关联规则挖掘

分析报告

赵文天 2120171105

一. 数据源以及数据预处理

1. 数据源

本次使用的数据集为 San Francisco Building Permits 数据集。该数据集种共有 43 个属性，198900 条数据。使用了数据集中的如下 11 个标称属性：

Permit Type
Curent Status
Existing Use
Existing Construction Type
Proposed construction type
Supervisor District
Neighborhoods - Analysis Boundaries
Zipcode
Structural Notification
Fire Only Permit
TIDF Compliance

2. 数据预处理

在预处理过程中，将每个样本中的所有不同的属性作为样本包含的不同项。这里忽略了所有缺失的属性。若两个不同属性的值相同，这两个属性也被看作不同的项。

二. 频繁项集以及关联规则

1. 频繁项集

我们使用 FP-tree 算法（FP growth 算法）提取频繁项集以及对关联规则进行挖掘，使用了 IBM developerworks 提供的 FP growth 算法实现。

在建立 FP 树后，对于项头表中的每个元素获取其在 FP 树中的前缀路径，构造对应的子树，构造条件模式基，从条件模式基递归挖掘得到频繁项集。这里没有对项数进行限制，最终得到所有的频繁项集。

在使用 FP 树进行频繁项集提取预计关联规则挖掘时，设置频繁项集的最小出现次数为 100，关联规则置信度最小值为 0.6。

所有频繁项集的结果在 frequent_pattern.json 文件中。该文件每一行为一个 JSON object，item 属性记录了频繁项的内容，项之间用逗号隔开。Support 属性记录了频繁项的支持度，即该项出现的次数占数据集中样本数的比例：

$$support(X) = \frac{sum(X)}{N\_samples}$$

其中，N_samples 为数据集中的样本总数。

例如，按支持度从大到小的顺序排序，前 5 个频繁项如下：

表 1 挖掘得到的部分频繁项集

项名称	支持度
Permit Type:8	0.05549

Existing Construction Type:5	0.04133
Proposed Construction Type:5	0.04080
Existing Construction Type:5, Proposed Construction Type:5	0.03744
Current Status:complete	0.03348

由表 1 可知，挖掘得到的支持度最大的频繁项集的支持度为 5.549%，可能的原因是原数据集中每个属性的值分布较为分散，一个属性中的同一个值出现的次数都较少。

2. 关联规则挖掘

在构建频繁项集后，根据频繁项集构建所有可能的关联规则，并计算其置信度，对每条置信度大于最小置信度的关联规则还要计算其支持度和提升度。关联规则是形如 X→Y 的蕴含式。

一条关联规则支持度，置信度以及提升度的计算方式如下：

支持度为项集 X∩Y 出现的概率：

$$support(X \to Y) = \frac{sum(X \cap Y)}{N\_samples}$$

置信度为包含 X 的事务中同时包含 Y 的比例：

$$confidence(X \to Y) = \frac{sum(X \cap Y)}{sum(X)}$$

提升度：表示事务包含 X 的条件下同时包含 Y 的概率与事务包含 Y 的概率之比，表示关联规则中 X 与 Y 的相关性，用于衡量关联规则是否有效。提升度>1 表示 X 与 Y 有正相关性，提升度<1 表示 X 与 Y 有负相关性，提升度=1 表示 X 与 Y 相互独立。其计算方法为：

$$lift(X \to \mathrm{Y}) = \frac{P(Y|X)}{P(Y)} = \frac{sum(X \cap Y)}{sum(X)sum(Y)}$$

挖掘得到的所有关联规则结果在 association_rules.csv 文件中。该文件每一行表示一条关联规则，x 属性为关联规则的前件，y 属性为关联规则的后件。项集中的每个项用逗号隔开，每项的格式为“属性名:值”。confidence, support, lift 属性分别为置信度，支持度以及提升度。

3. 关联规则评价

按支持度降序将关联规则排序后，得到的前几个关联规则如下：

表 2 挖掘得到的部分频繁项集

X	Y	support	confidence	lift
Proposed Construction Type:5	Existing Construction Type:5	0.037436	0.917447	22.19954
Existing Construction Type:5	Proposed Construction Type:5	0.037436	0.905839	22.19954
Proposed Construction Type:5	Permit Type:8	0.026657	0.653277	11.77284
Existing Construction Type:5	Permit Type:8	0.025304	0.612287	11.03415
Proposed Construction Type:5	Permit Type:8, Existing Construction Type:5	0.024897	0.610153	24.11273

由表 2 中的示例可知，挖掘得到的关联规则的置信度和提升度均较高，但支持

度较低。这说明挖掘得到的规则都是有效规则。支持度普遍较低的原因可参考对表1的说明，频繁项集的支持度均较低。