

数据挖掘-FP-Growth

生活中的数据本身包含着各种联系，大数据的出现给我们带来了一种新的思维方式，那就是跳出先因后果的思维方式，用大量的数据统计结果去倒推原因，找到无关因素间的隐蔽联系，例如众人皆知的营销案例“啤酒和尿布”。本文将介绍目前业界经典关联规则算法 FP-Growth，借助分析购物篮中商品数据，找出商品之间的关联关系，帮助用户挖掘出数据最大的价值，获得更大的经营收益。

FP-Growth 简介

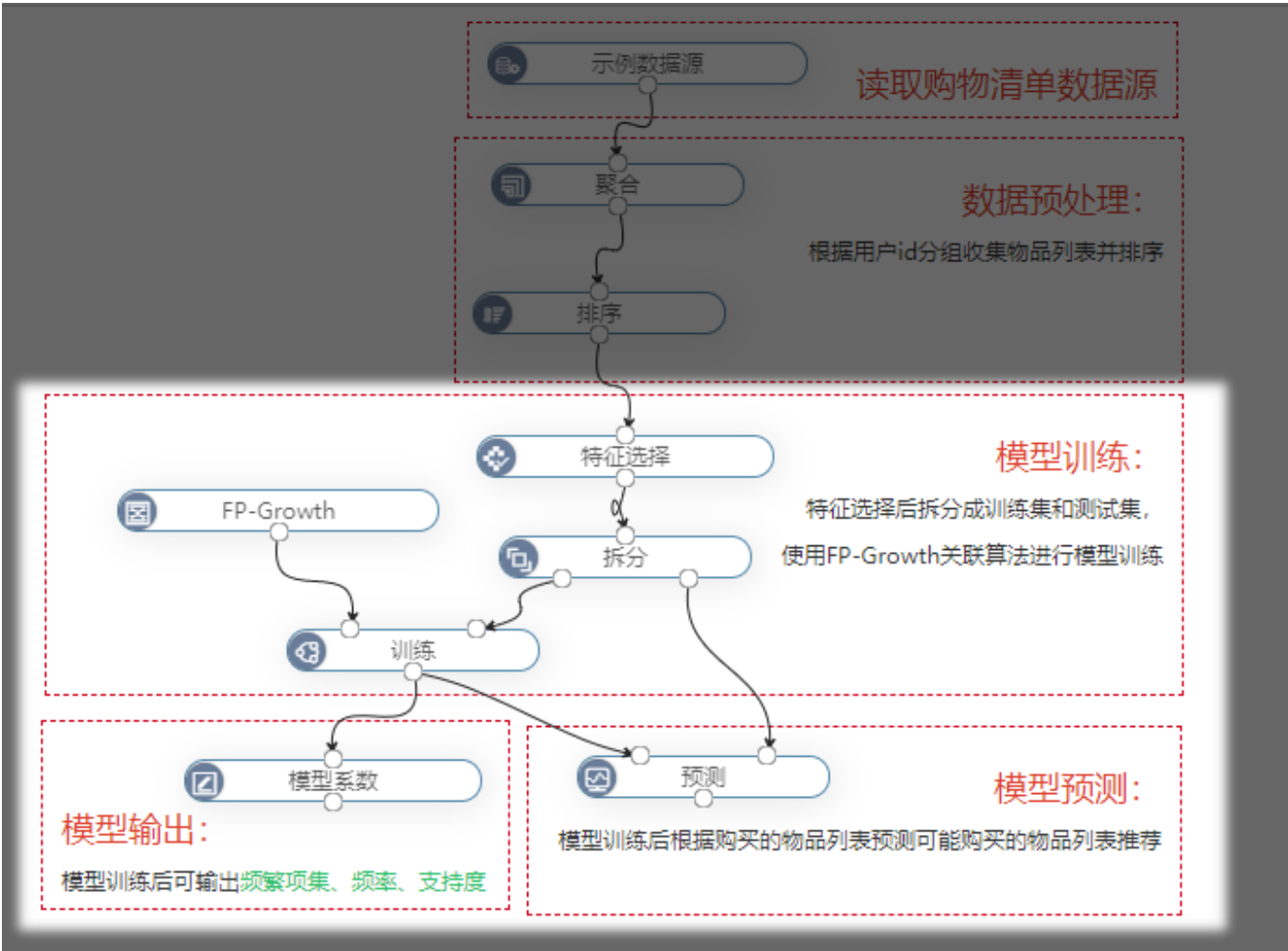
FP-Growth (Frequent Pattern Growth) 作为一种经典的关联规则算法，可以高效自动地从数据中挖掘出频繁集合，进而挖掘出潜在的关联规则，这些在企业营销决策的制定过程中具有重要的参考价值。另外，FP-Growth采用高级的数据结构，使其在大数据集上处理效率显著优于其他算法 (Apriori)，从而加快整个数据挖掘的过程。

应用场景

关联规则的挖掘算法，典型应用是零售业的购物篮分析，挖掘交易数据中的伴随购买、购买推荐关系链等商机，有助于制定营销策略。

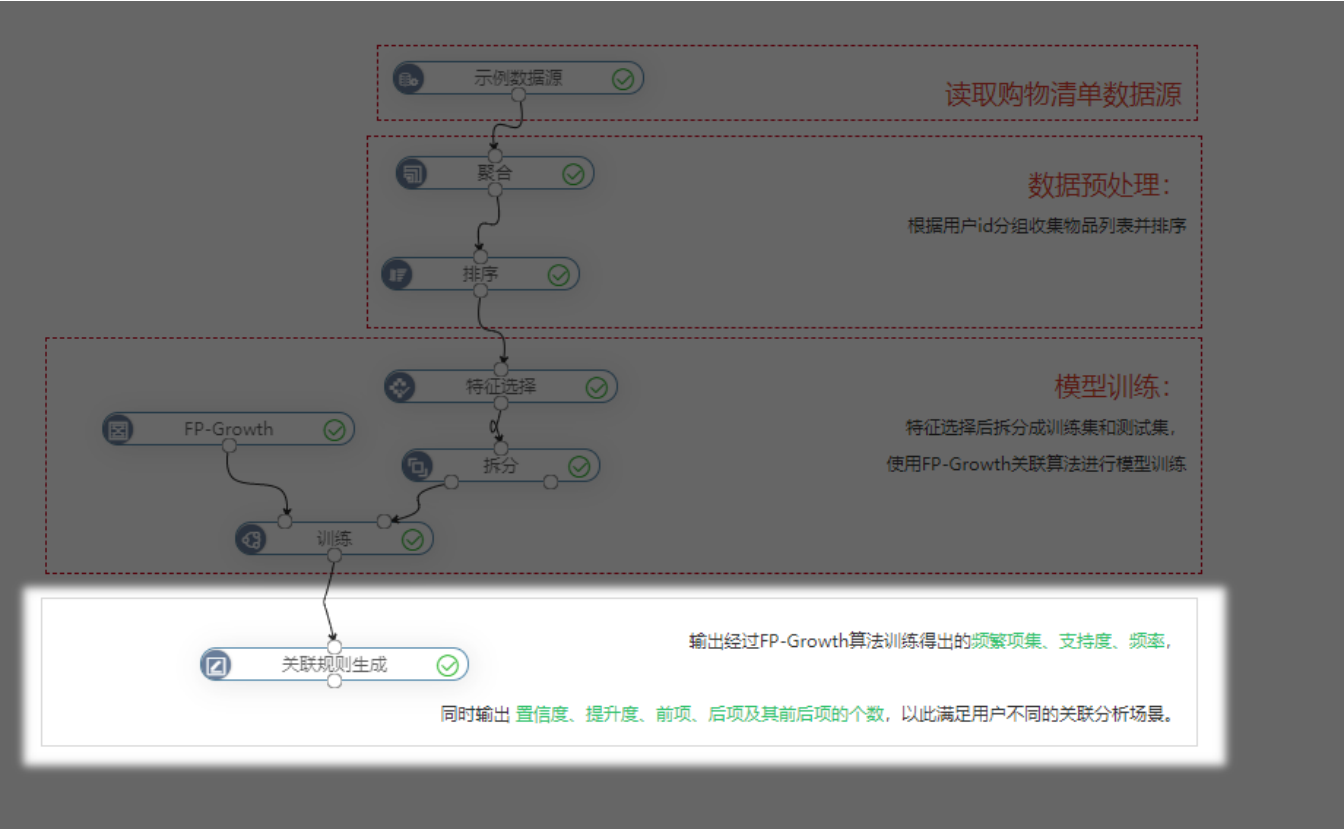
应用案例

(1) 我们可以基于产品内置的购物清单数据，针对购物篮商品利用FP-Growth算法模型输出相关关联分析指标（频繁项集、频率、支持度），同时结合预测功能得出用户可能购买的物品，请点击[购物篮分析](#)查看示例详情。



A: 频繁项集	# 频率	# 支持度
蔬菜罐头	219	0.33282674772036475
冻肉	217	0.32978723404255317
鱼类	207	0.31458966565349544
啤酒	207	0.31458966565349544
水果	202	0.3069908814589666
白酒	193	0.2933130699088146
糖果	193	0.2933130699088146
肉罐头	145	0.22036474164133737
冻肉,蔬菜罐头	130	0.19756838905775076
软饮料	130	0.19756838905775076
鲜肉	128	0.1945288753799392
啤酒,蔬菜罐头	122	0.18541033434650456
乳制品	122	0.18541033434650456
啤酒,冻肉	120	0.182370820668693
啤酒,冻肉,蔬菜罐头	107	0.16261398176291794
水果,鱼类	100	0.1519756838905775

(2) 基于产品内置的实验案例“购物篮分析”，结合 关联规则生成 功能，也可轻松输出更详尽的关联分析指标（置信度、提升度、前项、后项等）。



当前显示 18 条 / 总共有 18 条数据

频繁项集	频率	支持度	前项	后项	置信度	提升度	前项数量	后项数量
^ Items	# freq	# support	^ antecedent	^ consequent	# confidence	# lift	# antecedent_quantity	# consequent_quantity
冻肉,啤酒	120	0.182370820668693	冻肉	啤酒	0.5529953917050692	1.7578307620383358	1	1
冻肉,啤酒	120	0.182370820668693	啤酒	冻肉	0.5797101449275363	1.7578307620383358	1	1
蔬菜罐头,冻肉,啤酒	107	0.16261398176291794	蔬菜罐头,冻肉	啤酒	0.823076923076923	2.6163507989594943	2	1
蔬菜罐头,冻肉,啤酒	107	0.16261398176291794	蔬菜罐头,啤酒	冻肉	0.8770491803278688	2.6594394500264413	2	1
蔬菜罐头,冻肉,啤酒	107	0.16261398176291794	冻肉,啤酒	蔬菜罐头	0.8916666666666667	2.6790715372907155	2	1
蔬菜罐头,冻肉,啤酒	107	0.16261398176291794	蔬菜罐头	冻肉,啤酒	0.4885844748858447	2.6790715372907155	1	2
蔬菜罐头,冻肉,啤酒	107	0.16261398176291794	冻肉	蔬菜罐头,啤酒	0.4930875576036866	2.659439450026441	1	2
蔬菜罐头,冻肉,啤酒	107	0.16261398176291794	啤酒	蔬菜罐头,冻肉	0.5169082125603864	2.6163507989594943	1	2
蔬菜罐头,啤酒	122	0.18541033434650456	蔬菜罐头	啤酒	0.5570776255707762	1.7708071382877815	1	1
蔬菜罐头,啤酒	122	0.18541033434650456	啤酒	蔬菜罐头	0.5893719806763285	1.7708071382877815	1	1
鱼类,水果	100	0.1519756838905775	鱼类	水果	0.4830917874396135	1.5736356244320082	1	1
鱼类,水果	100	0.1519756838905775	水果	鱼类	0.49504950495049505	1.5736356244320084	1	1
蔬菜罐头,冻肉	130	0.19756838905775076	蔬菜罐头	冻肉	0.593607305936073	1.7999705405803506	1	1
蔬菜罐头,冻肉	130	0.19756838905775076	冻肉	蔬菜罐头	0.5990783410138248	1.7999705405803503	1	1
糖果,白酒	94	0.14285714285714285	糖果	白酒	0.48704663212435234	1.6605009530457193	1	1
糖果,白酒	94	0.14285714285714285	白酒	糖果	0.48704663212435234	1.6605009530457193	1	1
蔬菜罐头,白酒	72	0.1094224924012158	蔬菜罐头	白酒	0.3287671232876712	1.1208744410533038	1	1
蔬菜罐头,白酒	72	0.1094224924012158	白酒	蔬菜罐头	0.37305699481865284	1.120874441053304	1	1

表头真名 表头别名 提示: 点击单元格可查看超出的内容。注意: 表头中 表示特征列。* 表示标准列

下载预览数据

相关概念

序号	概念	定义	说明	业务意义
1	频繁项集	频繁项集是指满足支持度最小阈值的所有项集。	项,是指我们分析数据中的一个对象;项集,就是若干项的项构成的集合,如集合{牛奶、麦片、糖}是一个3项集。	频繁项集表示频繁地并且同时出现在交易数据集中的商品的集合,可用于发现共现事项或者发现事物的热点信息。

2	支持度 Support	支持度是指几个频繁关联的数据在数据集中出现的次数占总数据集的比重。	假设数据集有 10 条记录，包含{ ‘牛奶’， ‘面包’ } 的有 5 条记录，那么{ ‘牛奶’， ‘面包’ } 的支持度就是 $5/10 = 0.5$ 。	支持度表示在关联规则中出现的频率，支持度高说明物品的需求量大。	在实际零售场景中，可通过支持度，置信度，提升度综合指标来看待商品组合，发现高价值关联商品，有助于提升客单价。
3	置信度 Confidence	置信度是指出现A后又出现B的条件概率，相当于A和B同时出现的概率占A出现概率的比值。	假设今天共有10笔订单，其中购买A的次数是8，同时购买A和B的次数是6，则其置信度是 $6/8=75\%$ 。 从关联规则的可信程度角度来看，“购买A的顾客又会购买B”这个商业推测，有75%的可能性是成立的。	置信度体现了衡量关联规则的可靠性，置信度高说明物品之间连带紧密。	
4	最小支持度	最小支持度作为支持度的阈值，满足最小支持度的项集才会输出	最小置信度、最小支持度只是在训练模型时起到筛选频繁项集的作用。		
5	最小置信度	最小置信度作为置信度的阈值，满足最小置信度的项集才会输出			