北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

数据挖掘
互评作业二

题　　目：频繁模式与关联规则挖掘

学　　院：计算机学院

专业名称：数字表演

学　　号：3120211084

姓　　名：刘子贤

任课教师：汤世平

1. 数据集说明

1.1 数据集选择

数据集名称：oakland-crime-statistics-2011-to-2016

1.2 详细说明以及处理

在这个数据集中，一共包含 6 个数据子集，分别为 2011-2016 年度的奥克兰犯罪情况。其详细属性如下：

```
2011数据集有以下属性 Index(['Agency', 'Create Time', 'Location', 'Area Id', 'Beat', 'Priority',
       'Incident Type Id', 'Incident Type Description', 'Event Number',
       'Closed Time'],
      dtype='object')
2012数据集有以下属性 Index(['Agency', 'Create Time', 'Area Id', 'Beat', 'Priority',
       'Incident Type Id', 'Incident Type Description', 'Event Number',
       'Closed Time', 'Location 1', 'Zip Codes'],
      dtype='object')
2013数据集有以下属性 Index(['Agency', 'Create Time', 'Location ', 'Area Id', 'Beat', 'Priority',
       'Incident Type Id', 'Incident Type Description', 'Event Number',
       'Closed Time'],
      dtype='object')
2014数据集有以下属性 Index(['Agency', 'Create Time', 'Area Id', 'Beat', 'Priority',
       'Incident Type Id', 'Incident Type Description', 'Event Number',
       'Closed Time', 'Location 1', 'Zip Codes'],
      dtype='object')
2015数据集有以下属性 Index(['Agency', 'Create Time', 'Location', 'Area Id', 'Beat', 'Priority',
       'Incident Type Id', 'Incident Type Description', 'Event Number',
       'Closed Time'],
      dtype='object')
2016数据集有以下属性 Index(['Agency', 'Create Time', 'Location', 'Area Id', 'Beat', 'Priority',
       'Incident Type Id', 'Incident Type Description', 'Event Number',
       'Closed Time'],
      dtype='object')
```

通过观察数据集可得知，这六年的数据属性基本一样，值得进行分析与预处理的有如下几个属性：Agency, Location, Area id, Beat, Incident Type id，Incident Type Descripe, Event Number，其中 2012 年和 2014 年的属性为 Location 1，经过特殊处理变为 Location。由于 Incident Type id 与 Incident Type Descripe 一一对应，我们只对

Incident Type id 进行分析。Event Number 对应每一行数据，不具备重复性，不对其进行分析。

发现有部分数据存在缺失值的情况。使用上次预处理的方法，舍去有缺失值的行后，由原来的 1046388 条数据变为剩下 859898 条数据，在此基础上进行实验。

2. 找出频繁模式

2.1 使用算法及简单介绍

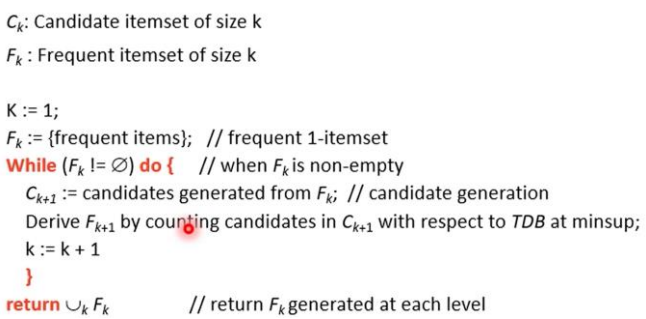
算法名称：Apriori 算法

算法介绍：Apriori 算法是第一个关联规则挖掘算法，也是最经典的算法。它利用逐层搜索的迭代方法找出数据库中项集的关系，以形成规则，其过程由连接（类矩阵运算）与剪枝（去掉那些没必要的中间结果）组成。（该部分引用自百度百科）。

在本实验中，使用 Apriori 算法来构建频繁项集。在本实验中，我们约定支持度的阈值为 10%，置信度的阈值为 50%。

```
min_sup = 0.1
min_conf = 0.5
```

算法的主要流程如课件中所示（如下图）：



相应代码为：

Apriori 主函数：

```
def apriori(self, dataset):             #算法主体
    C1 = self.C1_generation(dataset)        #生成单元数候选项集
    dataset = [set(data) for data in dataset]
    F1, support_data = self.Ck_low_support_filtering(dataset, C1)
    F = [F1]
    k = 2
    while len(F[k-2]) > 0:
        Ck = self.apriori_gen(F[k-2], k)        #当候选项元素太于2时，合并时检测是否子项集满足频繁
        Fk, support_k = self.Ck_low_support_filtering(dataset, Ck)      #过滤支持度低于阈值的项集
        support_data.update(support_k)
        F.append(Fk)
        k += 1
    return F, support_data
```

单元素候选集生成函数：

```
def C1_generation(self, dataset):       #生成单元数候选项集
    C1 = []
    progress = ProgressBar()
    for data in progress(dataset):
        for item in data:
            if [item] not in C1:
                C1.append([item])
    return [frozenset(item) for item in C1]
```

过滤低支持度函数：

```
def Ck_low_support_filtering(self, dataset, Ck):        #过滤支持度低于阈值的项集
    Ck_count = dict()
    for data in dataset:
        for cand in Ck:
            if cand.issubset(data):
                if cand not in Ck_count:
                    Ck_count[cand] = 1
                else:
                    Ck_count[cand] += 1

    num_items = float(len(dataset))
    return_list = []
    support_data = dict()
    # 过滤非频繁项集
    for key in Ck_count:
        support = Ck_count[key] / num_items
        if support >= self.min_sup:
            return_list.insert(0, key)
        support_data[key] = support
    return return_list, support_data
```

合并筛选函数：

```
def apriori_gen(self, Fk, k):        #当候选项元素大于2时，合并时检测是否子项集满足频繁
    return_list = []
    len_Fk = len(Fk)

    for i in range(len_Fk):
        for j in range(i+1, len_Fk):
            # 第k-2个项相同时，将两个集合合并
            F1 = list(Fk[i])[:k-2]
            F2 = list(Fk[j])[:k-2]
            F1.sort()
            F2.sort()
            if F1 == F2:
                return_list.append(Fk[i] | Fk[j])
    return return_list
```

最终产生的频繁项集结果保存在“频繁项集.json”文件中，效果如下图所示：

3. 导出关联规则（包含评价）

基于 2 中使用 Apriori 算法得出的频繁项集，我们计算关联规则以及使用评价指标来评价它们。本实验使用的是课件中的 Lift 和 Jaccard 两种指标进行评价。

Measure	Definition	Range	Null-Invariant?
$\chi^2(A,B)$	$\sum_{i,j}\frac{(e(a_i,b_j)-o(a_i,b_j))^2}{e(a_i,b_j)}$	$[0,\infty]$	No
$Lift(A,B)$	$\frac{s(A\cup B)}{s(A)\times s(B)}$	$[0,\infty]$	No
$Allconf(A,B)$	$\frac{s(A\cup B)}{max\{s(A),s(B)\}}$	$[0,1]$	Yes
$Jaccard(A,B)$	$\frac{s(A\cup B)}{s(A)+s(B)-s(A\cup B)}$	$[0,1]$	Yes
$Cosine(A,B)$	$\frac{s(A\cup B)}{\sqrt{s(A)\times s(B)}}$	$[0,1]$	Yes
$Kulczynski(A,B)$	$\frac{1}{2}(\frac{s(A\cup B)}{s(A)}+\frac{s(A\cup B)}{s(B)})$	$[0,1]$	Yes
$MaxConf(A,B)$	$max\{\frac{s(A\cup B)}{s(A)},\frac{s(A\cup B)}{s(B)}\}$	$[0,1]$	Yes

其中计算的公式为：

支持度：

$$Sup(X) = \frac{count(X)}{all\_data}$$

置信度：

$$conf(X \to Y) = \frac{Sup(X \cup Y)}{Sup(X)}$$

Lift:

$$left(X \to Y) = \frac{Sup(X \cup Y)}{Sup(X) * Sup(Y)}$$

Jaccard:

$$\text{Jaccard}(X \to Y) = \frac{Sup(X \cup Y)}{Sup(X) + Sup(Y) - Sup(X \cup Y)}$$

以上计算对应的代码为：

```
def cal_conf(self, freq_set, H, sup_rata, strong_rules_list):  # 评估规则
    prunedH = []
    for reasoned_item in H:
        sup = sup_rata[freq_set]
        conf = sup / sup_rata[freq_set - reasoned_item]
        lift = conf / sup_rata[reasoned_item]
        jaccard = sup / (sup_rata[freq_set - reasoned_item] + sup_rata[reasoned_item] - sup)
        if conf >= self.min_conf:
            strong_rules_list.append(
                (freq_set - reasoned_item, reasoned_item, sup, conf, lift, jaccard))
            prunedH.append(reasoned_item)
    return prunedH
```

计算并保留达到置信度阈值的函数为：

```
def generate_rules(self, F, sup_rata):
    strong_rules_list = []
    for i in range(1, len(F)):
        for freq_set in F[i]:
            H1 = [frozenset([item]) for item in freq_set]
            if i > 1:
                self.rules_from_reasoned_item(freq_set, H1, sup_rata, strong_rules_list)
            else:
                self.cal_conf(freq_set, H1, sup_rata, strong_rules_list)
    return strong_rules_list
```

4. 分析挖掘结果

由于电脑问题，跑完全部数据花费的时间实在是太长了。由下图可见，原本预计的时间是 5 小时能跑完，但跑了一晚上之后发现还需要 13 小时。于是我放弃跑完全部数据，取前 50000 个数据进行实验。

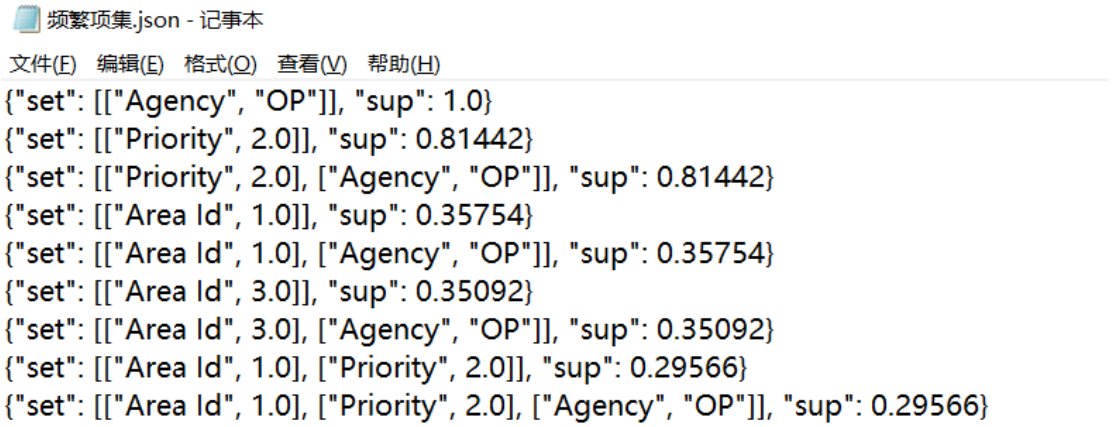
```
综合数据集有以下属性 Index(['Agency', 'Location', 'Area Id', 'Beat', 'Priority', 'Incident Type Id',
       'Incident Type Description', 'Event Number'],
      dtype='object')
 74% (637973 of 859898) |############      | Elapsed Time: 8:51:17 ETA:  12:50:05
```

```
172             return data_all.head(50000)
```

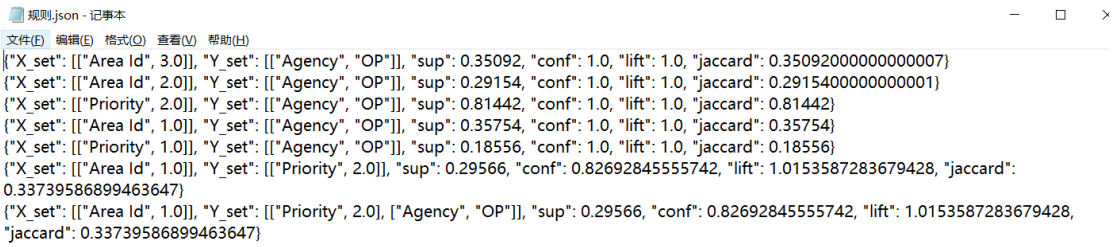
更换后的时间如下所示，只需几分钟：

```
综合数据集有以下属性 Index(['Agency', 'Location', 'Area Id', 'Beat', 'Priority', 'Incident Type Id',
       'Incident Type Description', 'Event Number'],
      dtype='object')
100% (50000 of 50000) |####################| Elapsed Time: 0:01:43 Time:  0:01:43
```

我们将得到的频繁项集放到了./results/频繁项集.json 文件中，按照支持度由大到小排列，形式如下图所示：



将得到的关联规则以及评价结果放到了./results/规则.json 文件中，按照置信度由大到小排列，形式如下图所示：



由于所有的 Agency 属性的值都是 OP，所以对其分析没有实际意义，我们跳过包含 Agency 属性的频繁项集与规则进行分析。

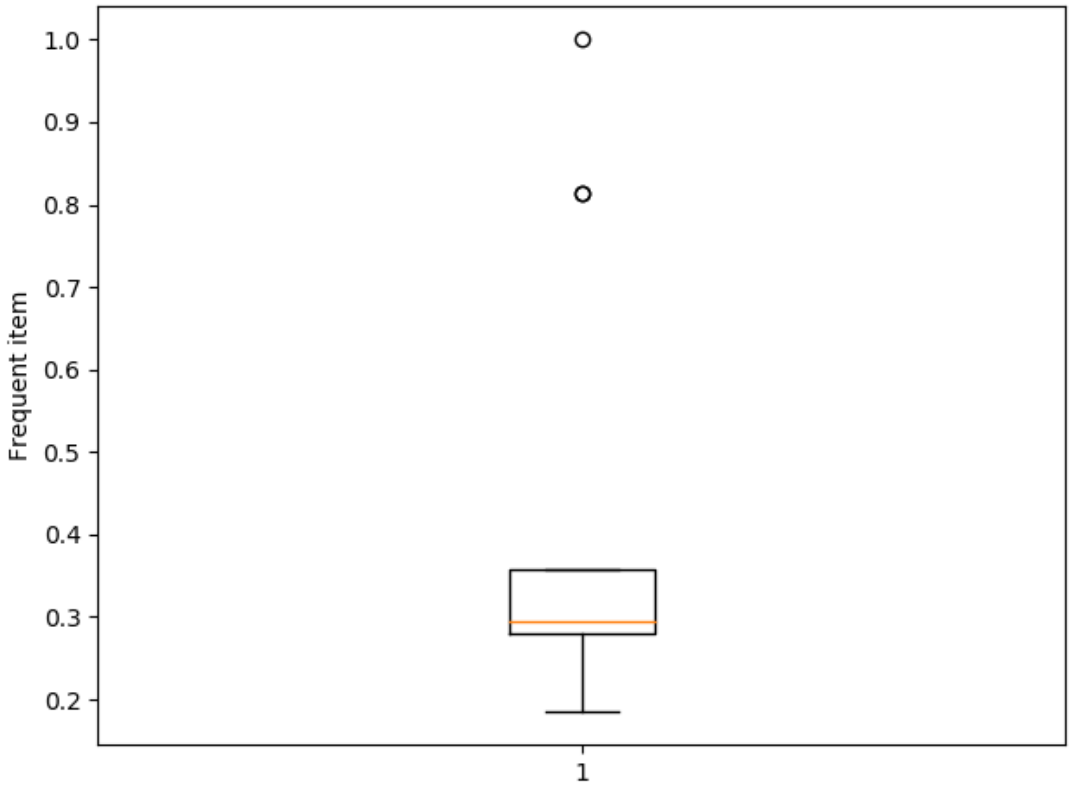
我们可以由频繁项集.json 得知，Area Id 为 1.0 时支持度最高，也就是说在该地区的犯罪事实出现最多。而且 Area Id 和 Priority 的关联度较高。

我们可以由规则.json 得知，["Area Id", 1.0]与["Priority", 2.0]的置信度较高，这说明犯罪的严重性与所在地有着较强联系。

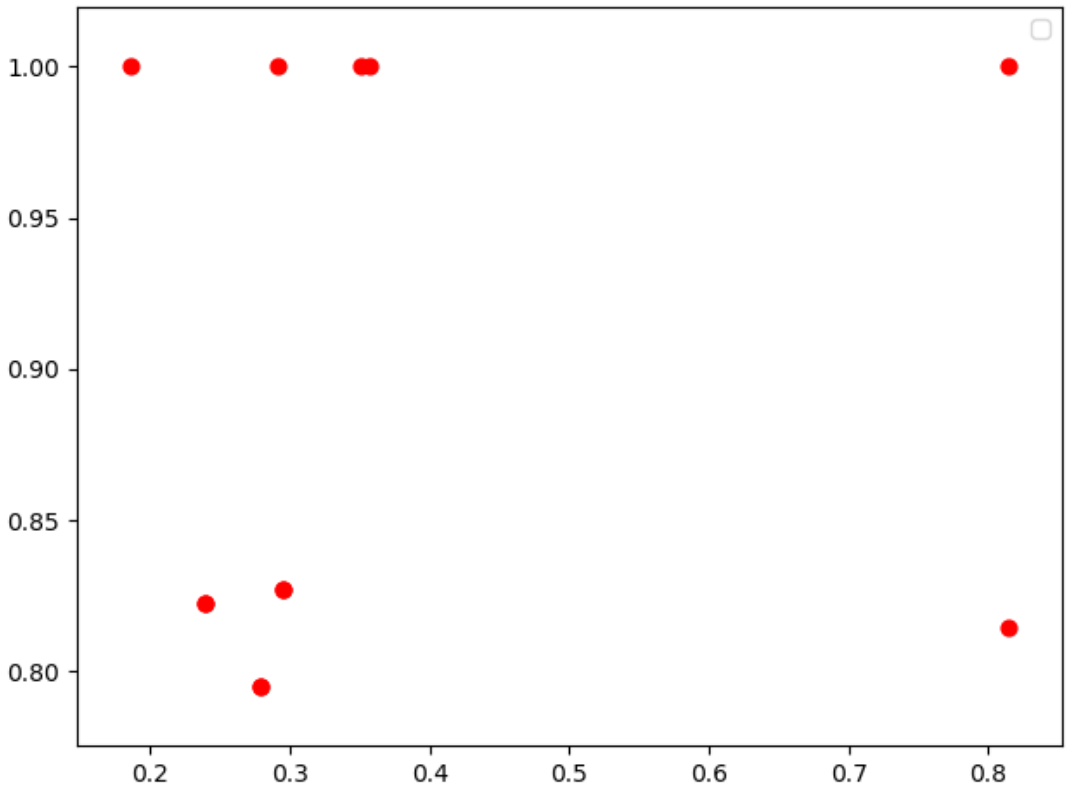
5. 可视化

分别使用盒图与散点图对频繁项集与规则进行可视化。

对频繁项集使用盒图可视化可得：



对规则的指出度与置信度使用散点图可视化可得：


1.00
0.95
0.90
0.85
0.80
0.2
0.3
0.4
0.5
0.6
0.7
0.8