# 摘要

**关键字**

# 引言

但由于近年来经济的快速发展，我国城市化进程加快，依赖化石燃料的交通基础设施也随之扩大。工业化建设和交通工具的普及导致大量污染物排放增加，空气污染日益严重。城市空气污染是我国以及其他发达或发展中国家共同面临的重大问题，大气的污染会对严重影响人类健康，伴随各种疾病的产生，例如：肺癌等。这些空气污染物包括：SO2、NO2、Pm2.5和Pm10等，Pm2.5对人体健康危害最大，因为这些超小颗粒物可以被吸入肺部，也可以溶解在血液中，导致肺癌对高发病率。此外，一些其他严重的环境问题也可能由空气污染引起，如酸雨和温室气体效应。例如，SO2和NO2是酸雨的主要原因，而CO2和N2O是温室气体效应的主要原因。

对于当前空气污染情况，有很多亟待解决的问题：如何及时并准确提供未来大气污染变化信息；如何观察周围环境污染的动态变化；如何针对未来环境保护提供一定的预防措施等等。但在现实生活中，大多数解决方法都是是通过净化空气或者控制污染物排放等，这些措施虽然能一定程度上改善大气环境，但无法从根源彻底解决空气污染的现状。只有在污染未发生之前就采取有效的措施进行防治，才能达到根本解决空气污染问题。因此，采取有效的评估方法，及时对空气质量进行评估及预测就显得十分重要。采取准确高效的评估算法的意义大致可以概括为如下两点：

1. 更加准确地分析出空气质量的变化情况，污染源和气候变化对空气质量的影响。

2. 预测空气污染的发生以及危害，并提前采取措施避免或减少污染。

由此看来，对空气质量的预测对我们保护周围大气环境有着重大意义

# 相关工作

# 本文算法

## 3.1 算法思想

### 3.1.1 C4.5算法

### 在决策树算法中，ID3算法最为著名，ID3算法以信息论为基础，核心思想是用信息增益作为选择属性的度量标准。C4.5算法对其进行了改进，主要克服了ID3算法选择偏向取值较多的属性等不足之处。

C4.5算法主要思想为：

假定S为训练集，S的目标属性C具有m个可能的类标号值，C={C1,C2,...Cm}，训练集S中，Ci在所有样本中出现的频率为（i=1,2,3,...,m），则该训练集S所包含的信息熵定义为：

假设用属性A来划分S，则划分后的样本子集的信息熵为：

则按属性A划分S的信息增益Gain(S,A)为样本集S的熵减去按属性A划分S后的样本子集的熵：

信息增益率使用“分裂信息”值将信息增益规范化。定义如下：

这个值表示通过将训练数据集D划分成对应于属性A测试的v个输出的v个划分产生的信息。信息增益率定义：

选择具有最大增益率的属性作为分裂属性。

### 3.1.2 改进C4.5算法

作为ID3算法的改进算法，C4.5在算法功能上增强了不少，但我们运用C4.5算法对空气质量进行检测时会存在很多不足之处：

1. 一般而言，我们获取的空气质量数据量都较大，并且属性值的取值也较多，根据泰勒展开式定理我们知道对数运算是比较复杂的， 因此在计算信息增益的过程中会，频繁的对数运算会严重影响我们算法的性能。
2. C4.5算法在对连续属性离散化时，要在任一属性的不同取值中插入若干个分割点，再计算所有分割点的信息增益率，选择其中信息增益率最大的分割阈值作为连续属性的最佳分割阈值。当决策树的节点数量比较多、连续属性数量比较多、连续属性中任一属性取值又较多时，算法的计算量少相当大的，这将会在很大程度上影响决策树的生成效率。
3. 采用传统的C4.5算法对空气质量数据进行预测，会忽略数据本身对结果的一些影响，只是简单地通过计算信息增益率来进行建模，未考虑各种污染物对空气影响的程度，因此我们需要进行对各个属性的增益率计算进行加权处理，降低了某些属性的信息熵，相应地提供了其他属性的信息熵。

针对如上三类问题，我们对C4.5算法进行了一定的改进。

1. **泰勒中值定理**：若在包含的某开区间(a,b)内具有n+1阶的导数，则当时，有：

将运算按照如上进行展开，

### 令

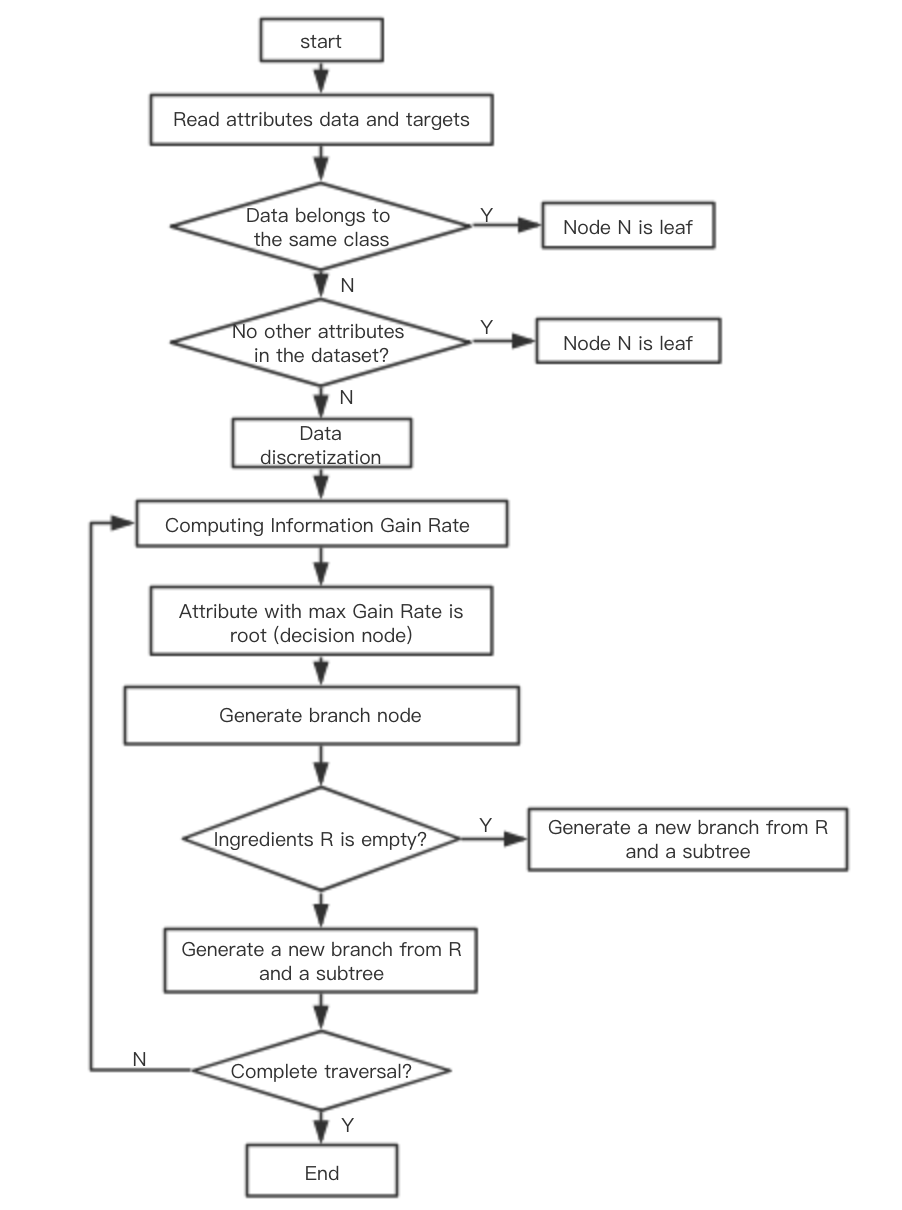
1. C4.5算法在连续属性离散化过程中，要对所有划分情况进行预测，这将占用很多时间，如何快速选择一个最佳的划分阈值已成为亟待解决的问题。Fayyad等证明：无论用于学习的数据集有多少个类别，不管类别的分布如何，连续属性的最佳分割点总在边界点处。

根据Fayyad边界点原理将连续型描述属性按升序排列，选取排序后某一连续型属性点相邻类边界点处的6个属性取值a1,a2,…a6作为测试属性值。其中：ai是类i中的最大值。计算相应的信息增益，选择信息增益最大的属性值作为最佳分割阈值进行划分。改进后的算法只需计算边界点处6个属性值的信息增益率，相当于传统C4.5算法遍历所有的属性值的信息增益率，其计算复杂度大大降低，并且当数据量不断增大时，其计算复杂度也不会增加很多。当遇到特殊情况即每个属性值仅代表1个类别时，改进算法的计算复杂度与C4.5算法相当。

1. 引入一个加权系数W（0<<1），该加权系数的大小由决策者根据先验知识领域知识来确定。通过加权我们可以降低了某些属性的信息熵，相应地提高了其他属性的信息熵。

如果指定某一属性的加权系数为W，则信息增益计算为：

## 3.2算法框架设计



## 3.3 算法步骤

改进后的决策树采用加权信息增益率最大的属性值作为划分节点，生成新的分支，然后通过递归调用该方法建立决策树节点的分支。

本文算法主要归纳为如下几个步骤：

* 读取文本数据，并保存属性实体空气数据以及对应的污染等级。
* 数据清洗。处理离群值以及缺省值。
* 对每个属性下的数据值进行离散化。
* 计算离散化后各属性值的加权信息增益率。
* 选择具有最大加权信息增益率的属性作为子树的根节点
* 若候选属性值不为空，则该根节点生成新的分支节点。
* 递归进行，直到候选节点为空，结束分支过程。

# 4. 实验

**（1）实验数据源**

数据采集作为数据挖掘的准备工作，采集过程中，首先要确定采集数据的作用和目的。例如：数据的总结、数据的分类以及数据的聚类等。确定了数据采集的任务后，就要确定使用哪一种挖掘算法。经过实验可以知道，对于同一挖掘任务可以采用不同的算法来实现。

通过相关网站，我们收集到某市几个月的平均空气质量信息表，如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SO2 | NO | NO2 | NOx | PM10 | PM2-5 | 空气等级 |
| 0.022 | 0.012 | 0.066 | 0.084 | 0.073 | 0.042 | II |
| 0.017 | 0.007 | 0.037 | 0.048 | 0.069 | 0.04 | I |
| 0.016 | 0.003 | 0.018 |  | 0.064 | 0.037 | I |
| 0.023 | 0.003 | 0.019 | 0.023 | 0.064 | 0.036 | I |
| 0.029 | 0.003 | 0.022 | 0.026 | 0.069 | 0.04 | I |
| 0.02 | 0.002 | 0.02 | 0.023 | 0.071 | 0.04 | I |
| 0.015 | 0.002 | 0.021 | 0.023 | 0.066 | 0.039 | I |
| 0.016 |  | 0.019 |  | 0.066 | 0.04 | I |
| 0.025 | 0.001 | 0.029 |  | 0.068 | 0.04 | I |
| 0.023 | 0.001 | 0.056 | 0.058 | 0.084 | 0.047 | II |
| 0.02 | 0.001 | 0.08 | 0.081 | 0.103 | 0.057 | II |
| 0.03 |  | 0.115 | 0.125 | 0.141 | 0.076 | II |
| 0.038 | 0.087 | 0.145 | 0.277 | 0.169 | 0.088 | III |
| 0.047 | 0.097 | 0.132 | 0.279 | 0.167 |  | III |
| 0.041 |  | 0.132 | 0.344 | 0.161 | 0.087 | III |
| 0.04 | 0.088 | 0.124 | 0.258 |  | 0.083 | III |
| 0.056 | 0.119 | 0.124 | 0.306 | 0.158 | 0.088 | III |
| 0.067 | 0.139 | 0.126 | 0.339 | 0.171 | 0.092 | III |

从上表可以看出，我们采集的数据有少部分缺失值，因此在进行数据挖掘之前，需要进行缺失值填充。

**（2）实验参数设置**

**（3）实验结果**

**（4）实验对比分析**

# 5. 结论

# 参考文献