基于决策树方法的专利被引影响因素研究

钟正 谢涛

本次研究基于决策树方法对可能影响专利被引的6个影响因素（专利申请年、专利公开年、发明人数量、专利权人数量、发明人国家、专利权人国家）与专利是否被引的潜在关系进行分析。研究发现，专利申请年是其中影响最为显著的因素，而后依次是发明人数量、专利公开年数量、专利权人数量，而其他2个指标的影响效果并不明显。

|  |  |  |
| --- | --- | --- |
| 影响因素 | 简介 | 可能对专利被引事件影响 |
| 专利申请年 | 指专利申请时的年份 | 随着时间的推移，申请年份越早的专利更可能被引用 |
| 专利公开年 | 指专利公开时的年份 | 随着时间的推移，公开年份越早的专利更可能被引用 |
| 发明人 | 指为发明创造的实质性特点做出创造性贡献的人 | 理论上讲，参与专利的发明人数量越多，发明的凝结技术就越多，创新性就越高，也容易被引用 |
| 发明人国家 | 指发明人所在的地区 | 研究证明，专利引用行为与地理位置存在关系 |
| 专利权人国家 | 指专利权人所在的地区 | 同上 |

表1 专利被引影响因素

1 数据来源与指标选取

* 1. 数据来源

本研究选取美国专利号为**43518**开头的专利数据作为研究对象，获得专利数据**1507**条，经过数据清理后得到**454**条标准化数据。

* 1. 数据指标选取

根据可能对专利被引产生影响的因素列表，拟定决策树的分析指标共6个，分别是：专利申请年PATN:APD、专利公开年PATN:ISD、发明人数量INVTSUM、专利权人数量ASSGSUM、发明人国家INVT:CNT、专利权人国家ASSG:CNT，下表是描述统计。

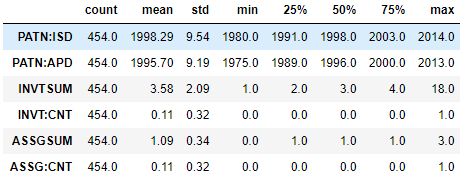


表2 分析指标描述性统计

1. 数据清洗与转换

运用python对数据进行数据预处理

主要核心操作有：

**数据清洗**

data.isnull().any(axis=0) //查看空值

data.isnull().any(axis=1)//判断数据行中是否存在缺失值

data.loc[data.isnull().any(axis=1)]//定位缺失值所在行

data3=data.dropna()

data//直接删除空值所在行

data\_null\_0=data.fillna(0,inplace=False)//定义缺失值为零

data.drop(index=1,axis=1)//删除特定行

**数据转换**

对发明人及专利人国家不是US的标记为非US

核心代码为：

data['INVT：CNT']=data['INVT：CNT'].str.replace('JPX','非US')

data

得到数据如下图

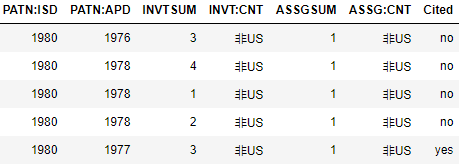


图1 待分析数据源表

1. 决策树模型构建及准确性评估、优化
   1. 模型构建

选取特征值和分析指标，并对数据做标化处理

核心代码：

target = df\_zhuanli['Cited']

data=df\_zhuanli.loc[:,'PATN:ISD':'ASSG:CNT']

得下图

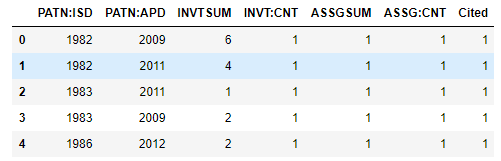


图2 待分析标化数据

**建立训练集和数据集，按照20%和80%区分**

核心代码

from sklearn.model\_selection import train\_test\_splittrain\_data, test\_data, train\_target, test\_target = train\_test\_split(data,\ target, test\_size=0.2, train\_size=0.8 ,random\_state=1234)

**建立决策树 深度为4 使用信息熵来衡量集合纯度**

核心代码：

from sklearn.tree import DecisionTreeClassifierclf = DecisionTreeClassifier(criterion='entropy', max\_depth=4, class\_weight=None, random\_state=1234 ) clf.fit(train\_data,train\_target)

* 1. 准确性评估与优化

对于分类问题，我们建立混淆矩阵通过计算Accuracy、Precision、Recall、F1-score四个指标对模型进行评估

核心代码

import sklearn.metrics as metrics

print(metrics.classification\_report(test\_target, clf.predict(test\_data)))

得下图

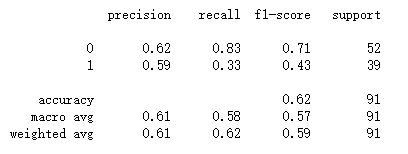


图3 模型评估指标

**由图中可知对于被引用=1情况，f1-score只有0.43，召回率只有0.33，说明在被引用的专利中，模型只能找出其中的33%，损失极大，所以提高被引用专利的权重**

核心代码如下：

clf.set\_params(\*\*{'class\_weight':{0:1,1:5}})//**调整被引用的权重为未被引用的五倍**

clf.fit(train\_data,train\_target)

print(metrics.classification\_report(test\_target, clf.predict(test\_data)))

得下图

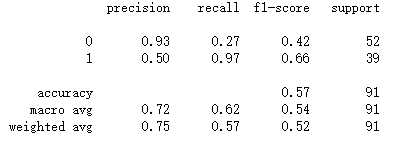


图4 模型优化后评估指标

可知，**优化后召回率达到97%**，符合预期，模型可行

1. 分析结果

计算出影响较大的特征值：

核心代码：

list(zip(data.columns, clf.feature\_importances\_))

得下图

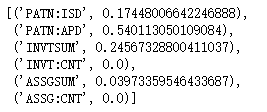


图5 特征指标影响权重

按挖掘模型生成的决策树分析结果

import pydotplusfrom IPython.display import Image import sklearn.tree as treedot\_data = tree.export\_graphviz(clf, out\_file=None, feature\_names=data.columns, class\_names=['0','1'], filled=True) graph = pydotplus.graph\_from\_dot\_data(dot\_data) Image(graph.create\_png())

如图：

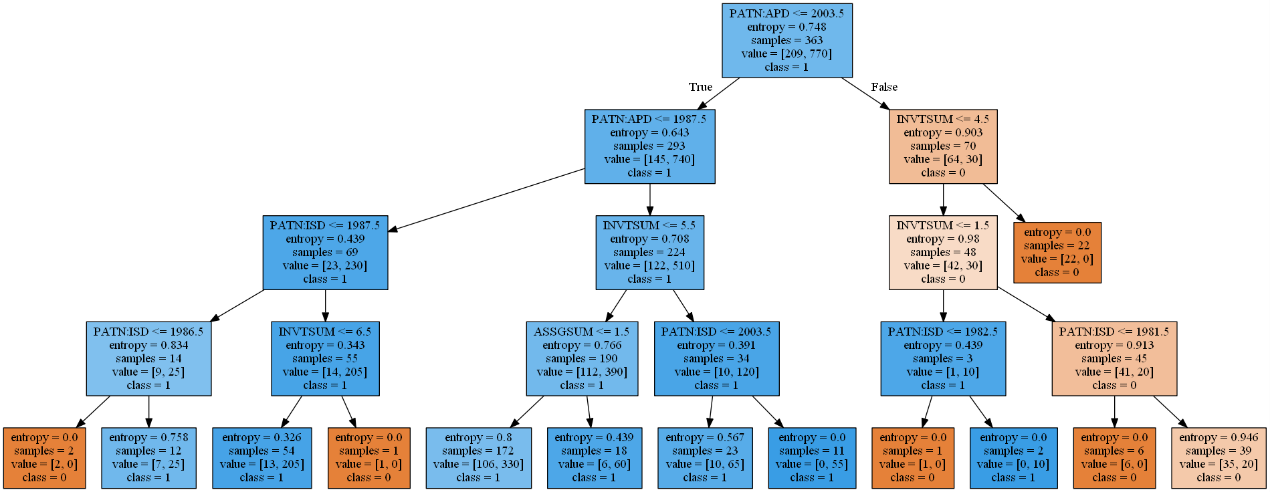


图6 决策树分析结果

根据决策树可分析得出部分规则，如表3

|  |  |
| --- | --- |
| 序号 | 规则内容 |
| 1 | 最早优先权年 ＜ 2003 是被引专利 ( （770/5） /363 ) 置信度 42.4% |
| 2 | 最早优先权年 ＞ 2003 and ＜ 2014 是被引专利( （30/5）/70) 置信度 8.5% |
| 3 | 最早优先权年 ＞ = 1987 and ＜ 2003 = ＞ 是被引专利（(740/5)/293） 置信度 50.5% |
| 4 | 最早优先权年 ＜1987 是被引专利( 848 /603) 置信度 66.67% |
| 5 | 最早优先权年 ＜1987 and 发明人数量＜6为被引专利 ( 41 /55) 置信度 74. 5% |
| 6 | 最早优先权年 ＜ 1987 and 发明人数量 ＞6 为被引专利置信度 76. 73% |
| 7 | 最早优先权年 ＜ 1987 and发明人数量 ＜6时，专利权人数量＜2时为被引专利 (66 /172) 置信度 38. 4% |
| 8 | 最早优先权年 ＜ 1987 and发明人数量 ＜6时，专利权人数量＞=2时，为被引专利（24/34）置信度70.5% |
| 9 | 最早优先权年 ＜ 1987 and，专利公开日＜ 1986是被引专利( 5 /14) 置信度 35.71% |
| 10 | 最早优先权年 ＜ 1987 and，专利公开日＜ 1987and＞ 1986是被引专利( 46 /69) 置信度 66.7% |
| 11 | 发明人数量 ＜ 2时，专利公开日＜ 1983，是被引专利的置信度为66.7% |
| 12 | 发明人数量 ＜ 2时，专利公开日＜ 1981，是被引专利的置信度为8.9% |

表3 专利被引影响因素决策树模型抽取的部分规则

根据以上规则可得出以下结论：

规则 1 ～ 规则 4 表明专利优先权年越早的专利更容易被引用。1987 年之前的专利被引率达到66.7% ， 而2003—2013 年的专利被引率仅仅为 8.5% 。根据专利被引率绘制被引率趋势图，见图 7，可以发现，专利引用率逐步下降，这进一步证明了时间因素对于专利被引行为的 影响作用。

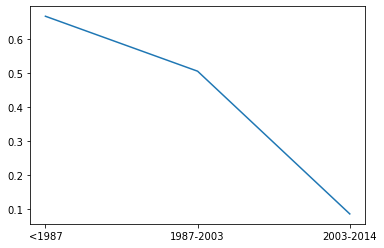


图7 优先权年的影响效果图

规则5与规则6表明，发明人数量对于专利被引有一定的影响，但是影响成度并不大，规则7与规则8表明，专利权人数量的增加，对于专利是否被引也具有一定的影响。

规则 9到规则12表明，专利公开年对于专利是否被引也具有较高的影响力，但不及专利申请年对于专利是否被引的影响程度大。

根据决策树预测模型的依赖关系网络绘制专利被引重要影响因素示意图，见图6

可以发现，能够对预测属性产生影响的属性由强到弱依次是: 专利申请年、专利公开年、专利权人数量、发明人数量。因此在本研究选取的6个影响专利被引的因素中，专利的专利申请年对被引影响最为显著的因素。而发明人国家、专利权人国家2个指标的影响效果并不显著。

发明人数量

专利公开日

专利权人数量

最早优先权年

图 6 专利被引重要影响因素 (重要程度由上往下逐渐变强)