利用历史购物数据分析超市商品摆放规则

小组成员： 康杨，刘元培，高建花，姜艳杰

利用历史购物数据分析超市商品摆放规则

摘要：随着信息化建设和应用进程的加快，商家的数据库中积聚了大量的销售信息。关联分析可以从销售数据库中找出各商品之间潜在的关联关系。利用数据挖掘技术，针对应用于超市销售关联规则的特点，能够从整体、宏观的角度去分析、解决问题，以帮助企业了解销售产品品类间的联系，进行优化产品品类配置，同时能够帮助企业的决策者根据发展形势的变化及历史数据做出正确的决策。

1. 背景

当今社会已经步入大数据时代，大量信息已经成为信息社会最重要的特征。如何更好地利用信息，如何从海量数据中发现知识创造价值是人类面对的一个重要课题。近年来我国的销售行业发展迅速，越来越多的大型企业或超市开始将目光投向数据挖掘技术，有效的利用数据挖掘技术为企业提供信息是各大零售巨头核心竞争力的重要组成部分。

数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘是一门交叉学科，它的目标是利用算法，从数据中抽取模式，将大容量数据转换为有用的知识和信息。它与传统的数据分析，如报表、查询、统计分析等数据分析考核的本质区别是它是没有明确假设的前提下去挖掘信息、发现知识。数据挖掘把人们对数据的应用从低层次的简单查询，提升到从数据中挖掘知识，提供决策支持。

关联分析的目的是为挖掘出隐藏在数据间的相互关系，从而促进信息的显化。关联规则是关联分析中的一种常用技术。关联规则是反应两项或多项属性之间存在的相关性，其目的是为了从一些项的属性值来预测其他属性值，同时可以应用推广所反映一个事件和其他事件之间依赖或者关联的知识。最为著名的关联规则挖掘方法是由美国 R.Agrawal 提出的 Apriori 算法。

要解决购物数据分析的问题，可以用数据挖掘中的关联规则技术。即通过量化的数字，描述产品 A 的出现对产品 B 的出现有多大影响。关联分析就是给定一组 Item 和一个记录集合，通过分析记录集合，推到出 Item 间的相关性。可以用四个属性来描述关联规则：

1） 可信度：在产品集 A 出现的前提下，B 出现的概率

2） 支持度：产品集 A、B 同时出现的概率

3) 期望可信度：产品集 B 出现的概率

4) 作用度可信度：对期望可信度的比值

目前大多数的关联分析都基于“支持度-置信度”的框架，其目的是抽取形如“if A then B”的规则。支持度高表示规则经常被使用，置信度高表示规则比较可靠，通过关联分析后可以得到关于产品 A、B 的关联程度，从而决定这两种货品在货架上的配置。

2. 问题描述

目前，商家往往不能根据累计的顾客购物信息获取有效的研究信息，更少有商家可以利用数据挖掘的方式从海量数据中获取有利的营销策略。如何从大量的购物数据中获取商品关联规则并令商家利益趋于最大化，则是我们本次实验的研究重点。

从网络上下载购物篮数据，对其中购物数据的基本属性如众数、中位数等进行统计分析，并依照购买顾客进行简单分类，完成初步的数据准备。经过参数训练，得到商品数据关联规则，估计满足商家利益最大化的摆放方式。

本实验采用了 Apriori 算法来测量测试样本中的关联规则，实验流程如下：

- 寻找开放数据集确定评测指标。
- 数据分析和数据预处理。
- 用 Apriori 算法进行实验，分析利润影响因素。

3. 方法

3.1 数据获取及预处理

使用 kaggle 数据集网站中的 Instacart Market Basket Analysis 数据集，目前正对该数据集进行预处理，提取出 product，aisle 和 department 三类数据集，从而进行对单类商品关联规则及大类商品的关联规则挖掘。各个数据集的商品信息如下图。

	A	B	C
1	order_id	product_id	name
2	1	49302	Bulgarian Yogurt
3	1	11109	Organic 4% Milk Fat Whole M
4	1	43633	Lightly Smoked Sardines in
5	1	22035	Organic Whole String Cheese
6	36	19660	Spring Water
7	36	43086	Super Greens Salad
8	36	46620	Cage Free Extra Large Grade
9	36	34497	Prosciutto, Americano
10	36	48679	Organic Garnet Sweet Potato
11	36	46979	Asparagus
12	38	21616	Organic Baby Arugula
13	96	20574	Roasted Turkey
14	96	40706	Organic Grape Tomatoes
15	96	27966	Organic Raspberries
16	96	24489	Organic Whole Strawberries

图 1:product 数据集

	A	B	C	D
1	order_id	aisle_id	name	
2	1	120	yogurt	
3	1	108	other creams cheeses	
4	1	95	canned meat seafood	
5	1	21	packaged cheese	
6	36	115	water seltzer sparkli	
7	36	123	packaged vegetables f	
8	36	86	eggs	
9	36	96	lunch meat	
10	36	83	fresh vegetables	
11	36	83	fresh vegetables	
12	38	123	packaged vegetables f	
13	96	96	lunch meat	
14	96	123	packaged vegetables f	
15	96	123	packaged vegetables f	

图 2：aisle 数据集

	A	B	C
1	order_id	departmer	name
2	1	16	dairy eggs
3	1	16	dairy eggs
4	1	15	canned goods
5	1	16	dairy eggs
6	36	7	beverages
7	36	4	produce
8	36	16	dairy eggs
9	36	20	deli
10	36	4	produce
11	36	4	produce
12	38	4	produce
13	96	20	deli
14	96	4	produce
15	96	4	produce

图 3：department 数据集

3.2 模型选取

经过分析和讨论，本实验采用经典算法 Apriori 算法。Apriori 算法的基本思想是，首先找出所有大于最小支持度的频繁项集，然后由频繁项集产生关联规则，这些规则满足最小支持度和最小可信度。Apriori 算法是用来发现频繁项集的一种方法。Apriori 算法的两个输入参数分别是最小支持度和数据集。该算法首先生成所有单个物品的相集列表，遍历之后去掉不满足最小支持度要求的项集；接下来对剩下的集合进行组合生成包含两个元素的项集，去掉不满足最小支持度的项集；重复该过程直到去掉所有不满足最小支持度的项集。其步骤是：依据支持度找出所有频繁项集（频度），依据置信度产生关联规则（强度），根据最后产生的关联规则，并考虑到利润因素，辅助商家做出商品的营销策略。Apriori 算法流程图如下图所示。

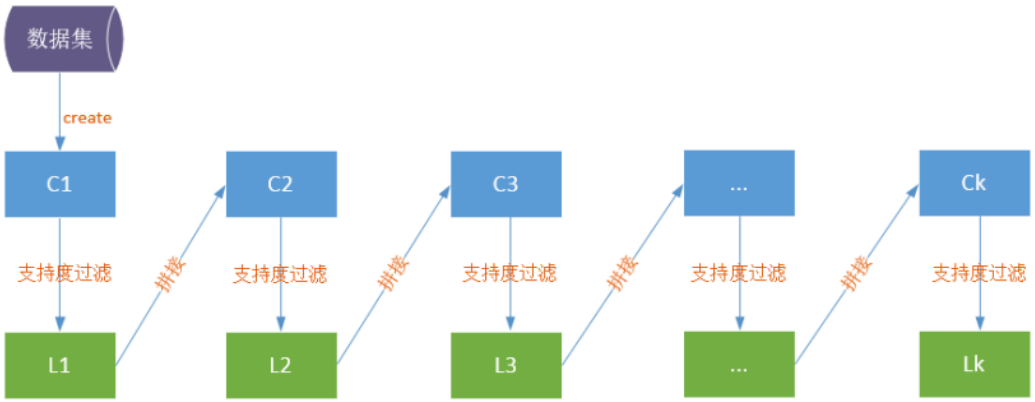


图 4：Apriori 算法流程图

发现频繁项集的过程如图 4 所示：

1) 由数据集生成候选项集 C1（1 表示每个候选项仅有一个数据项）;再由 C1 通过支持度过滤，生成频繁项集 L1（1 表示每个频繁项仅有一个数据项）。

2) 将 L1 的数据项两两拼接成 C2。

3) 从候选项集 C2 开始，通过支持度过滤生成 L2。L2 根据 Apriori 原理拼接成候选项集 C3；C3 通过支持度过滤生成 L3……直到 Lk 中仅有一个或没有数据项为止。

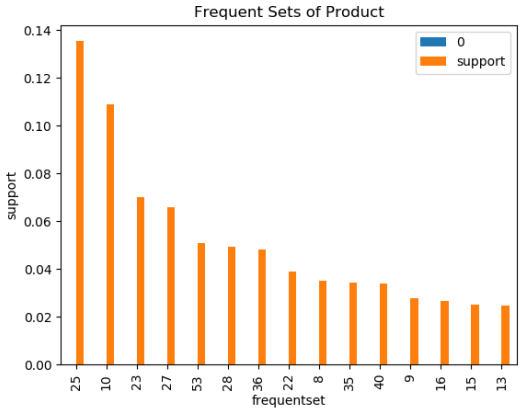
Apriori 算法实现的伪代码如下：

```
Apriori算法
输入:数据集D;最小支持度minsupport
输出:频繁项集L
L[1] = {频繁1项集}//初始化频繁项集L[1]（也称单品项集）
for(k=2;L[k-1]不为空;k++){
    C[k] = candidate_gen(L[k-1]);//根据L[k-1]产生新的候选频繁项集C[k]
    For all transactions t ∈ D; {//对所有的交易记录做循环
        C = subset(C[k],t);//找出当前交易记录t和候选频繁项集C[k]的交集
        For all candidates c ∈C do//对存在的候选频繁项集的交集进行支持度计数
        c.count ++;
    }
    L[k] = {c ∈ C[k] | c.count >= minsup};//保留大于最小支持度的频繁项集到L[k];
}
Answer = L=∪L[k]=L[1] ∪ L[2] ∪ ... ∪ L[k]

candidate_gen(L[k-1])
输入:(k-1)-项集
输出:k-候选集C[k]
for all itemset p ∈ L[k-1]
    for all itemset q ∈ L[k-1]
        if(p.item1 = q.item1,p.item2=q.item2,...,p.item(k-2)=q.item(k-2),p.item(k-1)<q.item(k-1))
            c=p∞q;
        if(has_infrequent_subset(c,L[k-1]) delete c;
        else add c to C[k];
    End for
End for
return C[k];

has_infrequent_subset(c,L[k-1])
输入:一个k-项集c,(k-1)-项集L[k-1]
输出:c是否从候选集中删除
for all (k-1)-subsets of c
    if S ∉ L[k-1]
        return true;
return false;
```

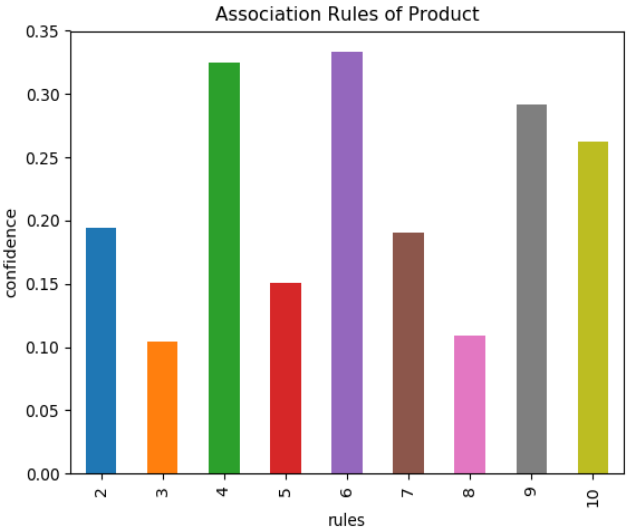
4．实验结果

利用 Apriori 算法进行频繁项集的挖掘，并且根据所挖掘到的频繁项集自动生成关联规则。本实验中设置最小支持度 minSupport = 0.2，最小置信度 minConf = 0.8。尽管最小支持度和最小置信度阈值有助于排出大量无趣规则的探查，但仍然会产生一些用户不感兴趣的规则，故本实验中采用提升度（lift）来作为相关性的度量。如果提升度小于 1，则表示关联规则的前项和后项是负相关的，如果值为 1，则前项和后项是相互独立的，如果值大于 1，意味着前项和后项是正相关的。

在 Instacart Market Basket Analysis 数据集中提取出的 product，aisle 和 department 三类数据集，我们分别找出其所有频繁项集，依据置信度产生关联规则，具体内容如下图所示（为更好的进行可视化操作，图中只表示置信度排名前 15 的频繁项集）。

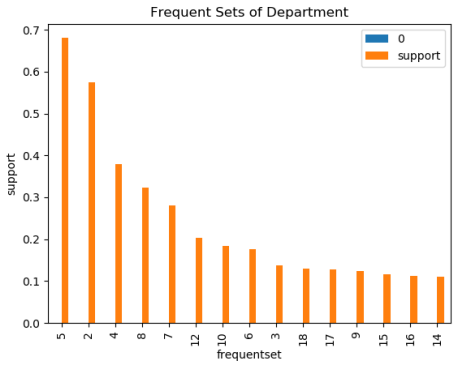


Product_id	Frequent sets	Support
25	Banana	0.1354
10	Bag of Organic Bananas	0.1090
23	Organic Strawberries	0.0702
27	Organic Baby Spinach	0.0657
53	Organic Avocado	0.0508
28	Organic Hass Avocado	0.0493
36	Large Lemon	0.0483
22	Strawberries	0.0390
8	Organic Raspberries	0.0349
35	Limes	0.0345
40	Organic Whole Milk	0.0340
9	Organic Blueberries	0.0276
16	Organic Zucchini	0.0266
15	Organic Yellow Onion	0.0252
13	Organic Garlic	0.0248

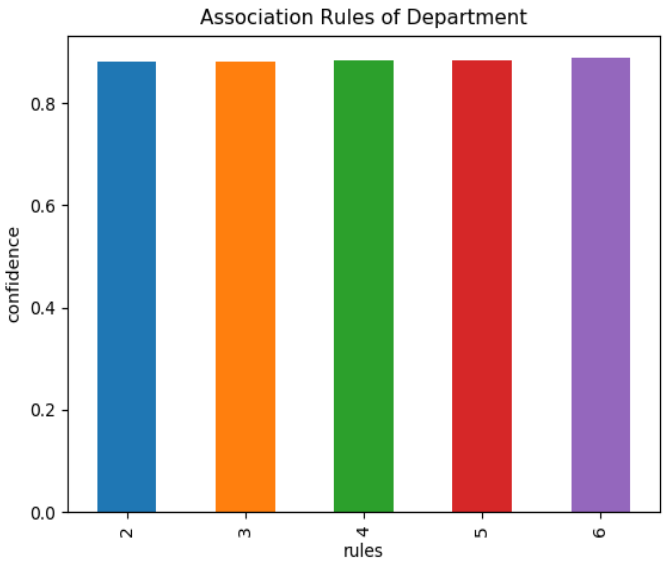


序号	Association rules of product	confidence
2	Organic Baby Spinach -> Banana	0.1944
3	Bag of Organic Bananas -> Organic Raspberries	0.104
4	Organic Raspberries -> Bag of Organic Bananas	0.3248
5	Bag of Organic Bananas -> Organic Hass Avocado	0.1507
6	Organic Hass Avocado -> Bag of Organic Bananas	0.3332
7	Organic Strawberries -> Banana	0.19
8	Banana -> Organic Avocado	0.1095
9	Organic Avocado -> Banana	0.2918
10	Large Lemon -> Banana	0.262

图 5：product 数据集的频繁项集 & 关联规则

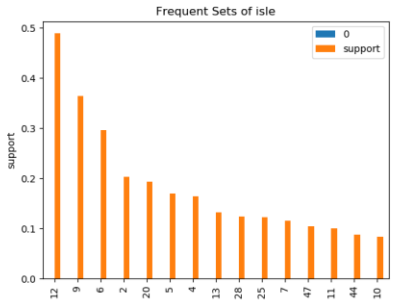


Department_id	Frequent sets	Support
5	produce	0.6804
2	dairy eggs	0.5754
4	beverages	0.3801
8	Snacks	0.3225
7	frozen	0.2804
27	(frozen, produce)	0.2081
12	Bakery	0.2028
10	pantry	0.1848
6	deli	0.1762
0	(deli, produce)	0.1431
25	(pantry, produce)	0.1399
3	canned goods	0.1374
9	meat seafood	0.1245
15	dry goods pasta	0.1171
14	breakfast	0.1113

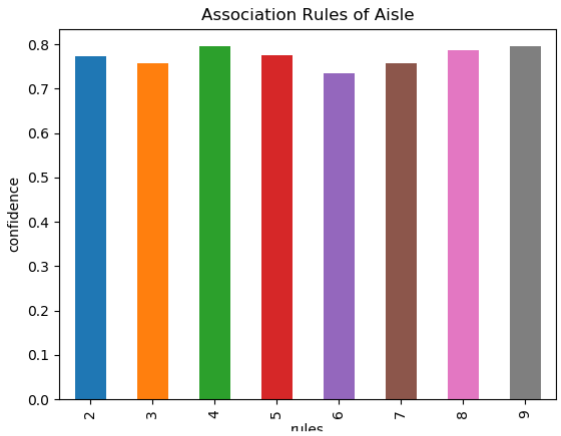


编号	Association rules of department	confidence
2	(snacks, dry goods pasta, dairy eggs) -> produce	0.8813
3	(canned goods, frozen, dairy eggs) -> produce	0.8813
4	(snacks, canned goods, dairy eggs) -> produce	0.8842
5	(dairy eggs, deli, frozen) -> produce	0.8839
6	(bakery, deli, dairy eggs) -> produce	0.8878

图 6：department 数据集的频繁项集 & 关联规则



Aisle_id	Frequent sets	Support
12	fresh fruits	0.4891
9	fresh vegetables	0.3637
6	packaged vegetables fruits	0.2965
2	yogurt	0.2029
20	milk	0.1929
5	water seltzer sparkling water	0.1703
4	packaged cheese	0.1646
13	soy lactosefree	0.1320
55	(fresh fruits, fresh vegetables, …)	0.1310
28	chips pretzels	0.1237
25	bread	0.1225
7	eggs	0.1159
11	refrigerated	0.1004
37	(packaged vegetables fruits, yogurt)	0.0872
10	frozen produce	0.0830



序号	Association rules of aisle	Confidence
2	(fresh herbs) -> fresh vegetables	0.773
3	(fresh vegetables, eggs) -> fresh fruits	0.7569
4	(milk, packaged vegetables fruits) -> fresh fruits	0.7954
5	(milk, fresh vegetables) -> fresh fruits	0.7763
6	(fresh vegetables, water seltzer sparkling water) -> fresh fruits	0.7355
7	(eggs, fresh vegetables) -> fresh fruits	0.7569
8	(packaged cheese, packaged vegetables fruits) -> fresh fruits	0.7876
9	(milk, packaged vegetables fruits) -> fresh fruits	0.7954

图 7：aisle 数据集的频繁项集 & 关联规则

通过对 aisle 数据集挖掘出的频繁项及关联规则的分析，商家可以按照图 8 所示来摆放物品，以实现利润的最大化（从入口到出口位置的商品是按照置信度从低到高进行摆放）。

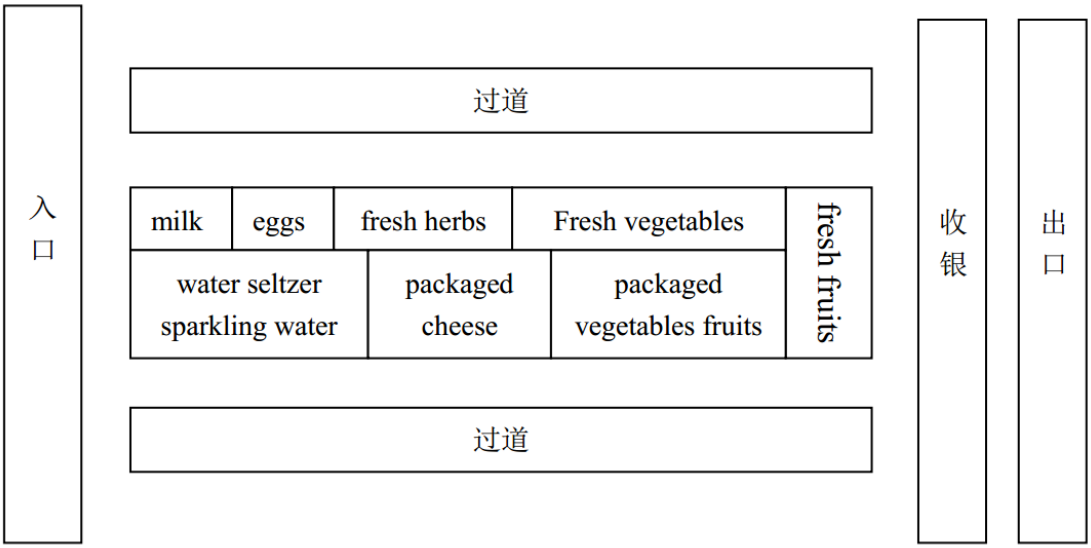


图 8：以过道为例的简单超市平面图

5. 总结

利用关联规则支持超市数据进行分析是一种科学的决策方式，在一定程度上可以避免主观或经验决策中一些先天不足，基于数据挖掘技术的决策系统将会在超市的数据分析中得到更加广泛的认可。数据挖掘技术必将为企业管理者的决策提供切实可靠的依据，使企业在激烈的竞争中掌握主动权，从而获得最大的利益。

参考文献：

[1]陈文伟. 数据仓库与数据挖掘

[2] Mehemd Kantardzic. 数据挖掘

[3]张云涛. 数据挖掘原理与技术

[4]陈玉婷，刘博等. 关联规则挖掘算法介绍

[5]王文琴等。基于关联规则的数据挖掘技术在 CRM 上的应用研究

[6]于芳. 关联分析在超市商品捆绑销售中的应用

[7] Jiawei Han, Micheline, Jian Pei. Data Mining: Concepts and Technology.

[8] Peter Harrington, Machine Learning in Action.

[9] Wes McKinney. Python for Data Analysis.

项目代码：https://github.com/kangyang94/bitdm/tree/master/code