

DataMining_homework2 实验报告

姓名：张映雪

学号：2120171101

小组：第 6 组

目录

1. 实验概述	3
2. 实验环境	3
3. 实验过程与结果分析	4
3.1 对数据集进行处理, 转换成适合关联规则挖掘的形式	4
3.2 Apriori 算法	4
4. 实验总结	7

1. 实验概述

1.1 实验所用数据集

- Building_Permits.csv

1.2 实验要求

- 对数据集进行处理，转换成适合关联规则挖掘的形式；
- 找出频繁项集；
- 导出关联规则，计算其支持度和置信度
- 对规则进行评价，可使用 Lift，也可以使用教材中所提及的其它指标

1.3 实验涉及算法

- Apriori 算法（由于作业提交时间较紧，没有继续尝试改进后的 Apriori 算法，课后会继续学习尝试这部分内容）

2. 实验环境

- 使用 python2.7
- 数据预处理所用的包： numpy、pandas

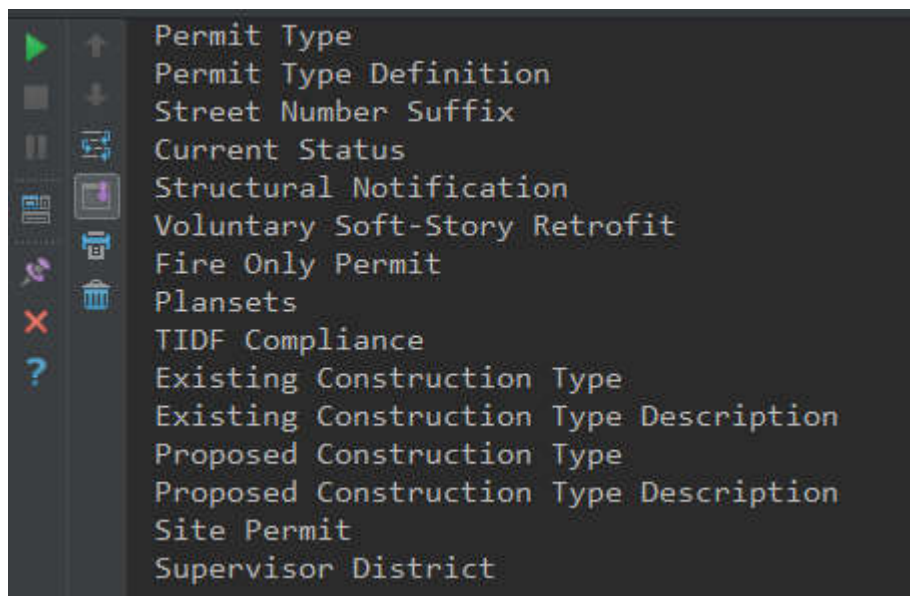
3. 实验过程与结果分析

3.1 对数据集进行处理，转换成适合关联规则挖掘的形式

首先对数据进行预处理。关联规则的挖掘时间主要是在生成频繁项集上，因为找出的频繁项集往往不会很多，利用频繁项集生成规则也就不会花太多的时间，而生成频繁项集需要测试很多的备选项集，在不加优化的情况下，所需的时间是 $O(2^N)$ 。而若将本数据集中所有属性进行处理，生成频繁项集所需要的时间太多，同时数据集中有些属性的属性值太过分散，用于关联规则挖掘的意义也不大，且因此我首先对数据集中的属性进行了筛选，找出属性值的取值比较密集的一部分属性，只对它们进行关联规则挖掘，代码实现如下：

```
def preprocess(file):  
    column_suitable = []  
    for column in file.columns:  
        if 0 < file[column].value_counts().__len__() < 20:  
            print column  
            column_suitable.append(column)
```

选出的属性有：



3.2 Apriori 算法

使用 Apriori 算法，首先计算出单个元素的支持度，然后选出单个元素置信度大于我们要求的数值，然后增加单个元素组合的个数，只要组合项的支持度大

于我们要求的数值就把它加到频繁项集中，依次递归。然后根据计算的支持度选出来的频繁项集来生成关联规则。

实验设置最低支持度为 0.3，最低置信度为 0.7。以下是一些重点模块的展示：

找出所有满足最小支持度的项集

```
# 求满足最小支持度的子集
def min_support(item_set, transaction_list, min_sup, freqs):
    _items = set()
    local_set = defaultdict(int)

    for item in item_set:
        for transaction in transaction_list:
            if item.issubset(transaction):
                freqs[item] += 1
                local_set[item] += 1
    for item, count in local_set.items():
        support = float(count)/len(transaction_list)
        if support >= min_sup:
            _items.add(item)
    return _items
```

Apriori 算法的核心代码;求解频繁项集和关联规则：

```
# 返回值：频繁项集和关联规则
def algo_apriori(data_iter, min_sup, min_confidence):
    item_set, transaction_list = getItemSetTransactionList(data_iter)
    # 所有项的频数 频繁项和非频繁项，1项和K项
    freqs = defaultdict(int)
    # 存储各元频繁项集合 key=K, value=K项频繁项集合
    large_set = dict()
    one_con_set = min_support(item_set, transaction_list, min_sup, freqs)
    current_l_set = one_con_set
    k = 2
    # 递归逐层求解
    while(current_l_set != set([])):
        large_set[k-1] = current_l_set
        current_l_set = joinSet(current_l_set, k)
        current_c_set = min_support(current_l_set, transaction_list, min_sup, freqs)
        current_l_set = current_c_set
        k = k + 1

    def get_support(item):
        return float(freqs[item])/len(transaction_list)

    get_items = []
    for key, value in large_set.items():
        Items.extend([(tuple(item), get_support(item))
                       for item in value])
```

```

get_rules = []
for key, value in large_set.items()[1:]:
    for item in value:
        _subsets = map(frozenset, [x for x in subsets(item)])
        for element in _subsets:
            remain = item.difference(element)
            if len(remain) > 0:
                confidence = get_support(item)/get_support(element)
                if confidence >= min_confidence:
                    get_rules.append(((tuple(element), tuple(remain)), get_support(
                        item), confidence, confidence/get_support(remain)))
return get_items, get_rules

```

通过实现 Apriori 算法，最终得到了频繁项集、关联规则。分别存在了文件 frequency_results.txt 和 rule_confidence_results.txt 中，部分片段截图如下

频繁项集：三张截图分别取自结果文件的开头、中间和结尾

frequent_item_set	support
('PT_8',)	0.899
('PTD_otc alterations permit',)	0.899
('PTD_otc alterations permit', 'PT_8')	0.899
('PCTD_wood frame (5)',)	0.575
('PCT_5.0',)	0.575
...	...
('ECTD_wood frame (5)', 'PCTD_wood frame (5)', 'CS_complete', 'PCT_5.0')	0.342
('ECT_5.0', 'PCTD_wood frame (5)', 'CS_complete', 'ECTD_wood frame (5)', 'PCT_5.0')	0.342
('PTD_otc alterations permit', 'CS_complete', 'PCTD_wood frame (5)')	0.329
('CS_complete', 'PCT_5.0', 'PT_8')	0.329
...	...
('P_0.0',)	0.318
('PT_8', 'P_0.0')	0.315
('PTD_otc alterations permit', 'P_0.0')	0.315
('PTD_otc alterations permit', 'PT_8', 'P_0.0')	0.315

关联规则：三张截图分别取自结果文件的开头、中间和结尾

rule	support	confidence	lift
('PCT_5.0',) -> ('PCTD_wood frame (5)',)	0.575	1.000	1.739
('PCTD_wood frame (5)',) -> ('PCT_5.0',)	0.575	1.000	1.739
('PTD_otc alterations permit',) -> ('PT_8',)	0.899	1.000	1.112
('PT_8',) -> ('PTD_otc alterations permit',)	0.899	1.000	1.112
...
('CS_complete', 'P_0.0') -> ('ECT_5.0', 'ECTD_wood frame (5)', 'PCTD_wood frame (5)')	0.194	0.905	1.607
('CS_complete', 'P_0.0') -> ('PCT_5.0', 'ECTD_wood frame (5)', 'PCTD_wood frame (5)')	0.194	0.905	1.607
('CS_complete', 'P_0.0') -> ('PCT_5.0', 'ECT_5.0', 'ECTD_wood frame (5)')	0.194	0.905	1.607
('CS_complete', 'P_0.0') -> ('PCT_5.0', 'ECT_5.0', 'PCTD_wood frame (5)')	0.194	0.905	1.607
('CS_complete', 'P_0.0') -> ('PCT_5.0', 'ECT_5.0', 'ECTD_wood frame (5)', 'PCTD_wood frame (5)')	0.194	0.905	1.607
('PT_8', 'P_0.0') -> ('ECTD_wood frame (5)',)	0.285	0.904	1.586
('PT_8', 'P_0.0') -> ('ECT_5.0',)	0.285	0.904	1.586
('PTD_otc alterations permit', 'P_0.0') -> ('ECT_5.0',)	0.285	0.904	1.586
...
('CS_complete',) -> ('PTD_otc alterations permit', 'PT_8', 'PCT_5.0', 'ECT_5.0', 'PCTD_wood frame (5)')	0.326	0.667	1.313
('CS_complete',) -> ('PTD_otc alterations permit', 'PCT_5.0', 'ECT_5.0', 'ECTD_wood frame (5)', 'PCTD_wood frame (5)')	0.326	0.667	1.313
('CS_complete',) -> ('PTD_otc alterations permit', 'PCT_5.0', 'ECT_5.0', 'ECTD_wood frame (5)', 'PT_8')	0.326	0.667	1.313
('CS_complete',) -> ('ECT_5.0', 'PCTD_wood frame (5)', 'ECTD_wood frame (5)', 'PTD_otc alterations permit', 'PCT_5.0', 'PT_8')	0.326	0.667	1.313

本次实验中，采用 Lift 对关联规则打分，且按分由高到低输出在 result_lift 中，以下为结果文件中前几行和最后几行截图：

rule				support	confidence	lift
('PT_8', 'ECTD_wood frame (5)', 'PCTD_wood frame (5)') -> ('PTD_otc alterations permit', 'PCT_5.0', 'ECT_5.0')				0.508	1.000	1.969
('ECT_5.0', 'PT_8', 'PCTD_wood frame (5)') -> ('PTD_otc alterations permit', 'PCT_5.0', 'ECTD_wood frame (5)')				0.508	1.000	1.969
('PCT_5.0', 'PT_8', 'ECTD_wood frame (5)') -> ('PTD_otc alterations permit', 'ECT_5.0', 'PCTD_wood frame (5)')				0.508	1.000	1.969
('PCT_5.0', 'PT_8', 'ECT_5.0') -> ('PTD_otc alterations permit', 'ECTD_wood frame (5)', 'PCTD_wood frame (5)')				0.508	1.000	1.969
('P_2.0',) -> ('PT_8',)				0.397	0.804	0.894
('P_2.0',) -> ('PTD_otc alterations permit',)				0.397	0.804	0.894
('P_2.0',) -> ('PTD_otc alterations permit', 'PT_8')				0.397	0.804	0.894

4 实验总结

经过这次关联规则挖掘的实验，我本人完成了 Apriori 算法的实现，真正掌握了 Apriori 算法，学会了挖掘关联规则的基本流程和代码实现。但由于时间比较紧张，没有尝试一些改进后的算法，在课下会继续学习和探索。