数据挖掘：关联规则挖掘

学院：计算机学院
专业：生物医学工程
姓名：陈晓珍
学号：2120170997

一. 实验环境

电脑：64 位 CPU：Intel5 Memory：8G
系统：Ubuntu14.04
语言：python

二. 数据集

由于使用的计算机是单机，没有安装集群，在实验的时候选择了两个数据集中数据量较小的数据集，有 19980 条数据记录的 Building_Permits 的数据集。

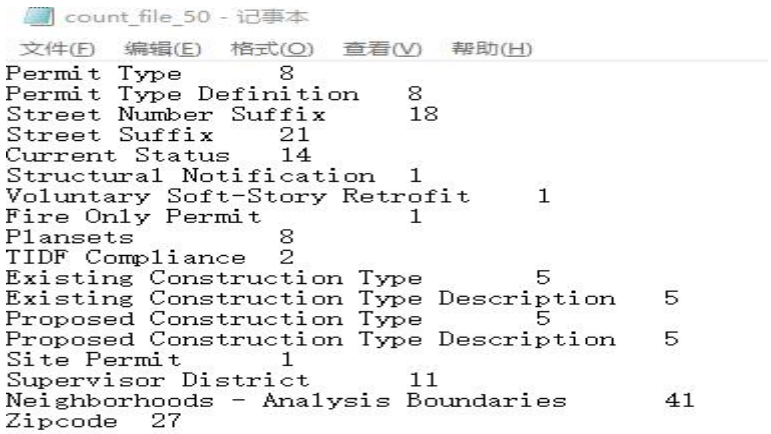
三. 实验

3.1 数据预处理

数据预处理的目的是将数据集进行处理以满足适合关联规则挖掘的形式。在此，对 Building_Permits 数据集进行剪切。一方面便于电脑处理数据速度快，一方面用于适合关联规则挖掘。

数据集中有许多数据空白，需要对数据集进行填充。

由于数据量比较大，单个电脑处理的数据速度比较慢，对数据进行部分处理，选择数据项内容比较少的进行关联规则挖掘。对每一项进行统计，输出了内容种类少于 50 的进行处理。具体文件可以见 count_file_50.txt 中。



3.2 频繁项集

发现关联规则要求项集必须满足的最小支持阈值，称为项集的最小支持度

（Minimum Support），记为 supmin。支持度大于或等于 supmin 的项集称为频繁项集，简称频繁集，反之则称为非频繁集。通常 k-项集如果满足 supmin，称为 k-频繁集，记作 Lk。关联规则的最小置信度（Minimum Confidence）记为 confmin，它表示关联规则需要满足的最低可靠性。

3.3 支持度，置信度，关联规则

支持度（suppport）：是交易集中同时包含 A 和 B 的交易数与所有交易数之比。表示所有集中，既有 A 又有 B 的概率。Support(A=>B)=P(A∪B)=count(A∪B)/|D|

置信度（confidence）：是包含 A 和 B 交易数与包含 A 的交易数之比。表示了这条规则有多大程度上值得可信。设条件的项的集合为 A,结果的集合为 B。置信度计算在 A 中，同时也含有 B 的概率（即：if A ,then B 的概率）。即 ：Confidence(A=>B)=P(B|A)=support(A∪B)/support(A)

关联规则（Association Rules）是反映一个事物与其他事物之间的相互依存性和关联性，如果两个或多个事物之间存在一定的关联关系，那么，其中一个事物就能通过其他事物预测到。关联规则是数据挖掘的一个重要技术，用于从大量数据中挖掘出有价值的数据项之间的相关关系。如果规则 R：X=>Y 满足 support(X=>Y) >= supmin 且 confidence(X=>Y)>=confmin，称关联规则 X=>Y 为强关联规则，否则称关联规则 X=>Y 为弱关联规则。、

3.4 APRIORI 算法

Apriori 算法是一种对有影响的挖掘布尔关联规则频繁项集的算法，通过算法的连接和剪枝即可挖掘频繁项集。

Apriori 算法将发现关联规则的过程分为两个步骤：

1.通过迭代，检索出事务数据库中的所有频繁项集，即支持度不低于用户设定的阈值的项集；

2.利用频繁项集构造出满足用户最小置信度的规则。

3.4.1 基本思想

Apriori 算法基本思想是通过对数据库的多次扫描来计算项集的支持度，发现所有的频繁项集从而生成关联规则。Apriori 算法对数据集进行多次扫描。第一次扫描得到频繁 1-项集的集合 L1，第 k（k>1）次扫描首先利用第（k-l）次扫描的结果 Lk 来产生候选 k-项集的集合 Ck，然后再扫描的过程中确定 Ck 中元素的支持度，最后再每一次扫描结束时计算频繁 k-项集的集合 Lk，算法当候选 k-项集的集合 Ck 为空时结束。

3.4.2 产生频繁项集的过程

产生频繁项集的过程主要分为连接和剪枝两步：

（1）连接步。为找 Lk，通过 Lk-1 与自身作连接产生候选 k-项集的集合 Ck。设和是 Lk-1 中的项集。记表示的第 j 个项。Apriori 假定事务或项集中的项按字典次序排序。对于（k-1）项集，意味将项排序，使< <…<。如果 Lk-1 的元素和的前（k-2）个对应项相等，则和可连接。即，如果（=）∩（=）∩…∩（=）∩（<）时，和可连接。条件<仅仅是保证不重复。连接和产生的结果项集为（，，…，，）。

（2）剪枝步。Apriori 算法的性质可知，频繁 k-项集的任何子集必须是频繁项集。由连接生成的集合 Ck 需要进行验证，去除不满足支持度的非频繁 k-项集。

3.4.3 主要步骤

(1)扫描全部数据，产生候选 1-项集的集合 C1；

(2)根据最小支持度，由候选 1-项集的集合 C1 产生频繁 1-项集的集合 L1；

(3)对 k>1，重复执行步骤④、⑤、⑥；

(4)由 Lk 执行连接和剪枝操作，产生候选（k+l）-项集的集合 Ck+1；

(5)根据最小支持度，由候选（k+l）-项集的集合 Ck+1，产生频繁（k+1）-项集的集合 Lk+1；

(6)若 L≠Φ，则 k=k+1，跳往步骤④；否则，跳往步骤⑦；

(7)根据最小置信度，由频繁项集产生强关联规则，结束。

3.5 关联规则评价及分析

提升度是指规则的支持度与规则后件出现的概率比值，反应规则前件和后件之间的正负相关性，公式：lift(A->B)=c(A->B)/P(B)。

实验中设置的最小支持度是 0.2，最小置信度 0.5.

计算出的支持度部分结果如下图，详细结果见 suppfile_150_0.20_0.5.txt。

```
item: ('complete', 'otc alterations permit', 'wood frame (5)', 'St')
      , 0.200
item: ('complete', '8', 'wood frame (5)', 'St')       , 0.200
item: ('complete', 'otc alterations permit', 'wood frame (5)', 'St',
'8')  , 0.200
item: ('5.0', 'complete', 'wood frame (5)', 'St', '8')      , 0.200
item: ('5.0', 'complete', 'otc alterations permit', 'wood frame (5)',
'St')       , 0.200
item: ('', 'complete', 'wood frame (5)', 'St', '8')   , 0.200
item: ('', 'complete', 'otc alterations permit', 'wood frame (5)', 'St')
      , 0.200
item: ('', 'complete', 'otc alterations permit', 'wood frame (5)', 'St',
'8')  , 0.200
item: ('5.0', '', 'complete', 'wood frame (5)', 'St', '8')  , 0.200
item: ('5.0', '', 'complete', 'otc alterations permit', 'wood frame
(5)', 'St')       , 0.200
item: ('5.0', 'complete', 'otc alterations permit', 'wood frame (5)',
'St', '8')  , 0.200
item: ('5.0', '', 'complete', 'otc alterations permit', 'wood frame
(5)', 'St', '8')  , 0.200
item: ('otc alterations permit', 'complete', '3.0')   , 0.203
item: ('8', 'complete', '3.0')      , 0.203
item: ('', 'complete', '3.0', 'otc alterations permit')     , 0.203
item: ('complete', '3.0', 'otc alterations permit', '8')    , 0.203
item: ('', 'complete', '3.0', '8')  , 0.203
item: ('', 'complete', 'otc alterations permit', '3.0', '8')      ,
0.203
item: ('2.0', '3.0', 'St')    , 0.207
item: ('', '3.0', '2.0', 'St')      , 0.207
item: ('1 family dwelling', '2.0', 'wood frame (5)')  , 0.207
item: ('', '1 family dwelling', '2.0', 'wood frame (5)')    , 0.207
item: ('5.0', '1 family dwelling', '2.0', 'wood frame (5)')       ,
0.207
```

计算出的置信度和 lift 部分结果如下图，详细结果见 conliftfile_150_0.20_0.5.txt。

```
Rule: ('2.0', 'St') ==> ('otc alterations permit', 'complete')     ,
confidence: 0.502,lift: 1.105
Rule: ('2.0', 'St') ==> ('8', 'complete')  , confidence: 0.502,lift:
1.105
Rule: ('2.0', 'St') ==> ('', 'otc alterations permit', 'complete')
       , confidence: 0.502,lift: 1.105
Rule: ('', '2.0', 'St') ==> ('otc alterations permit', 'complete')
       , confidence: 0.502,lift: 1.105
Rule: ('2.0', 'St') ==> ('otc alterations permit', 'complete', '8')
       , confidence: 0.502,lift: 1.000
Rule: ('2.0', 'St') ==> ('', '8', 'complete')    , confidence:
0.502,lift: 1.105
Rule: ('', '2.0', 'St') ==> ('8', 'complete')    , confidence:
0.502,lift: 1.105
Rule: ('2.0', 'St') ==> ('', 'otc alterations permit', 'complete', '8')
       , confidence: 0.502,lift: 1.000
Rule: ('', '2.0', 'St') ==> ('otc alterations permit', 'complete', '8')
       , confidence: 0.502,lift: 1.000
Rule: ('otc alterations permit', 'St') ==> ('wood frame (5)',)     ,
confidence: 0.503,lift: 0.864
```

四．挖掘结果

4.1 数据挖掘处理源程序

见 propress.py。

4.2 关联规则挖掘源程序

见 Apriori.py。