

数据挖掘导论大作业开题报告

选题

本次课程综合实验的选题为“淘宝母婴购物数据可视化分析”，本题为天池大数据竞赛学习赛的一道选题（[淘宝母婴购物数据可视化分析学习赛天池大赛-阿里云天池 \(aliyun.com\)](https://tianchi.aliyun.com)）。



背景介绍

该背景介绍部分摘自天池竞赛赛事网页

母婴用品是淘宝的热门购物类目，随着国家鼓励二胎、三胎政策的推进，会进一步促进了母婴类目商品的销量。与此年轻一代父母的育儿观念也发生了较大的变化，因此中国母婴电商市场发展形态也越来越多样化。随之引起各大母婴品牌更加激烈的争夺，越来越多的母婴品牌管窥到行业潜在的商机，纷纷加入母婴电商，行业竞争越来越激烈。

本次课程综合实验将基于“淘宝母婴购物数据”进行分析。

数据集介绍

本次学习赛提供了两个数据集，分别为“tianchi_mum_baby_trade_history.csv”和“tianchi_mum_baby.csv”，两个数据集分别包含不同信息：

数据集	内容
'tianchi_mum_baby_trade_history.csv'	购物行为表
'tianchi_mum_baby.csv'	母婴信息表

每个数据集中的数据及特征值分别如下：

母婴信息表 ('tianchi_mum_baby.csv') :

字段	字段说明
user_id	用户id
birthday	出生年月日
gender	性别(0-男孩 1-女孩 2-不明)

< < 1-10 > > 953 rows × 3 columns pd.DataFrame			
	user_id	birthday	gender
1	415971	20121111	0
2	1372572	20120130	1
3	10339332	20110910	0
4	10642245	20130213	0
5	10923201	20110830	1
6	11768880	20120107	1
7	12519465	20130705	1
8	12950574	20090708	0
9	13735440	20120323	0

购物行为表 ('tianchi_mum_baby.csv') :

字段	字段说明
user_id	用户id
auction_id	交易id
category_1	商品一级目录
category_2	商品二级目录
buy_mount	购买数量
day	交易时间 (年月日)

< < 1-10 > > 29971 rows × 6 columns pd.DataFrame						
	user_id	auction_id	category_2	category_1	buy_mount	day
0	786295544	41098319944	50014866	50022520	2	20140919
1	532110457	17916191097	50011993	28	1	20131011
2	249013725	21896936223	50012461	50014815	1	20131011
3	917056007	12515996043	50018831	50014815	2	20141023
4	444069173	20487688075	50013636	50008168	1	20141103
5	152298847	41840167463	121394024	50008168	1	20141103
6	513441334	19909384116	50010557	50008168	1	20121212
7	297411659	13540124907	50010542	50008168	1	20121212
8	82830661	19948600790	50013874	28	1	20121101

环境

本次实验使用的python环境为之前几次实验中用anaconda新搭建的环境'newConda'，python版本为3.8，已经安装了之前实验中使用到的各种基本packages，在本次实验中预计将安装聚类、分类、频繁模式挖掘相关的官方库以调用进行处理。

实验中.ipynb文件使用DataSpell 2023.1.2进行编写和运行。

预计实验过程

数据预处理

本次实验的数据集对预处理过程要求可能较高，因为存在部分缺失值、异常值等。将对数据集进行数据清洗，在进行聚类前对数据集进行进行数据变化，将原始数据变换为适合挖掘的形式。

可视化

可视化将会是本次实验的重点之一，因此本次学习赛的题目就为“淘宝母婴购物数据可视化分析”。本次实验中，预计将对数据集依照各种特征值进行可视化处理，方便直观地挖掘什么时间（哪一年或一年中的哪个月、哪一天）交易量最大、什么年龄的婴儿家庭中购买相关母婴产品较多、什么类商品对交易量的贡献最大等。

聚类和分类

参考之前实验中的结果，本次将首先对数据集进行聚类操作（预计使用K-Means），之后再进行分类（具体算法未确定，将在实验过程中根据表现进行选择），并尝试预测并进行购物车推荐。

频繁模式挖掘

预计将使用Apriori算法进行频繁模式挖掘，尝试发现数据集中隐藏的信息，例如什么样的家庭喜欢哪一类的产品。