天津大学

数据挖掘实验报告

天池教学赛——产品关联分析

学　院　　智能与计算学部

专　业　　人工智能

年　级　　2019级

学　号　　3019244132

姓　名　　游奕桁

日　期　　2021年10月25 日

一、任务描述

- 赛题以购物篮分析为背景，要求选手对品牌的历史订单数据，挖掘频繁项集与关联规则。通过这道赛题，鼓励学习者利用订单数据，为企业提供销售策略，产品关联组合，为企业提升销量的同时，也为消费者提供更适合的商品推荐。
- 需要使用关联分析（比如Apriori算法） 挖掘订单中的频繁项集及关联规则
  - 频繁项集、关联规则的计算会用到支持度、置信度、提升度等指标。
  - 频繁项集：即大于最小支持度的商品或商品组合。
  - 关联规则：在频繁项集中，满足最小置信度，或最小提升度的推荐规则

二、数据

- 数据源：order.csv，product.csv，customer.csv，date.csv ，分别为订单表，产品表，客户表，日期表。
- 使用panda库导入数据

```
order_df = pd.read_csv('./order.csv', encoding='gbk')
customer_df = pd.read_csv('./customer.csv', encoding='gbk')
date_df = pd.read_csv('./date.csv', encoding='gbk')
product_df = pd.read_csv('./product.csv', encoding='gbk')
```

- 查看订单前五条数据，发现挖掘关联规则只需要使用到'订单日期'、'客户ID'、'产品名称'三个属性即可。

```
order_df.head(5)
```

	订单日期	年份	订单数量	产品ID	客户ID	交易类型	销售区域ID	销售大区	国家	区域	产品类别	产品型号名称	产品名称	产品成本	利润	单价	销售金额
0	2016/1/1	2016	1	528	14432BA	1	4	西南区	中国	大中华区	配件	Rawlings Heart of THE Hide-11.5	棒球手套	500.0	1199.0	1699.0	1699.0
1	2016/1/2	2016	1	528	18741BA	1	4	西南区	中国	大中华区	配件	Rawlings Heart of THE Hide-11.5	棒球手套	500.0	1199.0	1699.0	1699.0
2	2016/1/2	2016	1	528	27988BA	1	4	西南区	中国	大中华区	配件	Rawlings Heart of THE Hide-11.5	棒球手套	500.0	1199.0	1699.0	1699.0
3	2016/1/5	2016	1	528	25710BA	1	4	西南区	中国	大中华区	配件	Rawlings Heart of THE Hide-11.5	棒球手套	500.0	1199.0	1699.0	1699.0
4	2016/1/6	2016	1	528	14999BA	1	4	西南区	中国	大中华区	配件	Rawlings Heart of THE Hide-11.5	棒球手套	500.0	1199.0	1699.0	1699.0

- 数据预处理：
  - 检查异常数据，空值及重复等

    订单数据集中不存在空值

```
pd.isnull(order_df).sum().values
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

    发现重复值，进行去除

```
order_df.duplicated().sum()
```

```
7
```

```
order_df = order_df.drop_duplicates()
order_df.duplicated().sum()
```

```
0
```

  - 将'订单日期'字段转化为日期格式
  - 用groupby来分出同一客户在同一日期购买的产品

```
order_df['订单日期']=pd.to_datetime(order_df['订单日期'])
order_df = order_df.groupby(['订单日期', '客户ID'])['产品名称'].unique()
```

- 分组后的结果：

```
订单日期        客户ID
2016-07-31  20199BA          [击打手套, 棒球手套]
            20550BA          [球棒与球棒袋, 头盔]
            20770BA       [帽子, 棒球服, 三角网架]
            20780BA    [打击T座, 帽子, 三角网架]
            21093BA       [帽子, 头盔, 棒球手套]
            21606BA            [头盔, 棒球手套]
            21914BA             [棒球服, 皮带]
            22222BA            [皮带, 捕手护具]
            22622BA          [装备包, 棒球手套]
            23184BA    [装备包, 垒垫, 球棒与球棒袋]
            23368BA                [棒球手套]
            23544BA            [袜子, 棒球手套]
            23545BA       [棒球服, 头盔, 棒球手套]
            23738BA          [头盔, 球棒与球棒袋]
            25402BA                [棒球手套]
            26725BA                [棒球手套]
            28585BA          [击打手套, 棒球手套]
            28853BA           [装备包, 棒球手套]
            29707BA            [头盔, 棒球手套]
            30810BA                   [垒垫]
Name: 产品名称, dtype: object
```

三、算法

- 使用Apriori算法对数据集中的关联规则和频繁项集进行挖掘
- Apriori算法伪代码

```
输入：数据集合D，支持度阈值α
输出：最大的频繁k项集
1）扫描整个数据集，得到所有出现过的数据，作为候选频繁1项集。k=1，频繁0项集为空集。
2）挖掘频繁k项集
   a）扫描数据计算候选频繁k项集的支持度
   b）去除候选频繁k项集中支持度低于α的数据集,得到频繁k项集。如果得到的频繁k项集为空，则直接返回频繁k-1项集的集合作为算法结果，算法结束。如果得到的频繁k项集只有一项，则直接返回频繁k项集的集合作为算法结果，算法结束。
   c）基于频繁k项集，连接生成候选频繁k+1项集。
3）  令k=k+1，转入步骤2。
```

- 实验调用mlxtend库中的apriori算法进行实现
  - 将订单数据转化为列表

```
transactions = []
for value in order_df:
    transactions.append(list(value))
```

  - 将订单数据列表重新编码成mlxtend库中Apriori函数支持的格式

```
te = TransactionEncoder()
encode = te.fit(transactions).transform(transactions)
data = pd.DataFrame(encode, columns=te.columns_)
```

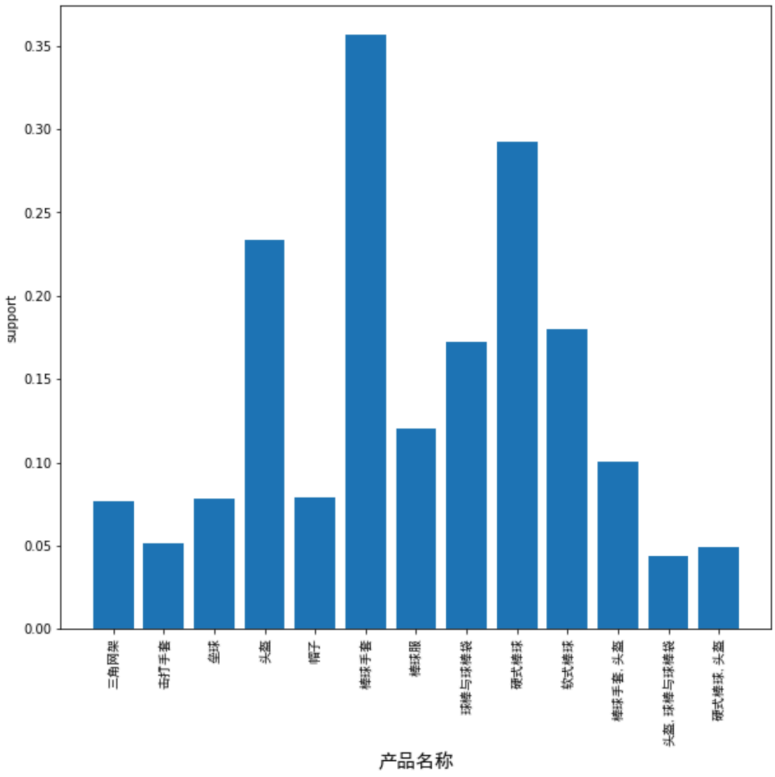
- 调用算法

```
support = apriori(data, min_support=0.04, use_colnames=True)
support.sort_values(by = "support")
rule = association_rules(support, metric='confidence', min_threshold=0.15)
```

四、实验结果

- 得到的频繁项集

	support	itemsets
0	0.076798	(三角网架)
1	0.051778	(击打手套)
2	0.078463	(垒球)
3	0.233145	(头盔)
4	0.079296	(帽子)
5	0.356434	(棒球手套)
6	0.120646	(棒球服)
7	0.172605	(球棒与球棒袋)
8	0.292128	(硬式棒球)
9	0.179955	(软式棒球)
10	0.100261	(棒球手套, 头盔)
11	0.043559	(头盔, 球棒与球棒袋)
12	0.048990	(硬式棒球, 头盔)



- 关联规则

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(棒球手套)	(头盔)	0.356434	0.233145	0.100261	0.281288	1.206494	0.017160	1.066985
1	(头盔)	(棒球手套)	0.233145	0.356434	0.100261	0.430036	1.206494	0.017160	1.129134
2	(头盔)	(球棒与球棒袋)	0.233145	0.172605	0.043559	0.186830	1.082416	0.003317	1.017494
3	(球棒与球棒袋)	(头盔)	0.172605	0.233145	0.043559	0.252360	1.082416	0.003317	1.025701
4	(硬式棒球)	(头盔)	0.292128	0.233145	0.048990	0.167700	0.719293	-0.019118	0.921368
5	(头盔)	(硬式棒球)	0.233145	0.292128	0.048990	0.210126	0.719293	-0.019118	0.896183

五、实验体会

- 通过本次实验，学会了天池实验室的基本使用，了解了用天池实验室参加比赛的基本流程。
- 在实践过程中，加深了对Apriori算法的理解，对支持度、置信度、提升度等指标的计算和含义又了更好的理解。