

关联规则—CARMA

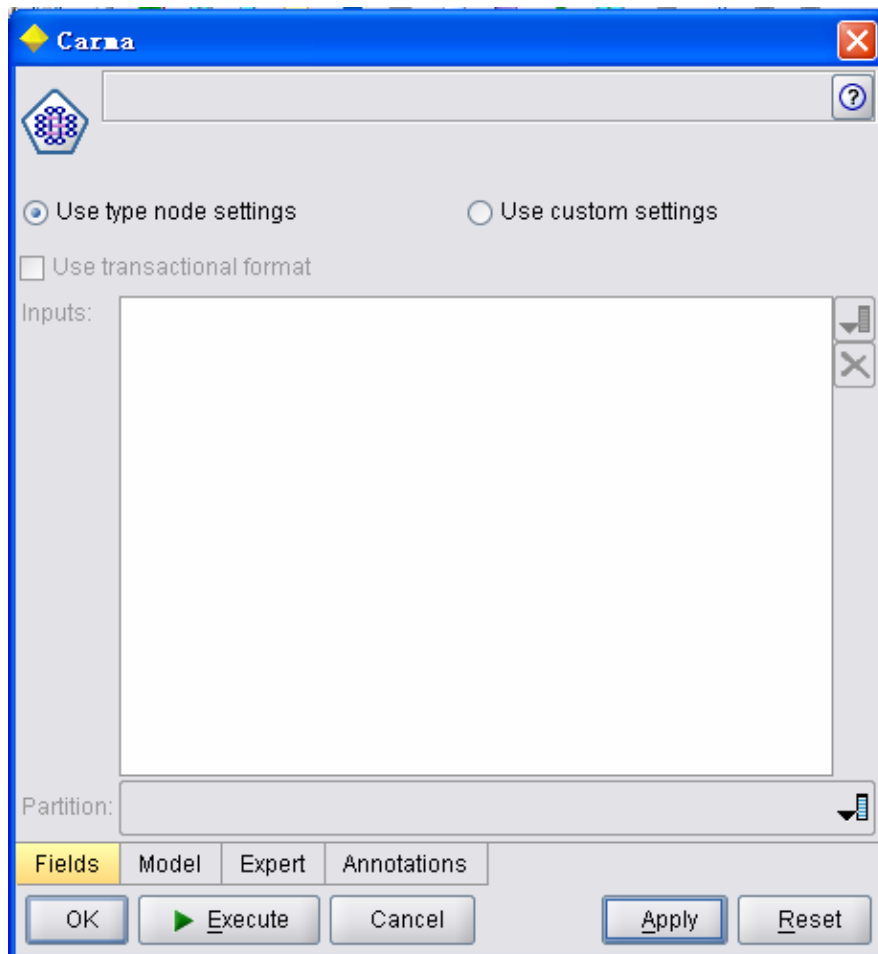
Continuous Association Rule Mining Algorithm

CARMA 算法简介

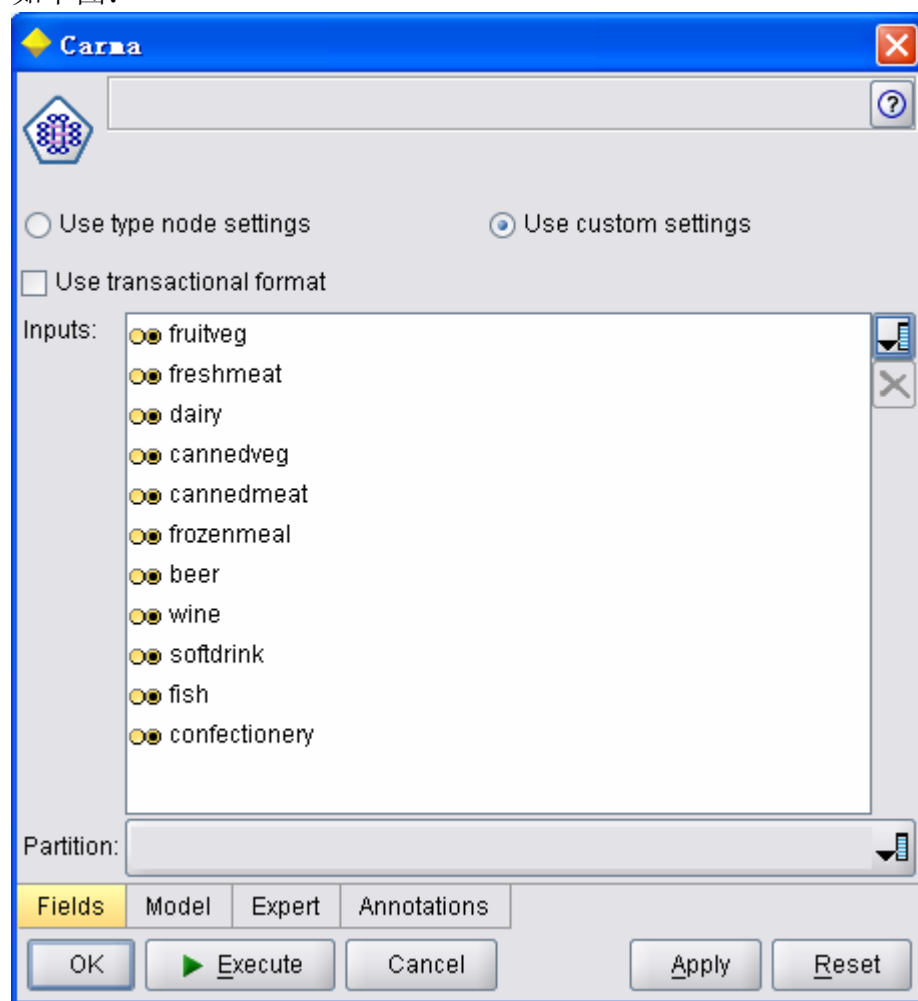
CARMA 是一种比较新的关联规则算法，它是 1999 年由 Berkeley 大学的 Christian Hidber 教授提出来的。它是一种占用内存少，能够处理在线连续交易流数据的一种新型的关联规则。CARMA 算法构造关联规则一般由两个阶段完成，经过第一阶段扫描，CARMA 算法产生一个满足给定支持度的所有大项目集；第二阶段，CARMA 算法则再次扫描数据流剔除掉上一阶段产生的项目集中比较小的项目集合，但是很多时候第二阶段不需要做。CARMA 算法还有一个很好的性质，就是在第一阶段扫描交易流的过程中可以不断的改变支持度，以控制输出的规则的大小和数目，这是其它关联规则所没有的。

CARMA 模型参数设置：

Carma 模型 Fields 选项卡：



勾选 Use type node settings 意思是使用 type node 定义的所有 flag 变量进行建模。
选择 Use custom settings 则可以让用户自己选择需要的 flag 变量进行建模，
如下图：



Use transactional format 是用来为模型指定要处理的数据格式的，关联规则要处理的数据一般分两种，一种是 transactional 的数据，一个是 tabular 的数据，系统默认格式为 tabular。如图：

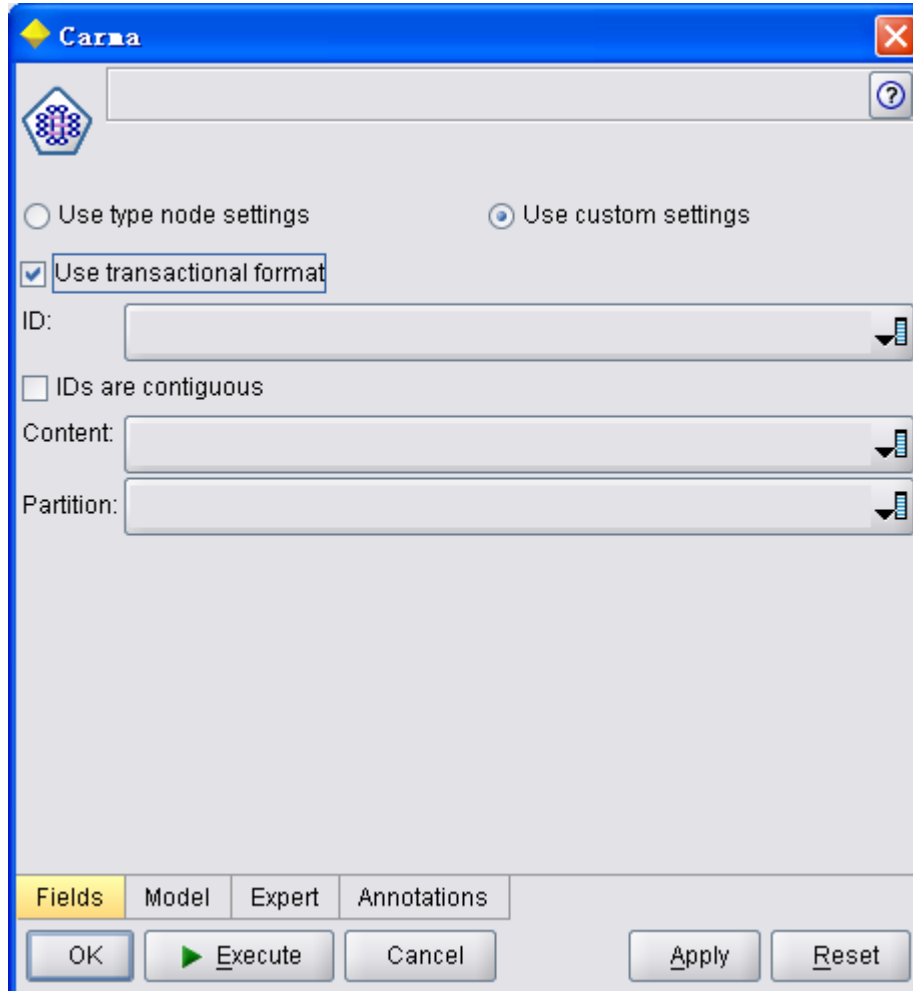
Transactional 数据格式：

Customer	Purchase
1	jam
2	milk
3	jam
3	bread
4	jam
4	bread
4	milk

Tabular 数据格式：

Customer	Jam	Bread	Milk
1	T	F	F
2	F	F	T
3	T	T	F
4	T	T	T

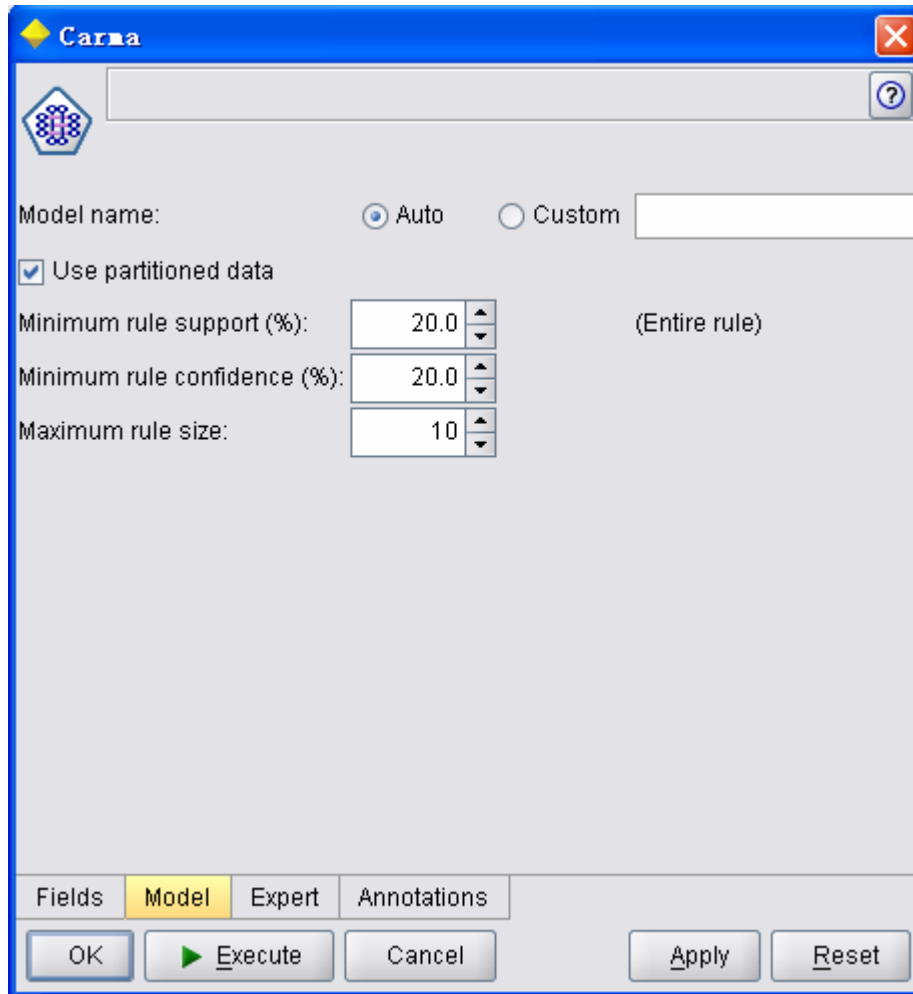
如果勾选了 Use transactional format 选项，Field 选项卡变成下图所示的形式，这个时候系统将按照 transactional 的数据格式来处理数据。



ID: 指定数据的唯一标识，比如：客户 ID。如果选定 IDs are contiguous，那么系统将默认数据已经按照 ID 号排序了，所以系统不再对其进行排序，缩短了建模的时间。

Content: 该项选择交易数据字段。

Carma 模型 Model 选项卡:



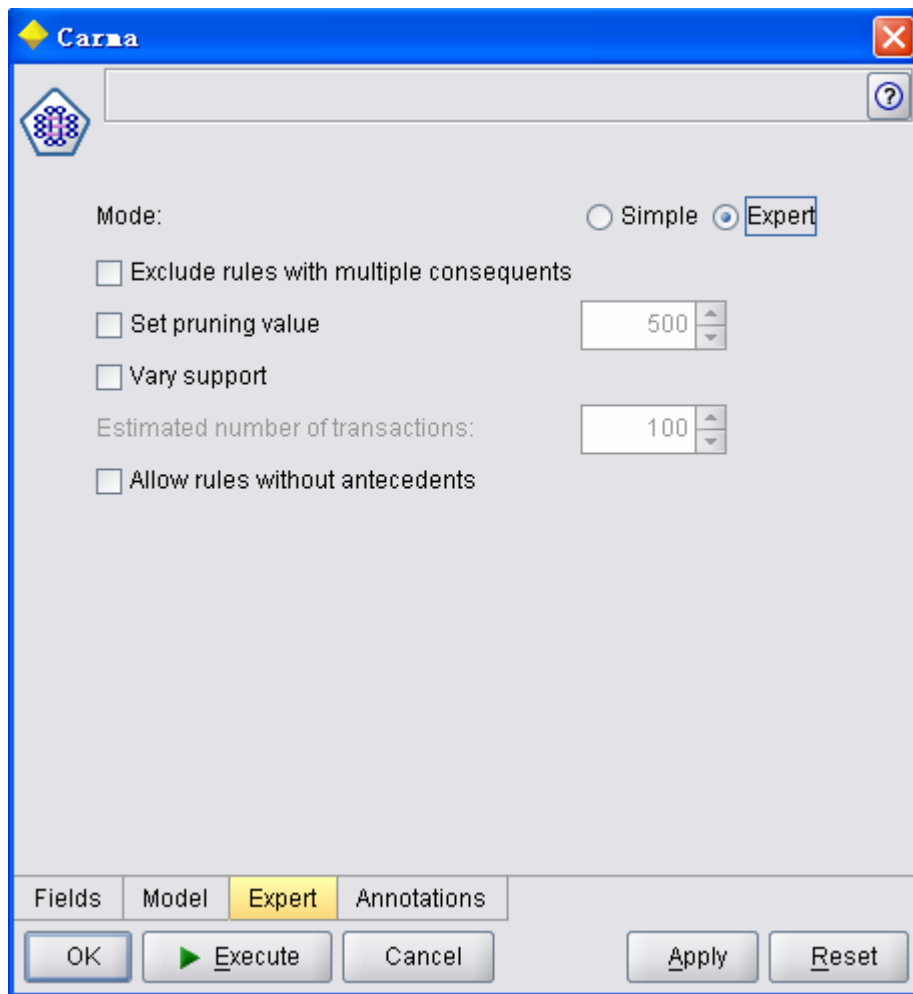
Use partitioned data: 使用该选项可以师兄对数据进行交叉验证等操作来检验模型的优劣，但是在使用之前必须用 `partition node` 对数据进行分割，将数据分割 `training` 和 `testing` 的数据。

Minimum rule support (%): 使用CARMA模型之前必须要设置一下Minimum rule support的值，系统默认的是 20%，这个值的大小直接关系到最后生成的规则的数目。需要注意的是在CARMA中，这个Minimum rule support (%)的定义与GRI和APRIORI是不一样的。在CARMA中，Minimum rule support指的是在整个训练集中包含某个规则的观测所占的比例，而在GRI和APRIORI中，它被定义为在整个训练集中包含某个规则中的antecedent的观测所占的比例。有一个特别需要注意的地方，在生成的CARMA模型中，用Brower查看生成的规则，其中clementine为每个规则给出的support并不是这里提到的rule support，它的support值与GRI中提供的support值是完全一致的，即antecedent support。可是为什么与APRIORI不一样呢？难道Apriori的不是antecedent support??（可能是参与建模的观测数目不一样导致的，在APRIORI里可能因为某些原因，参与建模的观测可能不到一千个。）

Minimum rule confidence (%): 使用CARMA模型需要设定的另一个参数就是规则的最小置信度。置信度定义为使用指定的规则对整个训练集中包含该规则antecedent部分的观测预测其consequent的正确率。系统默认是 20%。

Maximum rule size: 设置每个规则中允许的最大元素个数，当只需要比较短的规则时，可以通过减小Maximum rule size来加速建模。

Carma 模型 Expert 选项卡:



Exclude rules with multiple consequents: 如果勾选这个选项的话，那么最后得到的规则集将不包含含有一个以上consequent的规则。比如，有规则bread & cheese & fish \Rightarrow wine&fruit，因为wine&fruit含有两个consequent，所以这条规则将被舍弃。

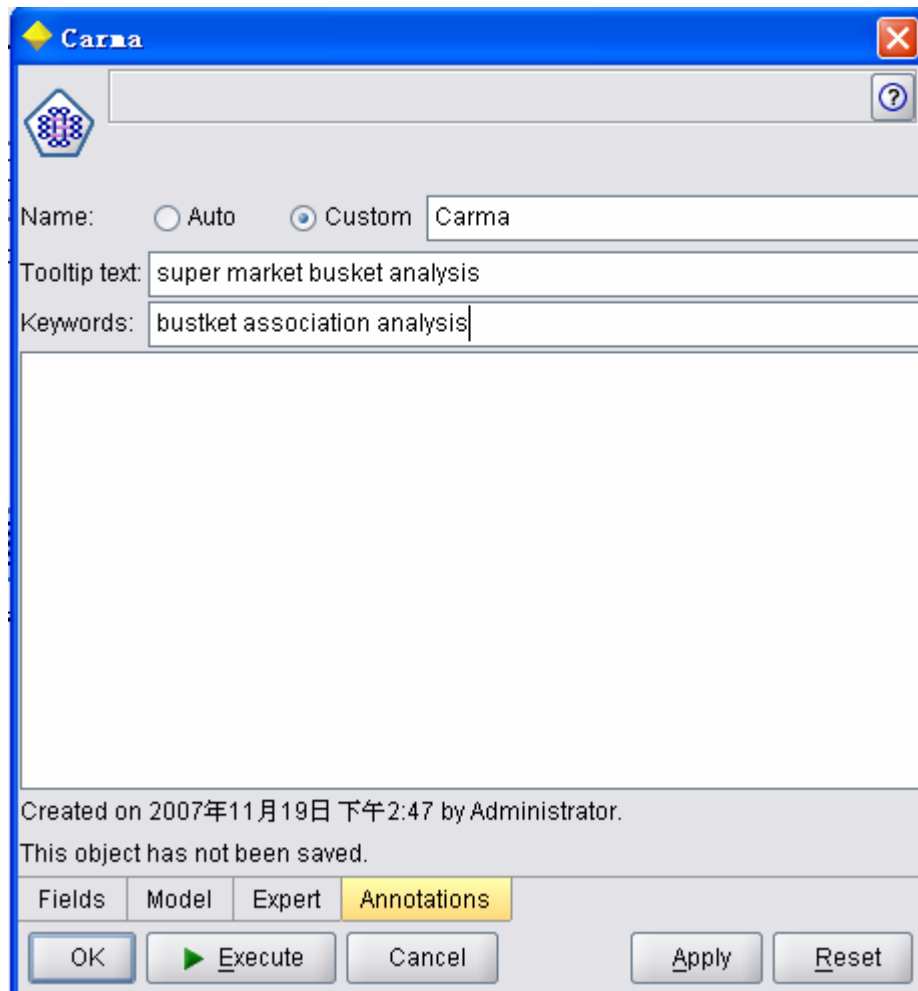
Set pruning value: 为了节省内存，CARMA 使用了一个可以周期性剔除当前出现频数比较小的规则。选择这个选项可以设定 CARMA 执行这个算法的周期，如果该值设定的比较小，可能可以减少该算法所占用的内存，但是可能会增加建模的时间；反之，如果该值设定的比较大，那么可能会增大该算法所占用的内存，但是可以减少建模所需要的时间。系统设定的默认值为 500。

Vary support: 该选项和上一个选项有一些相似的地方，选择该选项，可以通过删除一些看似出现频率很高（其实不是）的规则集来提高建模的效率。其原理是，为模型设定一个很大的初试support值，这样可以将很多出现频率不够高的规则排除在外，然后再逐渐减小support的值，直到减为Model选项卡中Minimum rule support的值。该选项还有一个子选项Estimated number of transactions，通过设定该选项的值来控制support值减小的速度。

Allow rules without antecedents: 选择该选项，可以允许规则集中出现没有

antecedent 的规则。在某些情况下，这个选项是很有用的，比如，如果想知道的就是顾客最经常购买的商品或商品组合，那么选择该选项就可以得到想要的结果。系统默认不使用该选项。

Carma 模型 Annotations 选项卡：

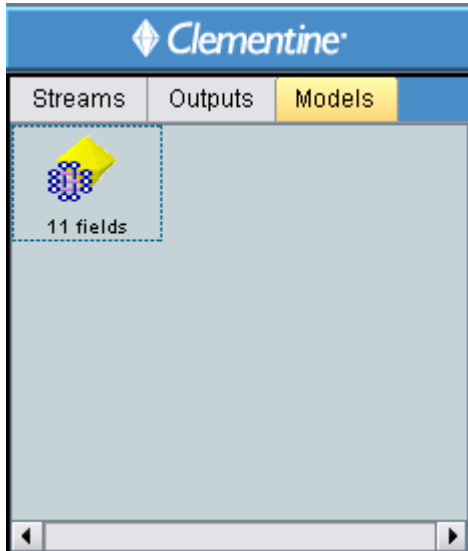


在该选项卡中可以为模型命名、加注释等。

生成 CARMA 模型

生成模型

生成的模型将会被放在 GE Palette 内，如图：



查看模型结果

对生成的模型点击右键，通过点选 browse 选项来查看生成的规则（Model 选项卡）。如图：

Consequent	Antecedent	Rule ID	Instances	Support %	Confidence %	Rule Support %	Lift	Deployability
frozenmeal	cannedveg beer	1	167	17.766	87.425	15.532	2.721	2.234
cannedveg	frozenmeal beer	2	170	18.085	85.882	15.532	2.664	2.553
beer	cannedveg frozenmeal	3	173	18.404	84.393	15.532	2.707	2.872
frozenmeal	beer	4	293	31.170	58.020	18.085	1.806	13.085
cannedveg	frozenmeal	5	302	32.128	57.285	18.404	1.777	13.723
frozenmeal	cannedveg	6	303	32.234	57.096	18.404	1.777	13.830
cannedveg	beer	7	293	31.170	56.997	17.766	1.768	13.404
beer	frozenmeal	8	302	32.128	56.291	18.085	1.806	14.043
beer	cannedveg	9	303	32.234	55.116	17.766	1.768	14.468
wine	confectionery	10	276	29.362	52.174	15.319	1.709	14.043
confectionery	wine	11	287	30.532	50.174	15.319	1.709	15.213
cannedveg	beer	12	293	31.170	49.829	15.532	2.707	15.638
frozenmeal	fruitveg	13	292	31.064	49.658	15.426	1.561	15.638
fish	fruitveg	14	299	31.809	48.495	15.426	1.561	16.383
cannedveg	frozenmeal	15	302	32.128	48.344	15.532	2.721	16.596
beer	frozenmeal	16	303	32.234	48.185	15.532	2.664	16.702

Rule ID: Rule ID 是在建立模型时为生成的每条规则指定的一个唯一编号。在以后使用时就可以很方便的通过 ID 获取需要的规则。

Instances: 对于每一条规则，clementine都会给出一个Instances值，它指的是所有记录中包含该规则的antecedent的记录的数量。比如：有一条规则 bread=>cheese，那么如果所有记录中有 100 条记录包含了bread，那么该规则的instances就是 100。

Support : support的定义和instances很接近，不同的是support描述的不是数量，是比例。比如，有 50%的记录都包含了bread，那么该规则的support值就是 50%。

Rule Support: Rule Support是在Support定义的基础上更进一步，它指的是所有记录中既包含某规则的antecedent，又包含consequence的记录所占的比例。比如：有 20%的记录既包含了bread，又包含了cheese，那么该规则的Rule Support就是 20%。

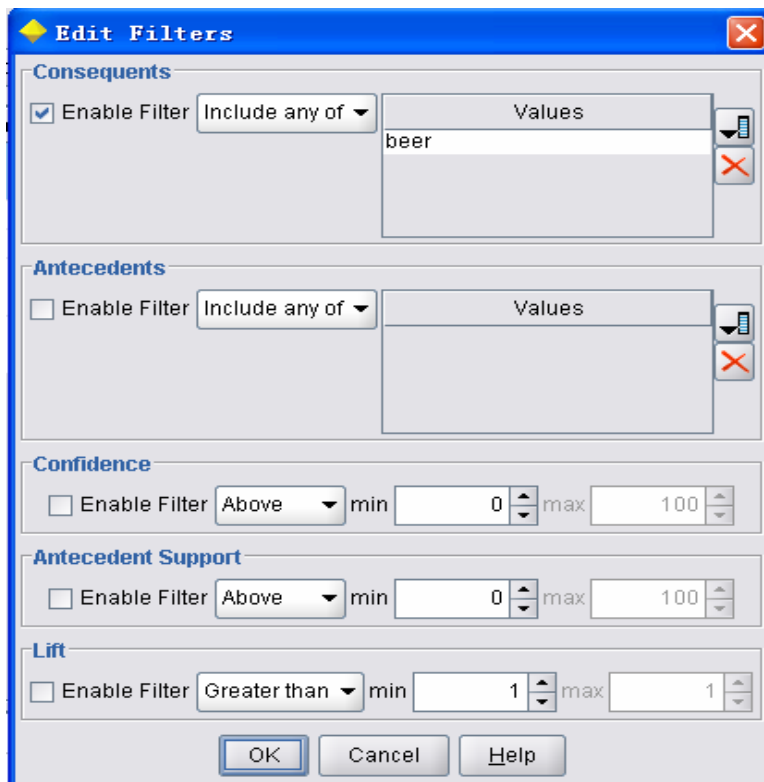
Confidence : confidence是由support和Rule Support共同定义而来的，它指的是Rule Support占Support的比例。比如：如果某规则的Rule Support是 20%，Support是 50%，那么该规则的Confidence是 40%。

Lift: 在已知某规则的 consequence 发生的先验概率的情况下，某规则的 Lift被定义为 Confidence 和该先验概率的比率值。比如：如果全部记录中有 10%买了 bread，那么如果此时有一个预测顾客买 bread 的规则 confidence 是 20%，那么该规则的 lift 就是 2；如果另外一个预测顾客购买 bread 的规则 confidence 是 11%，那么该规则的 lift 就是 1.1。也就是说，如果某规则的 lift 值越大，越偏离 1，那么该规则的预测效果就越好。

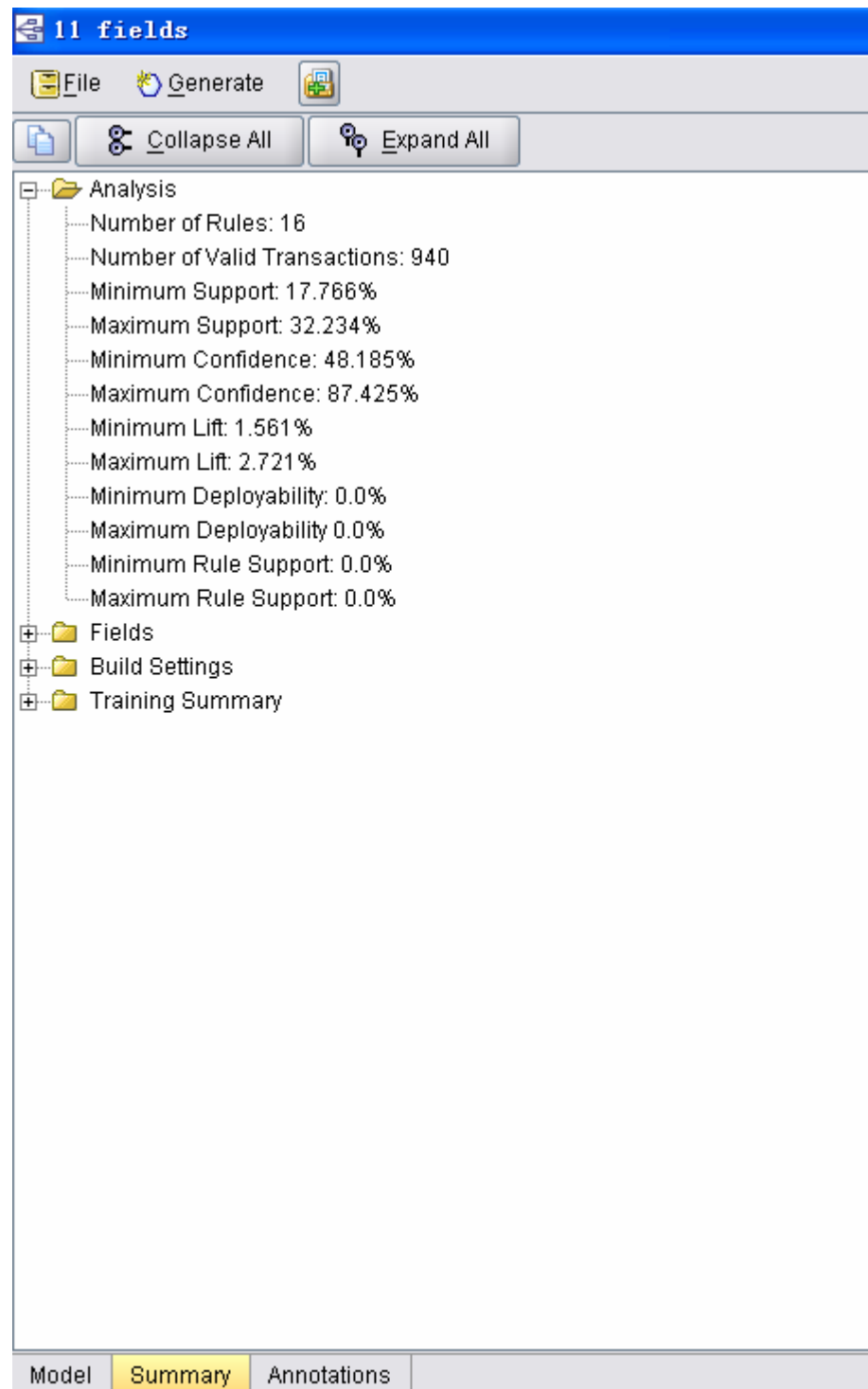
Deployability: 某规则的Deployability被定义为该规则的Support和Rule Support之差。

筛选按钮 :

可以使用筛选按钮来对 Model 选项卡中显示的规则按照指定的规则进行筛选，这样就可以在 Model 选项卡中仅显示想要的规则，便于生成仅含有特定规则的节点。该筛选节点提供了四种筛选方法，可以依据 Consequent、Antecedents、Confidence、Antecedent Support 以及 Lift 五个方面来对规则进行筛选，以及如图：



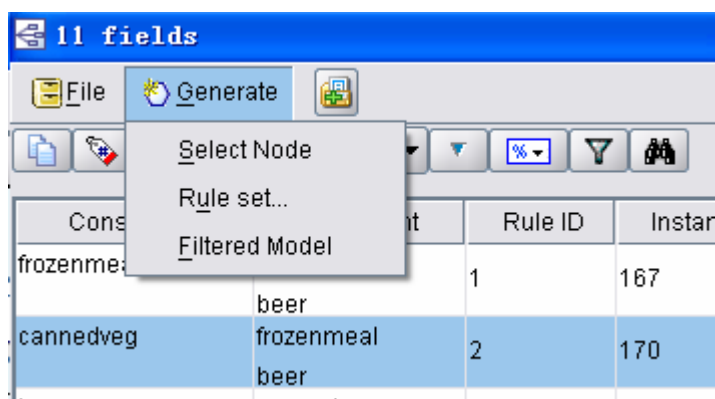
查看模型总结选项卡



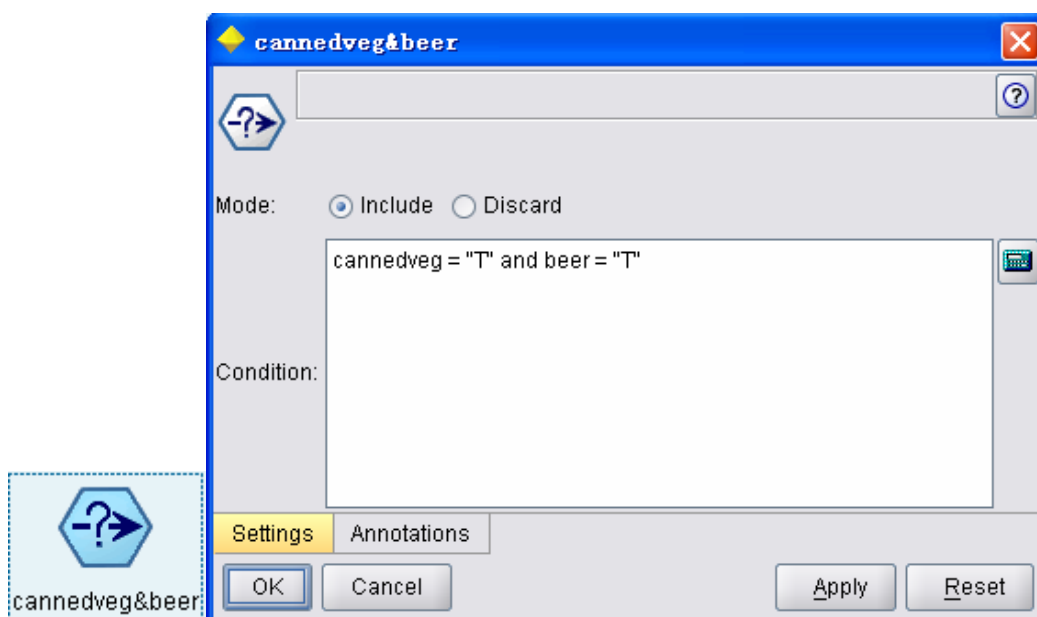
该选项卡给出了生成的 Carma 模型的各项信息的汇总。

生成节点

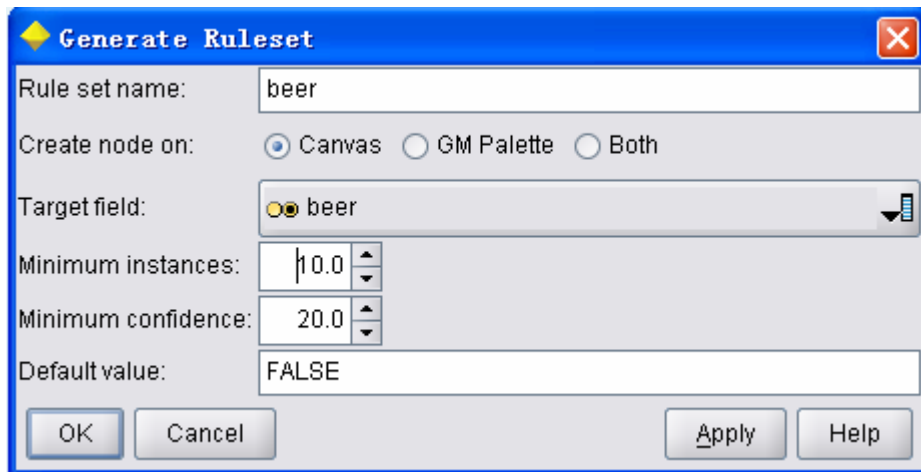
在生成的 Carma 模型中可以通过 Generate 菜单生成不同的节点 (Select Node、Rule set 以及 Filtered Model)，这些节点可以被加入到数据流中建立进一步的模型。



Select Node: 选中某条想要研究的规则，点击 Generate 菜单的 Select Node 生成一个选择节点，可以将其加入到数据流中用于筛选需要的记录。比如：选中第二条记录（如上图），生成选择节点，则该节点可以筛选出所有购买了 frozenmeal 和 beer 的记录。



Rule set: 点选该项可以为指定的 consequence 从规则集中提取出一组规则生成一个规则节点，进而将这个规则节点放入数据流中对每行记录是否含有 consequence 进行预测。如果不使用 Rule set，用生成的 Carma 模型也可以得到同样的效果。如图：



Rule set name: 为生成的规则集指定一个名称。

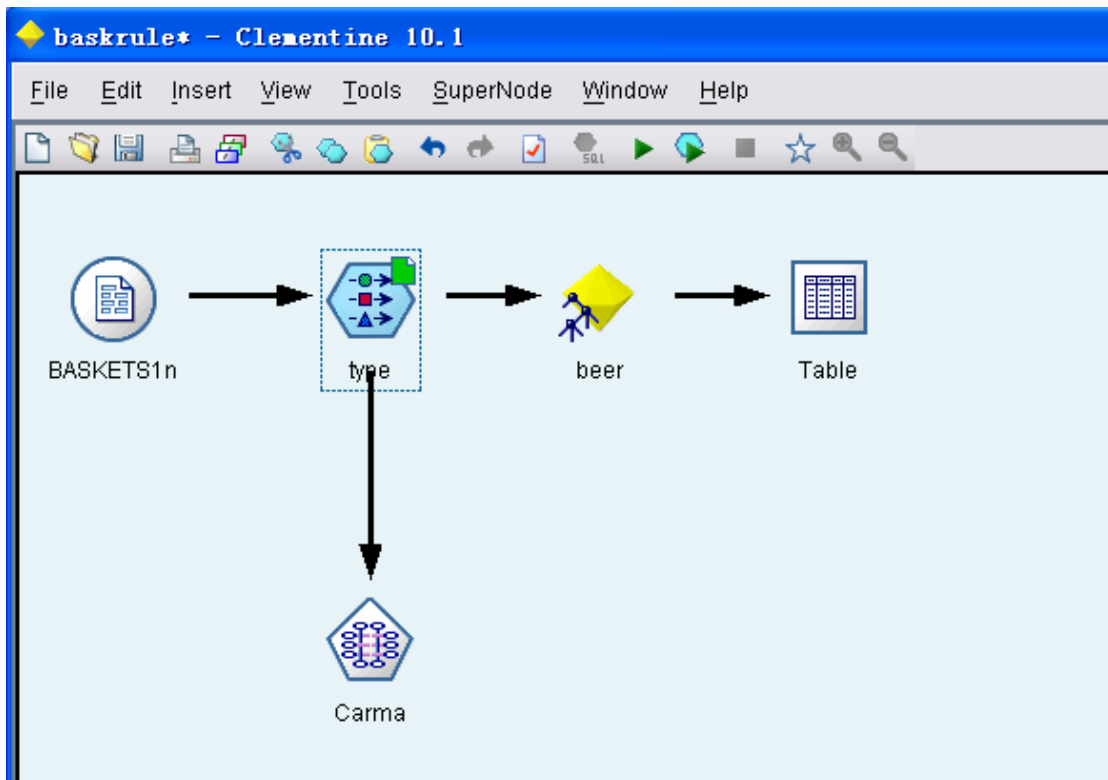
Creat node on: 指定生成的规则节点放置的位置。Canvas 就是放在中央面板上，GM Palette 就是放在右上角的模型窗口内。

Target field: 这个选项是十分重要的，它将指明需要预测的变量，进而根据这个变量来从 Carma 的规则集中提取出相应的对这个变量进行预测的规则。

Minimum instances: 这个选项为选择规则设定一个阈值，即只选择那些 instances 大于设定值的规则。

Minimum confidence: 同上，对规则的 confidence 设定一个阈值。

Default value: 设定一个预测的默认值。即规则节点判断某个观测不是指定的 consequence 时就预测为 Default value（上图中设定为了 FALSE）。下图将规则节点加入到数据流中进行预测：



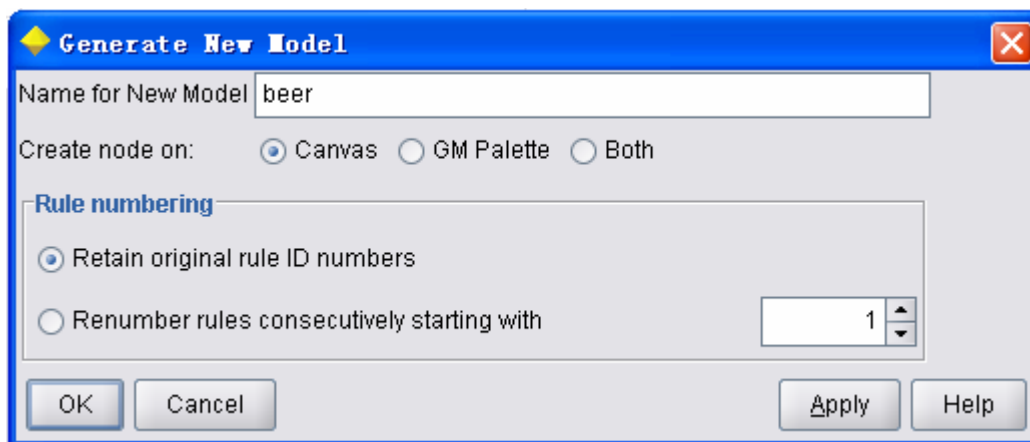
将 table 表打开查看预测的结果，如下图：

	cardid	value	pmethod	sex	ho...	income	age	fruitveg	fres...	dairy	cannedveg	ca...	frozenmeal	beer	wi...	soft...	fish	co...	\$A-beer	\$AC-beer
1	39808	42.712	CHEQUE	M	NO	27000	46	F	T	F	F	F	F	F	F	F	F	F	F	0.500
2	67362	25.357	CASH	F	NO	30000	28	F	T	F	F	F	F	F	F	F	F	T	F	0.500
3	10872	20.618	CASH	M	NO	13200	36	F	F	T	F	T	F	T	F	F	T	F	T	0.653
4	26748	23.688	CARD	F	NO	12200	26	F	F	T	F	F	F	F	T	F	F	F	F	0.500
5	91609	18.813	CARD	M	YES	11000	24	F	F	F	F	F	F	F	F	F	F	F	F	0.500
6	26630	46.487	CARD	F	NO	15000	35	F	T	F	F	F	F	F	T	F	T	F	F	0.500
7	62995	14.047	CASH	F	YES	20800	30	T	F	F	F	F	F	F	F	T	F	F	F	0.500
8	38765	22.203	CASH	M	YES	24400	22	F	F	F	F	F	F	T	F	F	F	F	F	0.500
9	28935	22.975	CHEQUE	F	NO	29500	46	T	F	F	F	F	T	F	F	F	F	F	T	0.563
10	41792	14.569	CASH	M	NO	29600	22	T	F	F	F	F	F	F	F	F	T	F	F	0.500
11	59480	10.328	CASH	F	NO	27100	18	T	T	T	F	F	F	F	T	F	T	F	T	0.551
12	60755	13.780	CASH	F	YES	20000	48	T	F	F	F	F	F	F	F	F	T	F	F	0.500
13	70998	36.509	CARD	M	YES	27300	43	F	T	F	F	T	T	F	F	F	T	F	T	0.563
14	80617	10.201	CHEQUE	F	YES	28000	43	F	F	F	F	F	F	F	F	F	T	F	F	0.500
15	61144	10.374	CASH	F	NO	27400	24	T	F	T	F	F	F	F	F	T	T	F	F	0.500
16	36405	34.822	CHEQUE	F	YES	18400	19	F	F	F	F	T	F	T	F	T	F	F	T	0.563
17	76567	42.248	CARD	M	YES	23100	31	T	F	F	T	F	F	F	F	F	T	F	T	0.551
18	85699	18.169	CASH	F	YES	27000	29	F	F	F	F	F	F	F	F	F	T	F	F	0.500
19	11357	10.753	CASH	F	YES	23100	26	F	F	F	F	F	F	T	F	F	T	F	F	0.500
20	97761	32.318	CARD	F	YES	25800	38	T	F	F	T	F	F	F	T	F	T	T	T	0.551
21	20362	31.720	CASH	M	YES	25100	38	F	F	F	F	F	T	F	F	F	T	F	T	0.563
22	33173	36.833	CASH	F	YES	24700	43	F	F	F	F	F	F	F	T	F	F	T	F	0.500
23	69934	31.179	CHEQUE	F	YES	21300	41	F	F	F	F	F	F	F	F	F	T	F	F	0.500
24	14743	21.681	CASH	M	YES	12400	48	T	T	T	T	T	T	T	T	F	F	F	T	0.653
25	83071	29.854	CASH	M	YES	18100	31	F	F	F	F	F	F	T	F	F	F	T	F	0.500
26	17571	15.270	CARD	F	YES	22900	23	T	F	F	F	T	F	T	F	F	F	F	T	0.563
27	37917	32.232	CHEQUE	F	NO	27000	32	F	F	F	F	F	F	F	T	F	F	T	F	0.500
28	11236	42.567	CARD	M	YES	26800	34	F	F	F	F	F	F	F	F	F	F	F	F	0.500
29	47914	44.591	CASH	F	YES	24700	32	F	T	F	F	F	F	F	T	F	T	T	F	0.500
30	58154	49.137	CHEQUE	M	NO	21300	50	F	F	F	F	T	F	F	F	F	F	F	F	0.500
31	35197	40.340	CASH	M	NO	27400	38	F	F	F	F	T	F	T	T	F	T	F	T	0.563
32	64892	39.000	CASH	F	YES	12900	46	F	F	F	F	F	T	F	T	F	F	F	T	0.563

表中最后两列即为该规则节点的预测结果，其中\$A-beer 为预测的结果，\$AC-beer 为该预测的置信度。

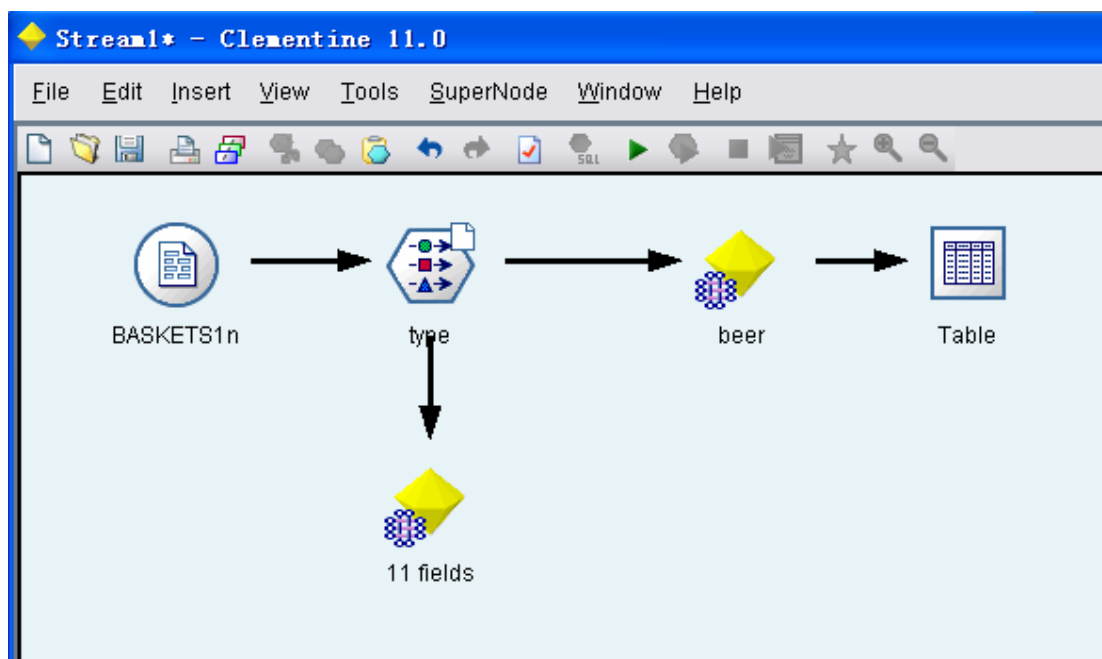
Filtered Model: 选择这一项可以生成一个含有该规则显示窗口中所有规则的 Carma 模型。Filtered Model 可以用选择的规则集来对某些特定的 antecedent

来预测其 consequent，而不用对每行观测都做预测。生成节点 Filtered Model 的设置窗口如图所示，它包含的规则为规则显示窗口（Model 选项卡）中的所有规则。



其中各个参数的设定和前面的节点是一样的。

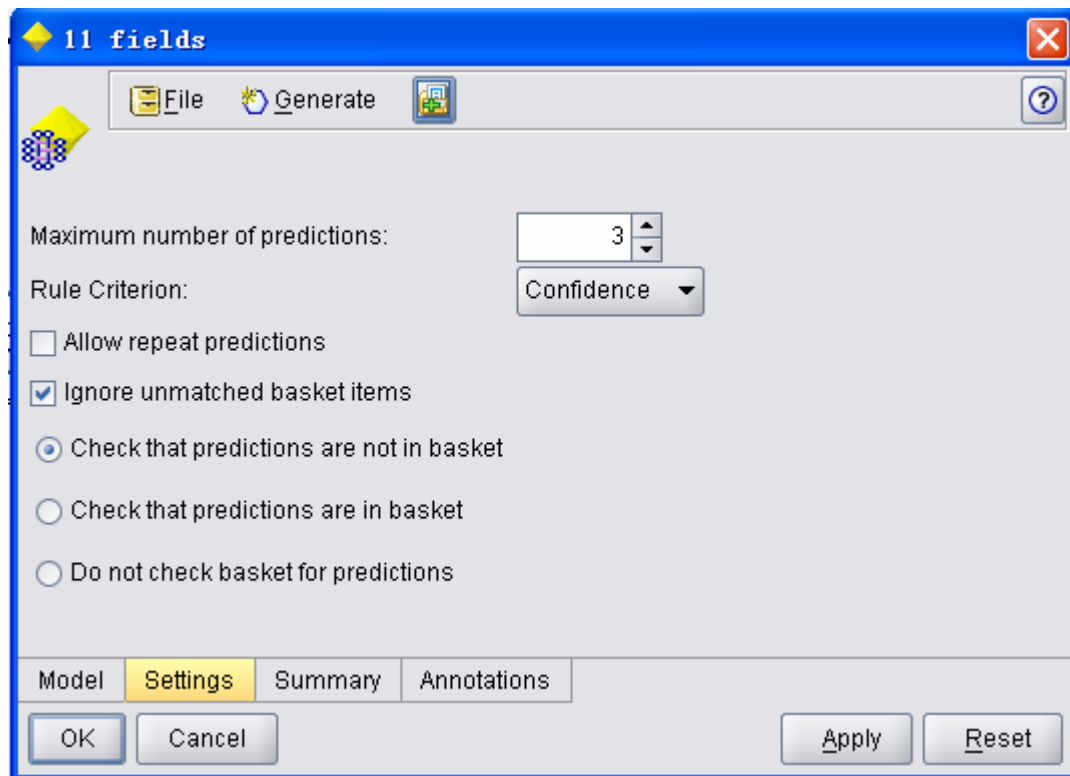
create a filtered model: 从关联规则模型中可以生成过滤节点，它包括了当前显示的所有规则。



用 CARMA 模型进行打分

Association Rule Model Settings:

当把生成的 CARMA 模型加入数据流后，可以打开模型对其进行编辑，其中 Settings 选项卡可以别用设置打分选项的细节。如图：



Maximum number of predictions: 该选项被用来设定对每一行数据所作预测的最大个数, 因为对于每一个数据行都可以有使用很多规则对其进行预测得出很多不同的结果, 该选项可以对所作预测的数据做限定, 使系统仅输出 confidence, support, lift 等最大的前几个结果结果, 该项一般和下面的 Rule Criterion 结合使用。

Rule Criterion: 该选项选择将预测结果作排序的标准, 和 Maximum number 哦发 predictions 结合使用, 标准包括: Confidence、Support、Rule support、(Support * Confidence)、Lift 以及 Deployability。

Allow repeat predictions: 设定是否允许输出相同的预测值, 因为通常都有很多条规则的 consequent 是一样的, 选择该选项就可以允许系统使用多条规则进行预测, 并且输出相同的结果。比如: 若选择此项, bread & cheese ⇨ wine 及 cheese & fruit ⇨ wine 该两条规则的结果都会被输出。

Ignore unmatched basket items: 选择该项可以在每一行包含的项目比规则中的 antecedent 要多时, 仍然可以使用该规则进行预测。比如: 如果某一数据行包括了 tent & sleeping bag & kettle, 那么如果选择该选项的话, 尽管 kettle 并没有包含在规则的 antecedent 集合中, 但是系统仍可以对其用规则 tent & sleeping bag ⇨ gas_stove 来进行预测。

Check that predictions are not in basket: 选定该选项, 系统会将 consequents 包含在数据行中的规则删去。比如, 用规则对数据行进行打分的目的是为了给客户推荐想买的家具, 那么如果该名顾客已经买了一个餐桌的话, 那么他基本上不可能再买餐桌了, 此时如果再给他购买餐桌的建议就显得可笑而无用论。选定该项可以避免这种情况的发生。

Do not check basket for predictions: 选择该选项可以在打分时使用所有的规则。

CARMA 模型与其它关联模型比较:

1. 算法上的区别

传统的关联规则算法，如 APRIORI、GRI、DIC 等从本质上说都是一种本地、离线的关联规则算法，它们需要有一个本地的完整的目标数据集，指定一个固定的支持度，然后对数据及进行多遍扫描来构造满足支持度的规则。但是 CARMA 算法是一种在线的关联规则算法，它可以以网络上不断产生实时交易流为数据源，通过一遍扫描，最多两遍来构造满足给定支持度的规则集，而且 CARMA 算法允许在扫描过程中，用户可以对支持度进行修改。

在内存的使用上，CARMA 算法比其他关联规则算法更具优势，它在执行过程中需要的内存要比其它关联规则算法少很多。但是速度稍慢是 CARMA 算法的一个缺点。

2. Clementine 中操作的区别

和其它的关联规则模型不同（比如 GRI、APRIORI），CARMA 模型不需要指定 IN 和 OUT 变量，它相当于将所有字段设置成 Both 建立 APRIORI 模型时的情况。如果还想找出对应于指定的商品或服务更合适的促销商品或服务的话，可以通过在结果集中去选择那些 antecedents 是你所指定的商品或服务的那些规则，通过查看它们对应的 consequent 来找出合适的促销品。

和 GRI、APRIORI 比起来，CARMA 模型的支持度（support）是对整个规则集而言的（支持度不仅对于 antecedent 有用，而且也对 consequent 有效），而不像 GRI、APRIORI 那样仅对 antecedent 做限制。另一点不同是，在 CARMA 中允许生成的规则集中的每条规则可以同时包含多个 consequents，在 GRI、APRIORI 中通常只能有一个 consequents。

Apriori 及 CARMA 等规则模型可以直接被嵌入数据流中对数据进行打分（预测），但是 GRI 不可以，不过可以先从 GRI 模型中生成一个 Ruleset，然后利用这个 ruleset 进行打分。