数据挖掘第二次作业

关联规则挖掘

曹倩雯

3120170497

关联规则挖掘结果及分析报告

1 基本任务：

- 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；

2 数据预处理

- 语言环境
  python 语言

- 数据集描述：
  选择数据集二，San Francisco Building Permits 进行关联规则挖掘

- 数据预处理：

为所有属性设置属性名分别为 a1、a2、……。在这些属性中，大体分为两种：标称属性和数值属性，其中数值属性不适合进行关联规则挖掘，因此，为了方便 Aprior 算法的处理，只对标称属性进行处理，即 Current Status，Structural Notification，Number of Proposed Stories，Fire Only Permit，TIDF Compliance，Proposed Construction Type，Site Permit 这七组。

但是这些数据的缺失值较为严重，如果仅仅单独取出这七组数据，部分数据展示如下：

	A	B	C	D	E	F	G
1	Structural I	Fire Only P	Site Permit	TIDF Comp	Proposed (	Number o	Proposed
2							
3							
4					constr type	6	
5					wood fram	2	
6							
7		Y			constr type	5	
8					wood fram	3	
9							

因此，需要利用数据间的相关性，填补这部分的缺失值，对于 Proposed Construction Type，使用众数进行填补；对于 Number of Proposed Stories，使用中位数进行填补，其余缺失值填补为设置好的空值或零值，部分数据展示如下：

	A	B	C	D	E	F	G
1	Structural I	Fire Only P	Site Permit	TIDF Comp	Proposed (	Number of	Proposed
2	None1	None2	None3	None4	wood fram	3	None5
3	None1	None2	None3	None4	wood fram	3	None5
4	None1	None2	None3	None4	constr type	6	None5
5	None1	None2	None3	None4	wood fram	2	None5
6	None1	None2	None3	None4	wood fram	3	None5
7	None1	Y	None	None4	constr type	5	None5
8	None1	None2	None3	None4	wood fram	3	None5
9	None1	None2	None3	None4	wood fram	3	None5

将得到的 csv 文件保存，用来做后续关联挖掘。

3　频繁规则统计与计算

- 找出频繁项集：

本次数据挖掘使用了 Apriori 算法，如果要发现强关联规则，就必须先找到频繁集。所谓频繁集，即支持度大于最小支持度的项集。本次数据挖掘令频繁项集的最小置信度阈值为 0.7、最小支持度阈值为 0.7，挖掘出的频繁项集结果为：

	items	support	Count
[1]	{large}	0.9999943	198900
[2]	{large, no}	0.9999854	198852
[3]	{no}	0.9999254	198812
[4]	{large, None2}	0.9721562	193542
[5]	{large, None5}	0.9701022	193522
[6]	{large, woodframe}	0.9625435	192121
[7]	{no, constr type}	0.9610258	191756

4 关联规则

设置关联规则的最小置信度阈值为 0.9，导出的关联规则为，部分如下图所示：

	association rules	lift
[1]	{3}--->{None4}	1.0335612
[2]	{3}--->{None1}	1.0221012
[3]	{3}--->{None2}	0.9988542
[4]	{3}--->{None5}	0.9412522
[5]	{wood frame}--->{None1}	1.0000521
[6]	{wood frame}--->{None2}	2.1525452
[7]	{wood frame}--->{None4}	1.0212422
[8]	{wood frame}--->{None3}	1.0052527