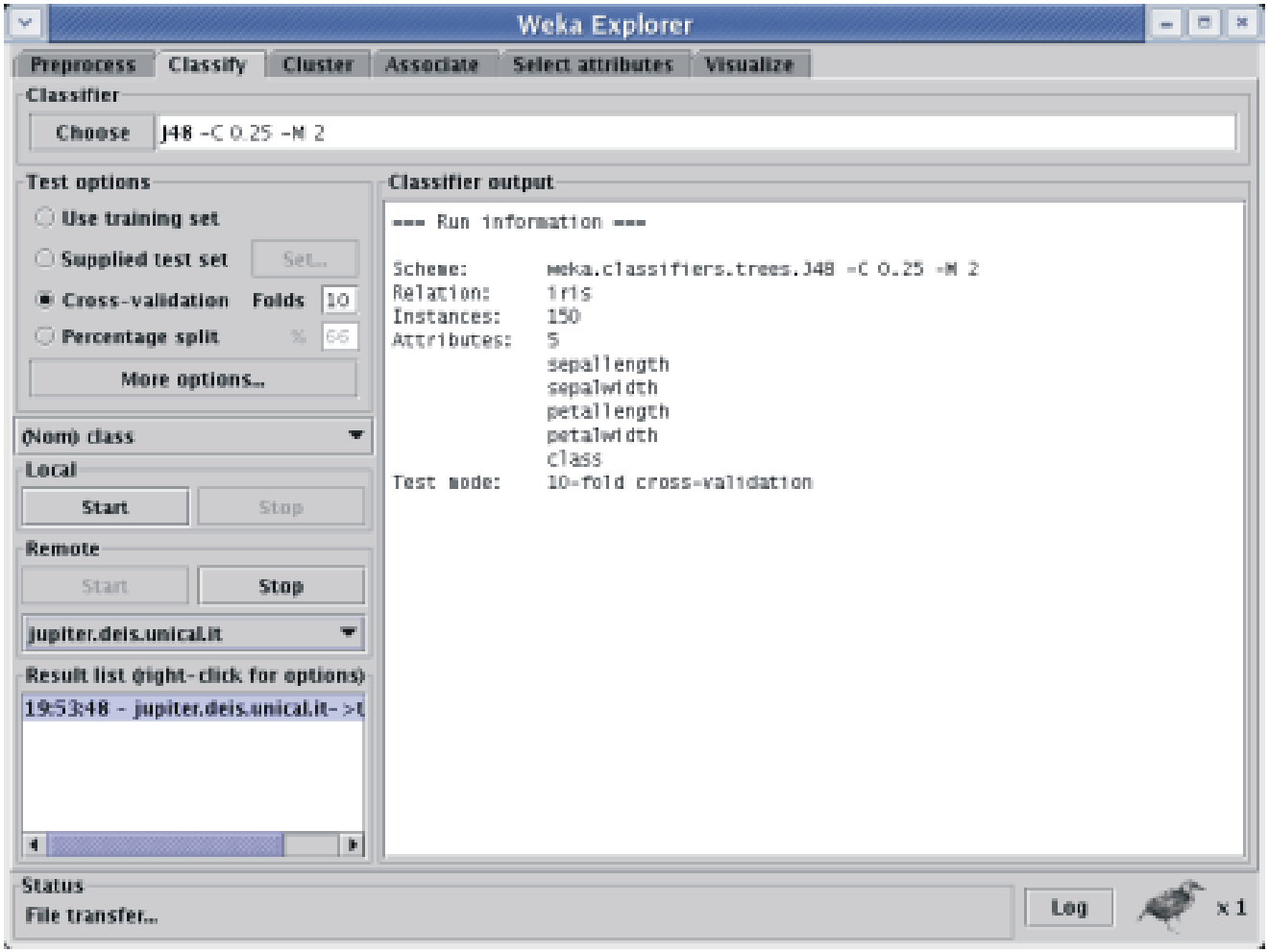
Example: Classification

- Select the panel corresponding to the task to be performed and click on "Choose" to select the algorithm to be used



Example: Classification

- Start the task by clicking the "Start" button



Example: Classification

- Whenever the results have been received, they are visualized in the "Classifier output" panel