Association Analysis of Supermarket Products Based on RapidMiner

Dongxiu Ye

Heyuan Polytechnic, China
Yedongxiu1988@163.com

Abstract:
With the development of information technology and database technology, information data has shown explosive growth. However, in the face of huge amounts of data, people are often confused and unable to discover the hidden knowledge and rules of the data, resulting in a waste of resources. This article uses the RapidMiner tool to analyze the data of the supermarket, analyze the products, and give examples of modeling and marketing suggestions, which has strong practical value.
Keywords: *association rules, bundle sale, supermarket, marketing*

基于 RapidMiner 的超市商品的关联分析

叶冬秀

河源职业技术学院，河源，广东，中国
Yedongxiu1988@163.com

中文摘要：
随着信息技术和数据库技术的发展，信息数据呈现爆炸式增长。然而面对海量的数据，人们往往无所适从，无法发现数据隐藏的知识与规则，从而导致资源的浪费。本文运用 RapidMiner 工具，分析了超市的数据，对产品进行关联分析，并给出建模和营销建议的实例，具有很强的实用价值。
关键词：关联规则　捆绑销售　超市　市场营销

1. 引言

数据挖掘（Data Mining），又称数据库中的知识发现（Knowledge Discovery in Database）[1]，顾名思义就是从大量的数据中挖掘出有用的信息。在最近几年里已被数据库界所广泛研究，其中关联规则（Association Rules）的挖掘是一个重要的问题。数据挖掘随着计算机应用的越来越广泛，每年都要积累大量的数据,运用数据挖掘技术，在这些数据当中我们可以找出金子来[2]。本文基于商家收集到顾客购买情况的数据，使用 RapidMiner 工具，来分析客户的购买行为及购买模式，挖掘客户购买产品的关联数据。本文的研究不仅验证了相关模型的可行性和准确性，也是对关联分析理论在超市数据中应用的一种重要探索，具有一定的理论意义和现实意义[3]。

2. RapidMiner 简介及为什么选择 RapidMiner？

2.1 RapidMiner 简介

Rapidminer 是世界领先的数据挖掘解决方案和完善的商业分析平台，是进行数据挖掘、文本挖掘和预测分析的强大工具。它提供大量各种各样的描述性技术和预测性技术，能简化数据挖掘过程的设计和评价，能从数据中能洞察商机，帮助做出盈利的决策。

2.2 为什么选择 RapidMiner

RapidMiner 不需要软件许可证费用，是中小型企业买得起的灵活的服务支持选项。另外，RapidMiner 还拥有最快的开发过程（即使针对非常复杂的数据挖掘流程）和操作可靠性的保证，被广泛认为是市场上最容易理解

的最灵活的数据挖掘解决方案。

RapidMiner 还有五方面的优势，优势一：有条理合逻辑的图形用户界面。RapidMiner 为分析进程的设计提供了一个强大而直观的图形用户界面。跟传统的 ETL（数据抽取、转换和加载）工具相比，RapidMiner 采用了更前卫的方式：每个转换部件、每个可视化部件、每个分析部件、每个预测部件，甚至报表中的每个显示部件都可以集成到一个进程中。这意味着所有任务都可以在一个工具中实现。更大的好处是部件之间可以互相交互，能够通过几次鼠标的点击集成到一起。即使对于复杂的元素间很多交互的商业流程，我们都可以借助 RapidMiner 简单地创建它而不需要编写任何代码； 优势二：前所未有的分析方便性和用户支持。依赖于元数据的管理和进程设计的智能分析，RapidMiner 紧密关注数据分析员的工作，并提供实时的帮助；优势三：无与伦比的分析技术的集成套装。RapidMiner 提供超过 1500 种关于数据集成、数据转换、分析、建模和可视化的方法。市场上没有其他产品能够提供如此多的方法来定义最优的分析进程。尤其是在属性选择和异常检测方面，市场上其他产品都不能提供 RapidMiner 提供的大部分方法；优势四：兼容主流标准。RapidMiner 支持多个标准，包括 PMML（预测模型标记语言），其允许 RapidMiner 与其他系统交换预测模型。优势五：丰富的扩展支持。Rapid-I 和第三方提供了大量的 RapidMiner 扩展包。它们都在 Rapid-I 市集公开发布。该市集是开源扩展和第三方私有扩展的发布平台。

3. 基于 RapidMiner 实现超市商品购买关联分析

3.1 商业理解

本文通过某超市在营业中收银机保存的顾客购买记录，来分析顾客购买的商品之间的关联性，从而为超市制定有效的营销策略提供依据。本文用 RapidMiner 工具制作了一个关联规则挖掘在零售业中的具体应用模型，这个模型主要引入了关联规则所获利润的概念讨论超市商品捆绑销售策略[4]。

3.2 数据理解

本文的数据是某超市在某段时间，1000 位顾客的购买清单。每位顾客都购买了 7 样产品，共 7000 条数据。

表 1 顾客购买清单

	A	B	C	D	E
1	CUSTOMER	TIME	PRODUCT		
2	0	0	hering		
3	0	1	corned_b		
4	0	2	olives		
5	0	3	ham		
6	0	4	turkey		
7	0	5	bourbon		
8	0	6	ice_crea		
9	1	0	baguette		
10	1	1	soda		
11	1	2	hering		
12	1	3	cracker		
13	1	4	heineken		
14	1	5	olives		
15	1	6	corned_b		
16	2	0	avocado		
17	2	1	cracker		
18	2	2	artichok		
19	2	3	heineken		
20	2	4	ham		
21	2	5	turkey		
22	2	6	sardines		
23	3	0	olives		
24	3	1	bourbon		
25	3	2	coke		
26	3	3	turkey		
27	3	4	ice_crea		
28	3	5	ham		

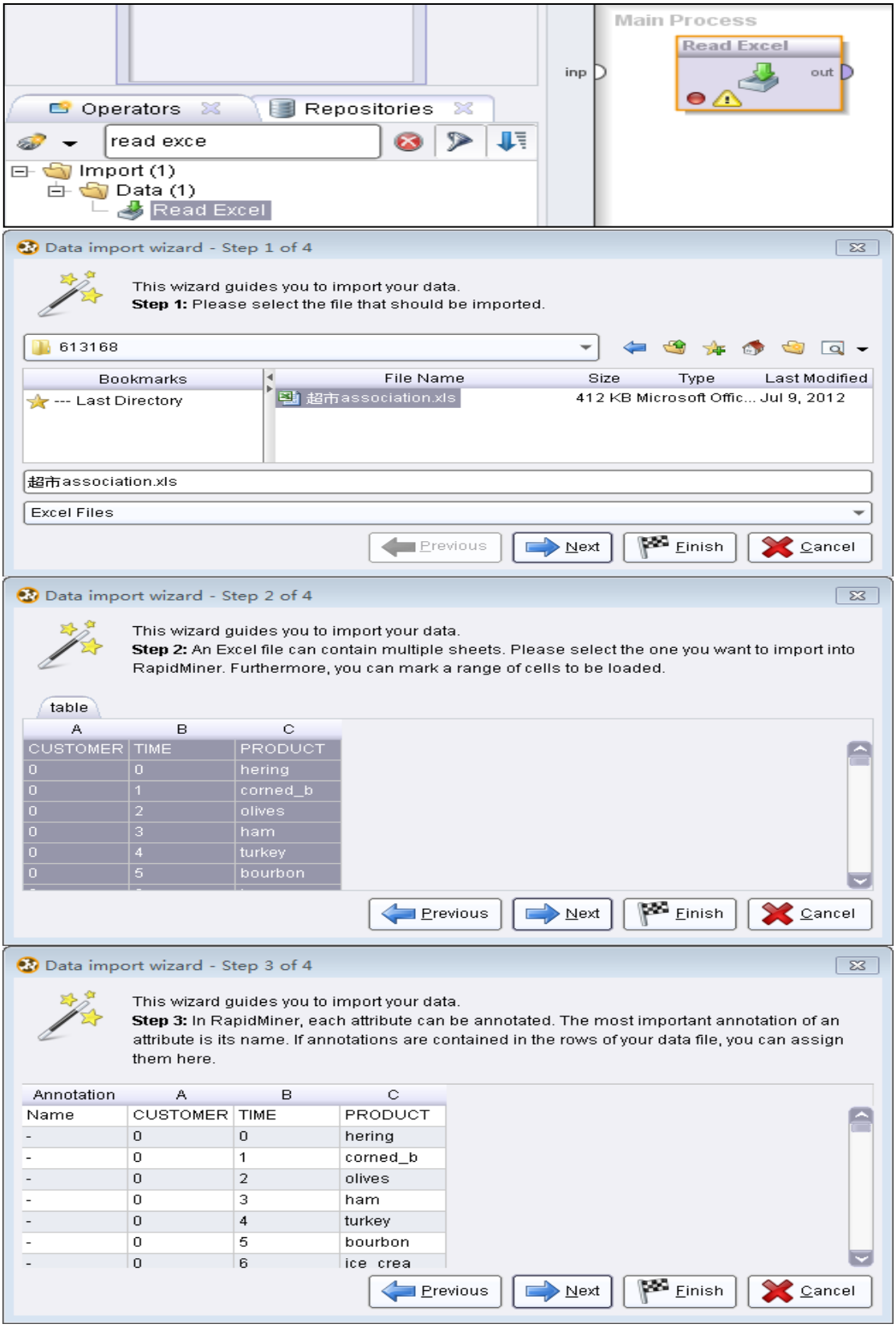
table

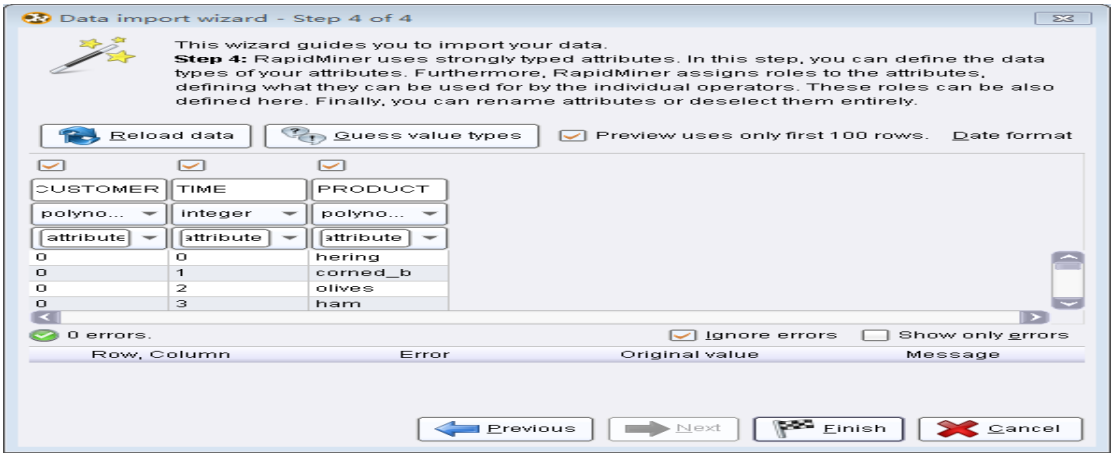
3.3 数据准备

每一行中：记录该顾客购买各个商品的情况。

3.4 建立模型与模型评估

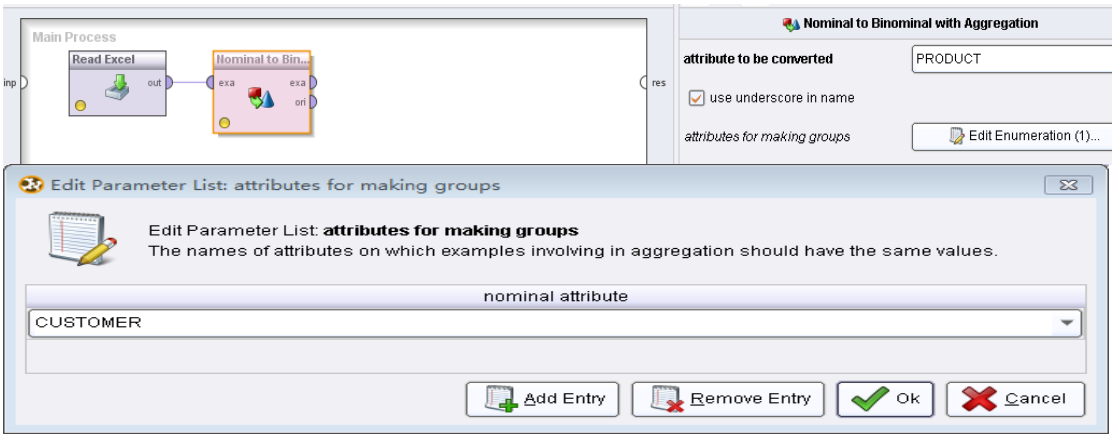
首先由 Read Excel 读入超市客户购买清单 Excel 表，步骤如下：


Operators
Repositories
read exce
Import (1)
Data (1)
Read Excel
inp
Main Process
Read Excel
out
Data import wizard - Step 1 of 4
This wizard guides you to import your data.
Step 1: Please select the file that should be imported.
613168
Bookmarks
--- Last Directory
File Name
Size
Type
Last Modified
超市association.xls
412 KB Microsoft Offic...
Jul 9, 2012
超市association.xls
Excel Files
Previous
Next
Finish
Cancel
Data import wizard - Step 2 of 4
This wizard guides you to import your data.
Step 2: An Excel file can contain multiple sheets. Please select the one you want to import into RapidMiner. Furthermore, you can mark a range of cells to be loaded.
table
A B C
CUSTOMER TIME PRODUCT
0 0 hering
0 1 corned_b
0 2 olives
0 3 ham
0 4 turkey
0 5 bourbon
Previous
Next
Finish
Cancel
Data import wizard - Step 3 of 4
This wizard guides you to import your data.
Step 3: In RapidMiner, each attribute can be annotated. The most important annotation of an attribute is its name. If annotations are contained in the rows of your data file, you can assign them here.
Annotation A B C
Name CUSTOMER TIME PRODUCT
- 0 0 hering
- 0 1 corned_b
- 0 2 olives
- 0 3 ham
- 0 4 turkey
- 0 5 bourbon
- 0 6 ice_crea
Previous
Next
Finish
Cancel

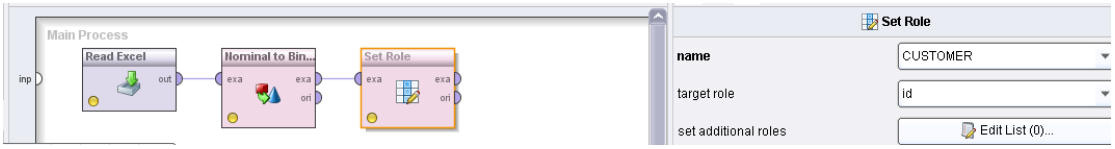


注意：第四步的 CUSTOMER 中的类型应改为 polynominal。

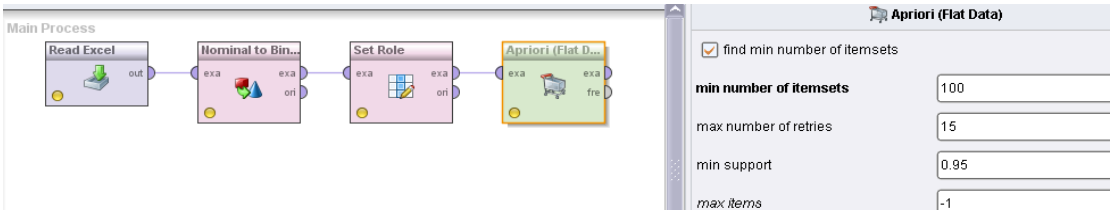
其次，引入 Nominal to Binominal with Aggregation 1）将 attribute to be converted 设成“PRODUCT” 2) 勾选“use underscore in name” 3) 点击 attributes for making groups 添加“CUSTOMER”，因此，把 Excel 中的 PRODUCT 转成购买标志。操作流程如下。



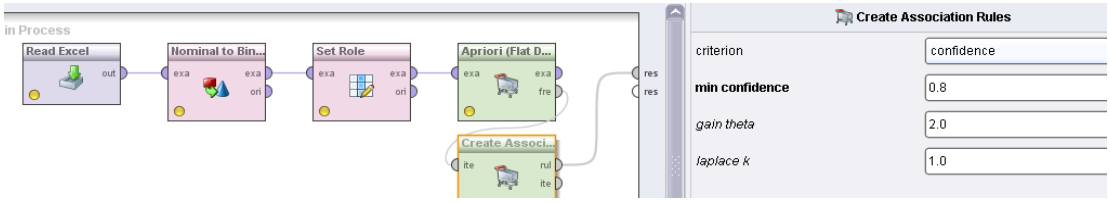
再次，引入部件 set role， 1) 将 name 设成“CUSTOMER”；2) 从 target role 设成“id”；该部件将 CUSTOMER 转变成非常规属性，不参与后续关联分析过程，注意参与关联分析的属性都是常规属性。



另外，引入部件 Apriori (Flat Data)，1) 勾选 find min number of itemsets；2) 将 min support 设成 0.95。操作图如下。



最后，为了产生关联规则，引入部件 Create Association Rules， 1) 从 criterion 中选择 confidence；2) 将 min confidence 设成 0.8，该部件产生关联规则 (置信度阈值为 80%)。图如下。



运行：把各个部件联系起来，并按运行键，就出现了运行结果。

结果产生了四条关联规则：

No.	Premises	Conclusion	Support	Confide...	LaPla...	Gain	p-s	Lift	Convi...
1	PRODUCT_soda	PRODUCT_heineken	0.257	0.808	0.954	-0.379	0.066	1.348	2.088
2	PRODUCT_artichok	PRODUCT_heineken	0.252	0.826	0.959	-0.358	0.069	1.378	2.305
3	PRODUCT_heineken, PRODUCT_soda	PRODUCT_cracker	0.234	0.911	0.982	-0.280	0.109	1.868	5.726
4	PRODUCT_cracker, PRODUCT_soda	PRODUCT_heineken	0.234	0.932	0.986	-0.268	0.083	1.555	5.915

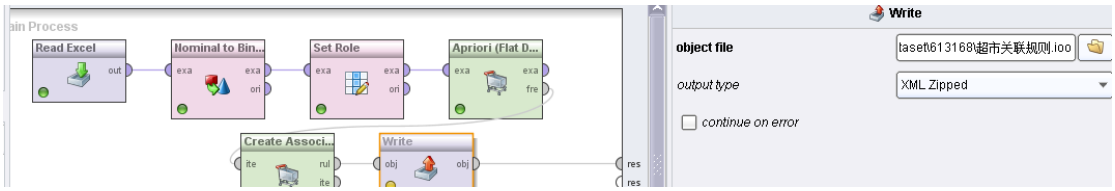
结果说明：

以 NO4 为例，对于已经买了饼干和汽水的顾客来说，他们就很有可能购买喜力啤酒。因为从分析结果来看，购买了饼干和汽水的顾客当中，有 93.2%(Confidence) 的顾客也同时购买了喜力啤酒；同时购买了饼干、汽水和喜力啤酒这三种商品的顾客占比为 23.4% (Support)。那么，采用该规则向客户推荐喜力啤酒比随机推荐喜力啤酒会提升效果 1.555 倍 (Lift)。

3.5 模型应用

（1）产生推荐规则并保存

由上，已经得出了推荐规则（即关联规则），下面只要引进一个 Write 部件，就可以把推荐规则保存下来。Write 将 object file 设成“超市关联规则”文件名，并保存关联规则集合到文件中。图如下。

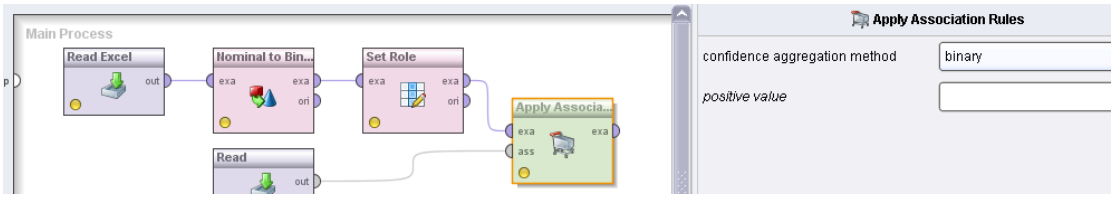


（2）对关联性强的商品做商品推荐

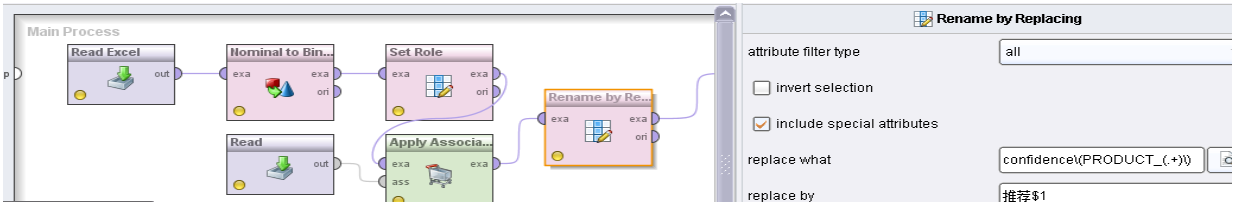
首先，引入 Read Excel、 Nominal to Binominal with Aggregation 和 Set role 部件，步骤同上。

其次，引入 Read，1）将 object file 设成“超市关联规则”文件名;2）从 io object 中选择 Association Rules; 读取已经保存的关联规则集合。

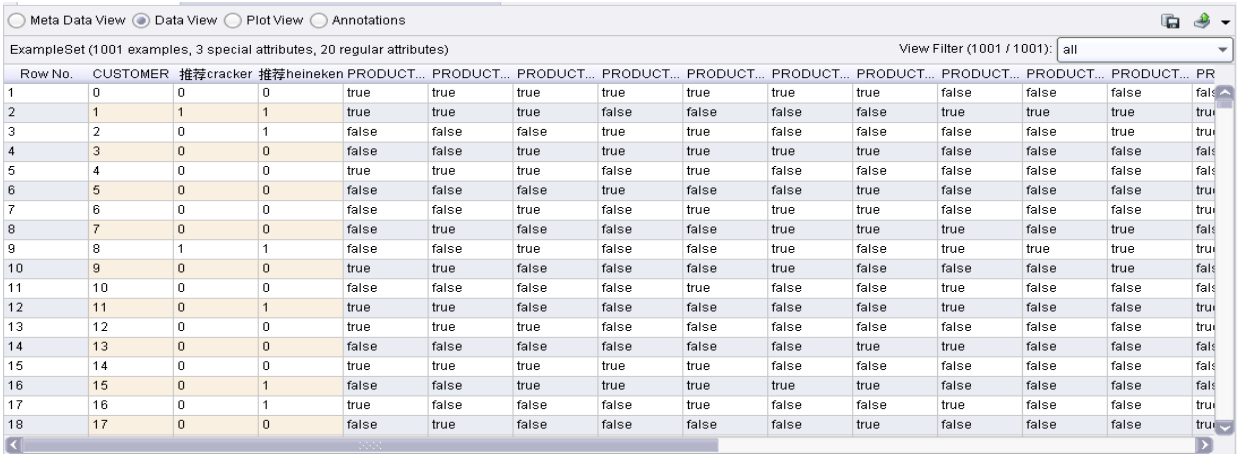
另外，引入 Apply Association Rules 部件，从 confidence aggregation method 中选择 binary，根据购买明细数据集应用关联规则集合，进行商品推荐。



最后，引入部件 Rename by Replacing， 1) 从 attribute filter type 中选择 all；2) 勾选 include special attributes； 3) 将 replace what 设成正则表达式 “confidence\(PRODUCT_(.+)\)”；4) 将 replace by 设成“推荐 $1”。该部件将形式为“confidence(PRODUCT_XXX)”的属性名改成“推荐 XXX”，这样便于理解。



（3）运行结果分析



Row No.	CUSTOMER	推荐cracker	推荐heineken	PRODUCT...	PRODUCT...	PRODUCT...	PRODUCT...	PRODUCT...	PRODUCT...	PRODUCT...	PRODUCT...	PRODUCT...	PRODUCT...	PR
1	0	0	0	true	true	true	true	true	true	true	false	false	false	fals
2	1	1	1	true	true	true	false	false	false	false	true	true	true	tru
3	2	0	1	false	false	false	true	true	false	false	false	false	true	tru
4	3	0	0	false	false	true	true	true	true	true	false	false	false	fals
5	4	0	0	true	true	true	false	true	false	false	false	false	false	fals
6	5	0	0	false	false	false	true	false	false	true	false	false	false	tru
7	6	0	0	false	false	true	false	true	true	true	false	false	false	tru
8	7	0	0	false	true	false	false	false	true	true	true	false	true	fals
9	8	1	1	false	false	true	false	false	true	false	true	true	true	tru
10	9	0	0	true	true	false	false	false	true	false	false	false	true	fals
11	10	0	0	false	false	false	false	true	false	false	true	false	false	fals
12	11	0	1	true	true	false	false	false	false	false	true	false	false	tru
13	12	0	0	true	true	true	false	false	false	false	false	false	false	tru
14	13	0	0	false	false	false	false	false	false	true	true	false	false	fals
15	14	0	0	true	true	true	true	true	false	false	false	false	false	fals
16	15	0	1	false	false	true	true	true	true	true	false	false	false	fals
17	16	0	1	true	false	false	false	true	false	false	true	false	false	tru
18	17	0	0	false	true	false	false	false	false	true	false	false	false	tru

以 ROW NO.16 和 17 为例，向顾客推荐 heineken 产品都显示为 1，即可向这两个顾客推荐 heineken 产品。从实际数据来看，顾客 16 是没有购买 heineken，而顾客 17 实际上是有买产品 heineken 的。因此，关联分析，就可以给顾客推荐顾客潜在想购买的商品。

注意，推荐 cracker 和推荐 heineken 全为 1 的客户，不推荐任何商品，因为该在实际情况下，两样产品都购买了。

4. 总结超市如何运用数据结果

基于顾客的购买清单，使用 RapidMiner 工具，我们做出了关联分析，并给出以下三点的建议。首先，超市可以把相关性强的商品，布置在相近的地方，达到优化商品布局的目的。其次，对于关联性强的商品，可以进行捆绑销售并对于捆绑的商品给与一定的优惠，从而设计促销方案。最后，在顾客进行购物的时候，销售人员还可以给与顾客感兴趣的产品做快速商品推荐。通过吸引更多的顾客，提高顾客的光顾次数，并且提高每个顾客的单次的消费金额，超市才得以获取更多的利润总额[4]。因此本文提出的超市商品捆绑销售模型具有一定的实用意义。

REFERENCES

[1] Hongxia Cheng, Research on Data Mining Algorithms Based on Association Rules, Computer Knowledge and Technology, 2007.

[2] Ying Peng, Data Mining Overview, Journal of Dehong Teachers’ College, 2009.

[3] Fang YU, The application research of association

rules mining in supermarket marketing analysis, Harbin Institute of Technology, 2010.

[4] Fang YU, Application of Association Analysis in Supermarket Bundle Sales, Market Modernization, 2010.