**数据挖掘大作业二：关联规则挖掘**

殷赫 生物医学工程 2220170871

**一、实验内容**

1. 对数据集进行处理，转换成适合关联规则挖掘的形式；
2. 找出频繁项集；
3. 导出关联规则，计算其支持度和置信度
4. 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

**二、实验程序**

**1）数据预处理部分**

由于电脑性能限制，这里只选取部分实验数据，选用Building\_Permits.csv数据集的

Permit Type、Current Status、Permit Type Definition三个属性，并对数据进行抽样选取，最后将数据的内容转化为数字标签便于进行关联分析。

def stats\_str(df):

result=dict(co.Counter(df))

return result

df = pd.DataFrame({"Permit Type":data['Permit Type'],

"Permit Type Definition":data['Permit Type Definition'],

"Current Status":data['Current Status']},

columns =['Permit Type','Permit Type Definition','Current Status'])

data2=df.values[0::1000,:]

dataindex=list(pd.Series(stats\_str(data2[:,0])).index)+list(pd.Series(stats\_str(data2[:,1])).index)+list(pd.Series(stats\_str(data2[:,2])).index)

for i in range(len(dataindex)):

data2[data2==dataindex[i]]=i

**2）关联分析程序**

a.createC1程序作用提取数据集中的相集

def createC1(dataSet):

C1 = []

for transaction in dataSet:

for item in transaction:

if not [item] in C1:

C1.append([item])

C1.sort()

return map(frozenset, C1)

b.scanD程序作用，计算支持度，并根据阈值进行剪枝

def scanD(D,Ck,minSupport):

ssCnt = {}

for tid in D:

for can in Ck:

if can.issubset(tid):

if not can in ssCnt:

ssCnt[can] = 1

else:

ssCnt[can] +=1

numItems = float(len(D))

retList = []

supportData = {}

for key in ssCnt:

support = ssCnt[key]/numItems

if support >= minSupport:

retList.insert(0,key)

supportData[key] = support

return retList, supportData

c. aprioriGen函数作用是进行联合，根据上一层的相集生成下一层新的相集

def aprioriGen(Lk, k):

retList = []

lenLk = len(Lk)

for i in range(lenLk):

for j in range(i+1, lenLk):

L1 = list(Lk[i])[:k-2]

L2 = list(Lk[j])[:k-2]

L1.sort()

L2.sort()

if L1==L2:

retList.append(Lk[i]|Lk[j])

return retList

d. apriori函数根据apriori算法原理对数据进行关联性分析

def apriori(dataSet, minSupport = 0.5):

C1 = list(createC1(dataSet))

D = list(map(set,dataSet))

L1, supportData = scanD(D,C1,minSupport)

L = [L1]

k = 2

while (len(L[k-2])>0):

Ck = aprioriGen(L[k-2], k)

Lk, supK = scanD(D, Ck, minSupport)

supportData.update(supK)

L.append(Lk)

k +=1

return L,supportData

L, suppData = apriori(data2)

**三、实验结果**

**1）实验数据预处理结果**

实验内容对应的数字标签，便于进行关联性分析

0：1,1：2,2：3,3：4,4：8、,5：'additions alterations or repairs',6：'new construction',7：'new construction wood frame',8：'otc alterations permit',9：'sign - erect',10：'cancelled',11：'complete',12：'expired',13：'filed',14：'issued',15：'reinstated',16：'withdrawn'

**2）频繁项集**

frozenset({11}), frozenset({8}), frozenset({4}), frozenset({4, 8})

对应的属性分别为complete、otc alterations permit、8、8和otc alterations permit

**3）关联规则**

frozenset({3}): 0.01507537688442211,

frozenset({9}): 0.01507537688442211,

frozenset({12}): 0.01507537688442211,

frozenset({4}): 0.8793969849246231,

frozenset({8}): 0.8793969849246231,

frozenset({14}): 0.3316582914572864,

frozenset({11}): 0.5527638190954773,

frozenset({13}): 0.08542713567839195,

frozenset({2}): 0.09045226130653267,

frozenset({5}): 0.09045226130653267,

frozenset({10}): 0.005025125628140704,

frozenset({1}): 0.010050251256281407,

frozenset({7}): 0.010050251256281407,

frozenset({16}): 0.005025125628140704,

frozenset({15}): 0.005025125628140704,

frozenset({0}): 0.005025125628140704,

frozenset({6}): 0.005025125628140704,

frozenset({4, 8}): 0.8793969849246231,

frozenset({8, 11}): 0.49246231155778897,

frozenset({4, 11}): 0.49246231155778897

可以得出关联规则：

S(4->8)=0.0517

C(4->8)=100%

**4）规则评价**

lift(A,B)=P(A交B)/(P(A)\*P(B))

lift(4,8)=P(4交8)/(P(4)\*P(8))= 1.137142857142857正相关