2024 春 数据分析及实践
实验四: 乳腺癌疾病预后关联规则挖掘分析

2024 年 5 月 16 日

马天开

tiankaima@mail.ustc.edu.cn

ID: 3 / PB2100030

读取数据、预处理

Q1. 缺失值

```
Missing values:
id              0
Class           0
age             0
menopause       0
tumor-size      0
inv-nodes       0
node-caps       8
deg-malig       0
breast          0
breast-quad     1
irradiat        0
```

Q2. 异常值

```
Unique values in tuomor-size:
['30-34' '20-24' '15-19' '0-4' '25-29' '50-54' '14-Oct' '40-44' '35-39' '9-May'
'45-49']
Unique values in inv-nodes:
['0-2' '8-Jun' '11-Sep' '5-Mar' '15-17' '14-Dec' '24-26']
```

手动修正方法:

- `tumor-size`:
  - `'14-Oct'` → `'10-14'`
  - `'9-May'` → `'9'`
- `inv-nodes`:
  - `'11-Sep'` → `'9-11'`
  - `'5-Mar'` → `'3-5'`
  - `'14-Dec'` → `'12-14'`

Q3. 数字索引替换

需要将所有 `str` 转换为 `int` 方便后续处理; 部分 `col` 值有重叠, 我们维护一个 `dict` 以便后续还原.

```
{
 0: 'Class=no-recurrence-events',
 1: 'Class=recurrence-events',
 2: 'age=30-39',
 // ...
}
```

关联规则挖掘

Q1. 实现 Apriori; 计算频繁项集 (threshold = 0.4)

```
[
  [{0}, {1}, {2}, {3}, {8}, {9}, {22}, {29}, {31}, {32}, {38}],
  [{0, 22}, {0, 29}, {0, 38}, {2, 29}, {2, 38}, {3, 29}, {8, 22}, # ... ],
  [{0, 29, 22}, {0, 38, 22}, {0, 29, 38}, {38, 29, 22}], [{0, 22, 38, 29}]
]
```

Q2. 挖掘 $X \to \{0\}$ 的强关联规则 (min_conf = 0.75)

Q3. 利用 `ind2val` 还原关联规则

关联规则	置信度	提升度
`[inv-nodes=0-2]` ⇒ `no-recurrence-events`	79.4%	1.12
`[node-caps=no]` ⇒ `no-recurrence-events`	77.4%	1.09
`[irradiat=no]` ⇒ `no-recurrence-events`	76.3%	1.08
`[age=30-39, node-caps=no]` ⇒ `no-recurrence-events`	77.1%	1.09
`[age=30-39, irradiat=no]` ⇒ `no-recurrence-events`	77.5%	1.09
`[node-caps=no, inv-nodes=0-2]` ⇒ `no-recurrence-events`	80.0%	1.13
`[inv-nodes=0-2, breast=left]` ⇒ `no-recurrence-events`	82.1%	1.16
`[irradiat=no, inv-nodes=0-2]` ⇒ `no-recurrence-events`	81.7%	1.15
`[node-caps=no, breast=left]` ⇒ `no-recurrence-events`	79.3%	1.12
`[node-caps=no, irradiat=no]` ⇒ `no-recurrence-events`	80.7%	1.14
`[irradiat=no, breast=left]` ⇒ `no-recurrence-events`	78.3%	1.11
`[node-caps=no, irradiat=no, inv-nodes=0-2]` ⇒ `no-recurrence-events`	82.4%	1.16