北京理工大學

项目终期报告

课程名称：数据挖掘

组号：第 7 组

成员：赵佰承、王颖、赵佳晨、韩义龙

关联分析在商家营销决策中的应用——终期报告

赵佰承：2120160969　王颖：2120160943
赵佳晨：3120160401　韩义龙：2720170002

摘要

对于给定某个时间段内商家的销售数据，关联分析的目标是从销售数据库中找出各商品之间潜在的关联关系。使用蛮力搜索的算法在寻找物品的不同组合时是一项十分耗时的任务，所以计算代价很高，为此我们采用 Apriori 算法来找出所有符合要求的频繁项集。本文还在关联规则的基础上附加了利润因素，由此辅助商家的营销决策，实现利润最大化。

1. 背景

近年来我国的零售行业发展迅速，越来越多的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。通过查看哪些商品经常一起购买，可以帮助商家了解用户的购买行为。这种从数据的海洋中抽取的知识可以用于商品的定价、市场促销、存货管理等环节。其重要技术就是利用关联分析挖掘潜在用户，提高销售额。从大规模的数据集中寻找物品间的隐含关系被称作关联分析或者关联规则学习。这里的主要问题在于，寻找物品的不同组合是一项十分耗时的任务，所需的计算代价很高，蛮力搜索方法并不能解决这个问题，所以需要用更智能的方法在合理的时间范围内找到频繁项集。

传统的关联规则的缺点就是没有考虑关联规则的商业价值。如果采用传统的关联规则评价标准，同时销售“一瓶昂贵的红酒和一盒鱼子酱”与销售“一盒牛奶和一袋面包”对于关联规则挖掘过程来说意义相差不大。但事实上，零售商会更倾向推销前者，显然该组合能为企业带来比后者更大的净利润。 利用开放出来的超市购物蓝数据进行关联规则分析，因为人们在选择购买商品的时候往往会购买多个商品，如果这些商品摆放在一

起，就会增加用户购买的可能性。

我们在关联分析 Apriori 经典算法的基础上，充分考虑了利润收益，提出了运用频繁项目集中各个项目的加权利润之和来表示交叉销售的利润的思想评估关联规则价值，并且在“数据堂”提供的实际电商销售数据上进行了频繁项集的挖掘，提出各种捆绑销售模型，具有很强的实用价值。

2.问题描述

本文解决的问题归纳如下：在已知的销售数据集，包括如买家 ID；商品名称，价格；销售时间等信息的基础上，进行关联性分析，挖掘到频繁项目集，给出频繁项目集的直方图，从而进行合理的产品摆放和个性化推荐，从而提高销售利润。

频繁项目集通常表示顾客一次购物时同时购买的商品组合，因此评估一项商品的商业价值时，除了考虑该商品本身所产生的独立利润外，同时也应考虑由于交叉销售时带来的额外利润。

3.运行环境

Windows7 64 位

Python 3.6

支持包 numpy、pandas、xlrd、matplotlib、Levenshtein

4.数据集及实现的技术方案

为了实现频繁项集的挖掘，本文利用 Apriori 算法来寻找不同的商品组合。图 1 展示了本实验方案处理的流程。

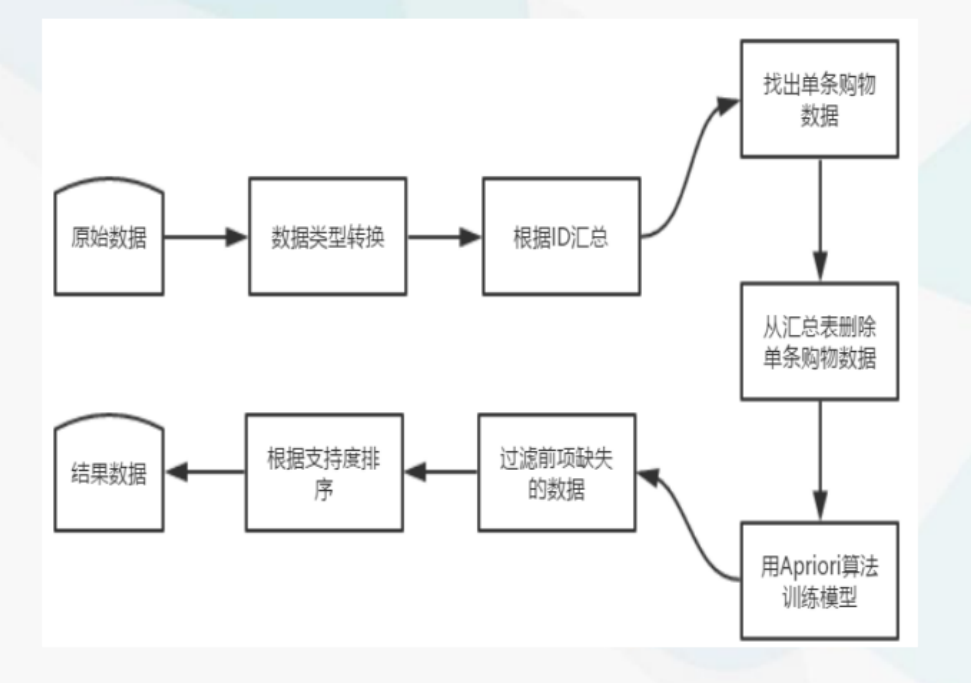

图 1. 技术方案流程图

4.1 数据集介绍及预处理

本实验选择网络收集的实际电商销售数据作为数据挖掘样本。选择来自于数据堂的 65535 条数据，其中包括了订单号、产品名称、付款时间、付款金额、付款类型以及支付类型等属性，属于典型的电商购物型数据。其数据格式如图 2 所示。

订单号	产品名称	付款时间	付款金额	付款类型	支付类型
20140929050517	女士深蓝色系带鞋	2014/11/3 15:51	477	手机货到付款	货到付款
20140929050517	女士平底小方头中长靴	2014/11/3 15:51	1332	手机货到付款	货到付款
20140929050517	女士黑色短款毛衣	2014/11/3 15:51	1332	手机货到付款	货到付款
20140929050517	女士黑色修身剪裁连衣裙	2014/11/3 15:51	1332	手机货到付款	货到付款
20141008060278	M-BINDU 摇滚风格银丝装饰针织衫	2014/11/3 15:51	1332	手机货到付款	货到付款
20141014068622	女士经典T字绣线优雅斜跨包手拎包	2014/11/10 15:05	2820	手机货到付款	货到付款
20141014068939	【秋冬新品】女士休闲平底中筒靴	2014/11/5 12:02	0	赠送	赠送

图 2.数据集样本数据格式

考虑数据集中存在商品中有赠品并且同一件商品由于颜色或尺寸的区分导致商品名称不同的情况。对于这样的数据在频繁项集的挖掘之前需要进行预处理，考虑采用的方法包括：

（1）将商品中属于赠品的条目剔除

（2）对于同一件商品由于颜色的不同导致的商品名字不同的情况，

将商品名均用其中某一个名称表示。

（3）对于同一件商品由于尺寸的不同导致的商品名字不同的情况，将商品名均用其中某一个名称表示。

4.2 频繁项集挖掘

考虑数据集特点和计算硬件条件，决定采用 Apriori 算法实现，而 Apriori 算法的基本思想是首先是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，这些规则必须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项集列表，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。Apriori 算法流程图 3 所示。

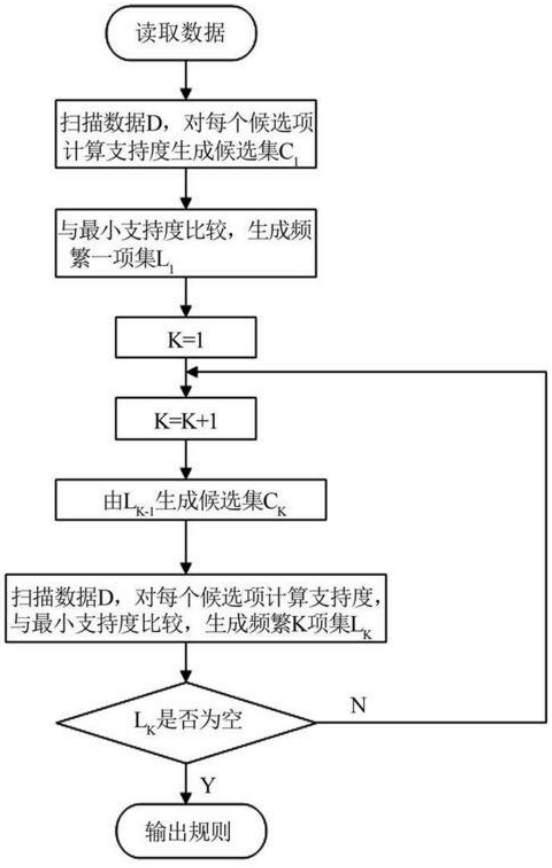

图 3. Apriori 算法流程图

Apriori 算法实现的伪代码如下：

```
输入：
    D：事务数据库；
    min_sup：最小支持度计数阈值。
输出：  L：D 中的频繁项集。
方法：
1)  L1 = find_frequent_1_itemsets(D);
2)  for (k = 2; Lk-1 ≠ ∅; k++) {
3)    Ck = aproiri_gen(Lk-1,min_sup);
4)    for each transaction t∈D{ //扫描 D 用来计数
5)      Ct = subset(Ck,t); //找出事务 t 中包含的所有候选 k 项集，
6)      for each candidate c∈Ct //对事务 t 包含的每个候选 k 项集的计数加一
7)        c.count++;
8)  }
9)    Lk={c∈Ck | c.count ≥ min_sup}
10) }
11) return L = ∪kLk;
procedure apriori_gen(Lk-1: frequent (k-1)-itemset; min_sup: support)
1) for each itemset l1∈Lk-1
2)   for each itemset l2∈Lk-1
3)     if (l1[1]=l2[1])∧...∧(l1[k-2]=l2[k-2])∧(l1[k-1]<l2[k-2]) then {
4)       c = l1 连接 l2; //连接步：产生 candidates
5)       if has_infrequent_subset(c,Lk-1) then
6)         delete c; // 剪枝步：移除非频繁的 cadidate
7)       else add c to Ck;
8) }
9) return Ck;
procedure has_infrequent_subset(c:candidate k-itemset; Lk-1:frequent
(k-1)-itemset)
// 使用先验知识
1) for each (k-1)-subset s of c
2)   if c∉Lk-1 then
3)     return TRUE;
4) return FALSE;
```

其中，Lk-1 表示频繁 k-1 项集。

4.3 考虑交叉销售利润的频繁项目集挖掘

在所得到的关联规则中，留下符合最小支持度和最小置信度的规则，并且根据每件商品的利润，计算出每一条关联规则的所对应的利润。设计优化的目标函数（本实验中所采用的目标函数为 $J = \alpha \cdot \sup port + profit$ ），

在最小支持度与利润之间做出权衡。

5.实验结果

本实验首先对数据集进行预处理，将同一件商品由于尺寸或者颜色不同而导致的商品名称的差异进行一般化。并将商品中的赠品项删除。然后，通过用户 ID 将销售数据合并成购物篮元组数据，为下一步的关联规则挖掘做准备。

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 = 0.015，最小置信度 minConf = 0.1。所挖掘得到的频繁项集的直方图如图 4 所示：

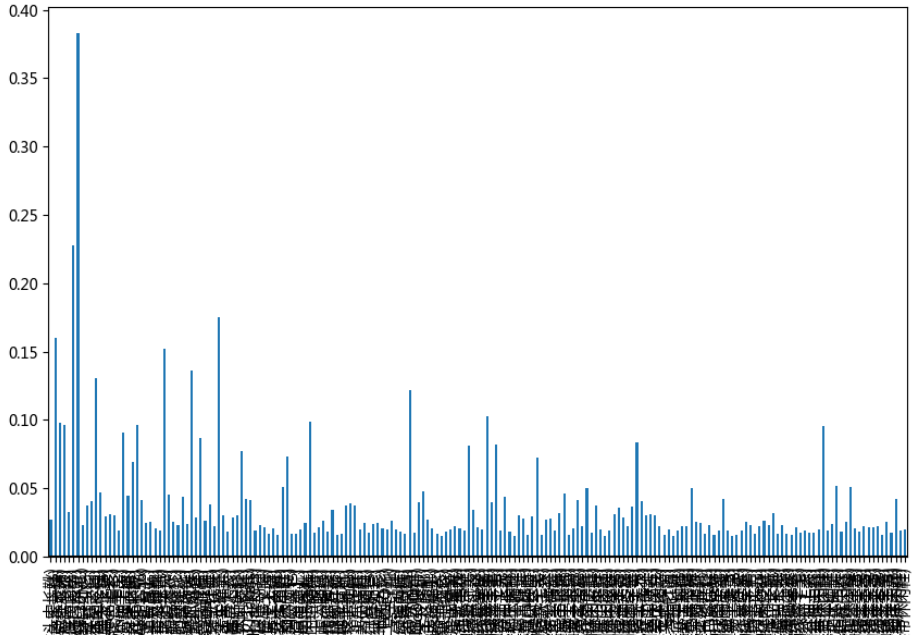

图 4. 关联规则直方图

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。如果提升度的值小于 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。计算出每一条规则对应的提升度，并且根据利润表，查询得到每一条关联规则各商品加权的利润，将其一起写

入关联规则中。最终得到的部分关联规则如图 5 所示:

前项	后项	support	confidenc	lift	利润
frozenset({'男士星际行者系列黑色幼线签字笔	frozenset({'女士进口长巾'})	0.102412074	0.267229	1.175524	129.7
frozenset({'女士进口长巾'})	frozenset({'男士星际行者系列黑色幼线签字笔'})	0.102412074	0.450504	1.175524	129.7
frozenset({'2013年新品女士logo印花拼接系带	frozenset({'女士裤子'})	0.095404932	0.781457	4.471235	87
frozenset({'女士裤子'})	frozenset({'2013年新品女士logo印花拼接系带休闲	0.095404932	0.545875	4.471235	87
frozenset({'【2014秋冬新款】男士夹克'})	frozenset({'男士星际行者系列黑色幼线签字笔'})	0.083546692	0.549159	1.432949	144.7
frozenset({'男士星际行者系列黑色幼线签字笔	frozenset({'【2014秋冬新款】男士夹克'})	0.083546692	0.218003	1.432949	144.7
frozenset({'男士星际行者系列黑色幼线签字笔	frozenset({'男士GANCIO饰扣休闲皮鞋'})	0.081794906	0.213432	1.639625	195.7
frozenset({'男士GANCIO饰扣休闲皮鞋'})	frozenset({'男士星际行者系列黑色幼线签字笔'})	0.081794906	0.628364	1.639625	195.7
frozenset({'男士星际行者系列黑色幼线签字笔	frozenset({'男士半开襟抓绒衣'})	0.081525401	0.212729	1.564577	129.7
frozenset({'男士半开襟抓绒衣'})	frozenset({'男士星际行者系列黑色幼线签字笔'})	0.081525401	0.599604	1.564577	129.7
frozenset({'男士星际行者系列黑色幼线签字笔	frozenset({'女士深蓝色系带鞋'})	0.07209271	0.188115	1.174099	148
frozenset({'女士深蓝色系带鞋'})	frozenset({'男士星际行者系列黑色幼线签字笔'})	0.07209271	0.449958	1.174099	148
frozenset({'[限量款]Virgin系列杏色铆钉装饰	frozenset({'女士裤子'})	0.051745048	0.707182	4.046261	113.7
frozenset({'女士裤子'})	frozenset({'[限量款]Virgin系列杏色铆钉装饰四粒	0.051745048	0.296068	4.046261	113.7
frozenset({'[限量款]Virgin系列杏色铆钉装饰	frozenset({'2013年新品女士logo印花拼接系带休闲	0.051071284	0.697974	5.717071	152.7
frozenset({'2013年新品女士logo印花拼接系带	frozenset({'[限量款]Virgin系列杏色铆钉装饰四粒	0.051071284	0.418322	5.717071	152.7
frozenset({'女士深蓝色系带鞋'})	frozenset({'女士进口长巾'})	0.050128015	0.312868	1.376285	146.3
frozenset({'女士进口长巾'})	frozenset({'女士深蓝色系带鞋'})	0.050128015	0.22051	1.376285	146.3
frozenset({'男士星际行者系列黑色幼线签字笔	frozenset({'女士绣花logo黑色抓绒休闲运动裤'})	0.049723757	0.129747	1.313576	139.7

图 5 挖掘得到的部分关联规则

对于所得到的关联规则，将关联规则前项和后项是负相关或者独立的规则剔除（即提升度 lift <= 1 的规则），选取优化函数 $J=\alpha\cdot\sup port+\beta\cdot profit$，此处可以将参数β 设为 1，优化函数 J 不变。此时优化目标函数为 $J=\alpha\cdot\sup port+profit$，在本实验中，选取参数α 为 100，计算每一条规则对应的 J，根据 J 的大下对关联规则进行排列，如图 6 所示。最后根据需求决定采用哪些关联规则应用于商家的营销决策中。

前项	后项	support	confidenc	lift	利润	目标J
frozenset({'男士GANCIO饰	frozenset({'女士深蓝色系带	0.016036	0.123188	1.70875	278	279.6036
frozenset({'女士深蓝色系	frozenset({'男士GANCIO饰扣	0.016036	0.100084	1.223598	278	279.6036
frozenset({'【2014秋冬新	frozenset({'男士GANCIO饰扣	0.018192	0.119575	1.461886	274.7	276.5192
frozenset({'男士GANCIO饰	frozenset({'男士星际行者系	0.018192	0.139752	1.672736	274.7	276.5192
frozenset({'女士摩登千鸟	frozenset({'男士星际行者系	0.021965	0.519108	1.354537	263.7	265.8965
frozenset({'男士GANCIO饰	frozenset({'男士半开襟抓绒	0.022504	0.172878	2.12054	259.7	261.9504
frozenset({'男士半开襟抓	frozenset({'男士GANCIO饰扣	0.022504	0.16551	2.023481	259.7	261.9504
frozenset({'男士GANCIO饰	frozenset({'女士进口长巾',	0.021291	0.163561	1.597088	259.7	261.8291
frozenset({'中性时尚光学	frozenset({'男士星际行者系	0.031802	0.46184	1.205102	246.9	250.0802
frozenset({'中性时尚光学	frozenset({'女士进口长巾'}	0.020752	0.30137	1.325706	245.2	247.2752
frozenset({'男款湖蓝色抓	frozenset({'男士星际行者系	0.019943	0.646288	1.686394	240.7	242.6943

图 6 优化后的部分关联规则

对于上列的关联规则，商家可以选择，将 “男士 GANCIO 饰扣休闲皮鞋”+“女士深蓝色系带鞋”+ “男士星际行者系列黑色幼线签字笔”进行捆绑销售，类似地，可以做出一系列的营销决策，并且可以后续的销售情况将参数α 调节至合适的值，以实现利润的最大化。

参考文献

[1] Jiawei Han, Micheline, Jian Pei. Data Mining: Concepts and Technology[D]. Morgan Kaufmann, 2011.

[2] Peter Harrington, Machine Learning in Action[D]. Manning Publictions Co., 2012.

[3] 于芳. 关联分析在超市商品捆绑销售中的应用[J]. 商场现代化, 2010, 1:33-34.

[4] Wes McKinney. Python for Data Analysis[D].O'Reilly Media, 2013.