# 第一章 绪论

## 选题的背景和意义

### 研究背景

良好的空气质量是我们人类赖以生存的基础，但由于近年来经济的快速发展，我们居住环境的空气质量也越来越差。空气污染是影响人类健康的最重要因素之一，空气中存在许多有危害的化学物质。这些化学物质有多种来源，一方面，有森林火灾、火山爆发、风蚀、花粉散播、有机化合物蒸发和天然放射性等自然资源。另一方面，人类工业活动代表了人工空气污染源。在许多类型的空气污染物中，氮氧化物、硫氧化物、碳单氧化物、臭氧和能蒸发并进入大气的有机化合物，大量的空气污染物会影响人口健康。为了预防污染，建立了区域、国家和国际大气污染监测网络，实时向人们通报主要污染物浓度。

空气质量预测是当代城市空气质量管理和信息系统的核心内容之一。这些系统通常是为了满足环境立法的需要而建立，专门针对决策者和公民。由于各地区具体特征、空气质量、气候条件以及地形之间关系的复杂性，可以用来精准预测空气质量总是很少的。因此，为了加强空气污染的防治，减少由于空气污染造成对人体健康和环境的危害，研究有效的空气评估方法对空气质量数据进行评估，并进一步制定改善措施减少大气污染、优化空气质量，就显得尤为重要。

### 研究意义

在经济高速发展的同时，空气质量也随之改变，对于当前空气污染情况，有很多亟待解决的问题：如何及时、准确地提供未来大气污染变化的信息；如何观察周围环境污染的动态变化；如何对未来的环境保护提供一定的措施等等。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。

## 1.2国内外研究现状

面对日益恶化的大气环境，国内外关于空气质量的评价方法的研究也日趋活跃。各种评估方法也不断衍生，无论是数据处理还是预测准确性等各方面都取得了显著的成绩。

目前在空气质量评价领域的方法有很多。例如：空气污染指数法、聚类分析法、决策树算法以及神经网络等等。

**空气污染指数**是评估空气质量状况的一空气污染指数组数字，它能告诉您今天或明天您呼吸的空气是清洁的还是受到污染的，以及您应当注意的健康问题。空气污染指数关注的是吸入受到污染的空气以后几小时或几天内人体健康可能受到的影响。空气污染指数划分为0－50、51－100、101－150、151－200、201-300和大于300六档，对应于空气质量的六个级别，指数越大，级别越高，说明污染越严重，对人体健康的影响也越明显

**聚类分析**是一组将研究对象分为相对同质的群组(clusters)的统计分析技术。依据研究对象（样品或指标）的特征，对其进行分类的方法，减少研究对象的数目。各类事物缺乏可靠的历史资料，无法确定共有多少类别，目的是将性质相近事物归入一类。各指标之间具有一定的相关关系。

**决策树**又称为判定树，是运用于分类的一种树结构，其中的每个内部节点代表对某一属性的一次测试，每条边代表一个测试结果，叶节点代表某个类或类的分布。决策树的决策过程需要从决策树的根节点开始，待测数据与决策树中的特征节点进行比较，并按照比较结果选择选择下一比较分支，直到叶子节点作为最终的决策结果。

**神经网络**，它是一种模拟人类大脑思维的训练方式，将空气检测指标等数据传入预先定义好的神经网络中，通过前向传播得到预测结果，若实际输出与期望输出不相符，则进行误差的反向传播，反向调整各神经网络元的权重，逐步降低误差值。这也是最常见的BP神经网络的思想。

## 1.3 研究内容与创新

### 1.3.1 研究内容

本文主要针对ID3、C4.5和CART决策树算法进行研究；讨论了决策树学习算法的基本思想、工作原理和适用范围。并基于决策树学习过程中的常见问题进行了综合分析及归纳：分类过程中的过拟合问题、样本采集的时候处理连续属性问题、采集过程中存在的缺省属性问题以及决策树产生之后的剪枝和优化等问题。

经过对各类问题对研究分析，对比各算法的优劣，找出解决相关问题的方法并提出了改进的决策树算法，同时基于真实数据的基础上，利用matlab进行仿真，并同其他算法进行对比，证明该算法的优先性和可行性。

### 1.3.2 创新点

虽然之前也有许多关于决策树对空气质量预测的研究，但基本只是在算法层面进行了进一步优化，未考虑实际生活环境中，影响空气质量的各因素之间的权重配比，据研究发现影响空气质量的主要因素为Pm2.5，也就是空气中大量的微小颗粒物。因此在本文的研究中引入了相关的平衡系数，其值的设定来源于相关的先验知识。另外由于空气物质含量数据是连续属性值，若数据量很大，我们需要对数据进行必要的离散化操作，提高系统性能。

## 1.4 研究方法与思路

### 1.4.1 研究方法

结合研究内容的特点，本文的研究主要采用以下几种方法：

1. 文献检索法

本文的研究首先需要阅读大量空气质量预测以及决策树算法的相关文献，并总结各文献中对该研究对创新点以及进展情况，找出以前研究对不足和避免内容的重复性。

2. 比较分析法

在论文中将国内外各种决策树算法以及空气质量预测的方法进行分析，对比各算法的优劣并总结，分析还存在什么问题，如何优化，例如：算法复杂度、算法准确度以及实用性等等。

3. 理论联系实际的方法

空气质量监测局采用空气质量指数对空气质量进行度量，监测指标为SO2、NO2、可吸入颗粒物、细粒颗粒物PM2.5、O3和CO等，其中主要影响因素为PM2.5，因此在设计算法模型时需要对PM2.5的数据进行额外处理，从而使得训练的模型更加准确。

1.4.2 研究思路

首先学习决策树算法的基础思想，了解信息熵、信息增益和信息增益率等概念。阅读国内外关于决策树算法研究的相关论文，提炼算法思想，再根据实际情况进行结合，取长补短。然后进行实验，调整参数值进行训练。算法改进之后，需要进行对比实验，将改进后的决策树算法与之前的或者他算法进行对比，观察其效果。最后将该算法应用到其他测试数据集，评估算法的泛化能力，得出结论。

# 空气质量评估方法研究

## 2.1空气污染指数法

空气污染指数，就是根据环境空气质量标准和各项污染物对人体健康、生态、环境的影响，将常规监测的几种空气污染物浓度简化成为单一的概念性指数值形式，它将空气污染程度和空气质量状况分级表示，适合于表示城市的短期空气质量状况和变化趋势。针对单项污染物的还规定了空气质量分指数。参与空气质量评价的主要污染物为细颗粒物、可吸入颗粒物、二氧化硫、二氧化氮、臭氧、一氧化碳等六项。

### 2.1.1空气污染指数的计算方法

API（AirPollution Index的英文缩写）是空气污染指数，我国城市空气质量日报API分级标准如表1：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 空气质量指数API | 污染物质量浓度() | | | | |
| SO2 | NO2 | PM10 | CO | O3 |
| 日均值 | 日均值 | 日均值 | 日均值 | 日均值 |
| 50 | 0.05 | 0.08 | 0.04 | 5 | 0.12 |
| 100 | 0.15 | 0.12 | 0.15 | 10 | 0.2 |
| 200 | 0.80 | 0.28 | 0.35 | 14 | 0.4 |
| 300 | 1.6 | 0.56 | 0.42 | 24 | 0.8 |
| 400 | 2.1 | 0.75 | 0.5 | 36 | 1 |
| 500 | 2.62 | 0.94 | 0.6 | 48 | 1.2 |

表1 空气污染指数对应的污染物浓度限值

设第i种污染物浓度,它的分指数为：

当第i种污染物浓度时，选择点及点来确定一线性函数，其分指数为：

### 2.1.2确定空气污染指数

对每一种污染物计算污染物指数之后，根据如下公式来确定API：

### 2.1.3空气质量评定

参考国家规定的空气污染物指数分级浓度，通过应该空气污染指数法对空气质量进行评价。该方法计算过程相对简单，同时准确率也相对较高。下图为空气质量分级标准：

|  |  |  |
| --- | --- | --- |
| 空气污染指数API | 空气质量级别 | 空气质量描述 |
| 0~50 | Ⅰ | 优 |
| 51~100 | Ⅱ | 良 |
| 101~200 | Ⅲ | 轻度污染 |
| 201~300 | Ⅳ | 中度污染 |
| >300 | Ⅴ | 重度污染 |

表2 空气质量分级标准

## 2.2聚类分析法

### 2.2.1基本思想

我们所研究的样品或指标之间存在程度不同的相似性（亲疏关系——以样品间距离衡量）。于是根据一批样品的多个观测指标，具体找出一些能够度量样品或指标之间相似程度的统计量，以这些统计量为划分类型的依据。把一些相似程度较大的样品聚合为一类，把另外一些彼此之间相似程度较大的样品又聚合为另一类，直到把所有的样品聚合完毕，这就是分类的基本思想。

**2.2.2 K-means聚类分析法**

k-means是划分方法中较经典的聚类算法之一。由于该算法的效率高，所以在对大规模数据进行聚类时被广泛应用。目前，许多算法均围绕着该算法进行扩展和改进。

k-means算法以k为参数，把n个对象分成k个簇，使簇内具有较高的相似度，而簇间的相似度较低。因此我们可以预先根据空气质量划分等级来初步确定K值，将空气污染物含量作为样本集，进行聚类，使得同一类的数据尽可能聚集到一起，不同数据尽量分离。 k-means算法的处理过程如下：首先，随机地选择k个对象，每个对象初始地代表了一个簇的平均值或中心;对剩余的每个对象，根据其与各簇中心的距离，将它赋给最近的簇;然后重新计算每个簇的平均值。这个过程不断重复，直到准则函数收敛。

因此可以看出该算法需要不断进行样本分类调整，不断更新新的聚类中心，当数据量非常大时，算法的时间开销比较大。

## 2.2.3神经网络方法

神经网络是一种生物模型，从数据处理的角度来看，它计算并采用与人脑相同的决策和结论，因而是一种有趣的生物模型。人类在解决诸如识别物体和面孔、识别环境、解决不同情况的问题、行动和反应等过程中表现出色的原因之一是大脑的组织和功能。从出生起，大脑就有能力根据早期的经验制定自己的规则。这些规则是通过加强神经元之间的旧连接和新连接来制定的。人脑中有几十亿个神经元相互连接，并通过电化学信号进行交流。神经网络就像人类大脑的复制品一样，可以解决机器学习领域的复杂问题。神经网络中的学习不是在每个神经元上进行的，尤其是作为一个系统，在全球范围内。所有的神经元一起学习，建立一个网络，可以高精度地解决问题。最常用的神经网络类型之一是反向传播神经网络，其中反向传播用于降低误差。

反向传播本身在计算输入数据结束后开始。每个神经元的激活功能由前一层的输出值和神经元之间的权重决定。首先需要做的是计算输出结果中的误差（预期结果和实际结果之间的区别）。计算误差是正向传播。在前向传播的误差已知后，使用后向传播进行最小化。这意味着算法正在从输出层向输入层反向传播，并且在传播的过程中，它会发现每个权重的错误。必须更改该值以最小化总误差。

# 决策树分类算法研究

## 3.1 决策树简介

决策树作为一种监督学习算法，被用作统计预测、数据挖掘和机器学习的模型。决策树算法有两种类型：分类树（预测结果是包含数据的类）和回归树（预测结果是实数）。分类树和回归树有一个共同的名字CART。分类树是目标变量包含离散值集的树模型。回归树是目标变量得到连续值（通常是实数）的决策树。决策树用于做出决策，并直观或显式地呈现它们，其目的是建立一个模型，该模型将根据先前学习的输入数据预测目标值。

决策树算法被构造成一个流，其中每个内部节点代表一个属性的测试，每个分支代表被测试的特性的值，每个叶子代表一个实例的分类，树的最高节点是根，内部节点（根）为测试数据的某些特征值。每个内部节点都与一个输入变量相匹配。根据输入值，每个叶子都是目标变量的值，显示在从根到叶的路径上。样本数据必须用于构造决策，然后根据树的模型预测输出值，具有最高值的信息位于树的顶部，学习树的过程可以通过根据测试特征值拆分子集中显示的源数据来执行。这种分割应该在每个子组上重复，它被称为递归划分。当节点中的子集具有与目标变量相似的值，并且拆分不增加预测值时，递归完成。

大多数已开发的决策树学习算法是一种核心算法的变体。该算法采用自顶向下的贪婪搜索遍历可能的决策树空间。这种方法也是 ID3 算法和 C4.5 算法的基础。

## 3.2 决策树的主要算法

### 3.2.1 ID3 算法

ID3算法是以信息论为基础，核心思想是选择属性时，用信息增益作为选择属性的度量标准，以使得在每一个非节点进行测试时，能获得关于被测试记录最大的类别信息。

**信息熵：**

用来度量一个属性的信息量。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}，假定训练集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：

（1）

熵越小表示样本对目标属性的分布越纯，反之熵越大表示样本对目标属性分布越混乱。

**信息增益：**

信息增益是划分前样本数据集的不纯程度（熵）和划分后样本数据集的不纯程度（熵）的差值。

假设划分前数据集为S，并用属性A来划分样本集S，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：

（2）

按属性A划分S后的样本子集的熵定义如下：假定属性A有k个不同的取值，从而将S划分为k个样本子集{S1,S2,..,Sk}，则按属性A划分S后的样本子集的信息熵为：

（3）

其中为样本子集中包含该的样本子集，为样本集S中包含的样本数。信息增益越大，说明使用属性A划分后的样本子集越纯，越有利于分类。

ID3算法虽然可以有效处理离散属性，但对于连续型属性却难以处理。连续型属性的处理是数据挖掘领域热门问题之一，影响着数据挖掘算法的准确性、复杂性和可理解性。同时，ID3算法容易产生过度拟合现象。C4.5算法解决了ID3算法产生过度拟合的缺点，并进一步改进，例如未知属性的处理、连续属性的离散化等。

**3.3 本章小结**

### 3.2.2 C4.5 算法

信息增益率使用“分裂信息”值将信息增益规范化。分类信息类似于Info(D)，定义如下：

（4）

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：

（5）

选择具有最大增益率的属性作为分裂属性。

### 3.2.3 CART 算法

CART 决策树使用"基尼指数" (Gini index) 来选 择划分属性.采用与式(4.1) 相同的符号，数据集 的纯度可用基尼值来度量:

Gini(D) = -（6）

直观来说， Gini(D) 反映了从数据集 中随机抽取两个样本，其类别标记 不一致的概率.因此， Gini(D) 越小，则数据集 的纯度越高. 采用与式(4.2) 相同的符号表示，属性 的基尼指数定义为

（7）

于是，我们在候选属性集合 中，选择那个使得划分后基尼指数最小的属 性作为最优划分属性，即

# 第四章 基于改进C4.5的空气质量评价方法

## 4.1基于决策树处理方案设计

C4.5算法的主要核心思想是采用信息增益率作为属性选择标准，但应用到空气质量预测上还存在少许缺点，例如：计算量大，时间复杂度较高以及未考虑样本数据本身的特性。

本文将针对这两点进行改进，提出一种新的决策树算法。

针对空气质量评估的研究，研究入口点是每日采集的空气质量数据信息，所以可以从采集的空气数据作为入口点进行分析研究，我们从大量空气质量数据中提取规则，随后通过这些规则来建立智能评估模型。这也就是数据挖掘方法的最佳应用。

在挖掘过程中可能存在的问题有：

1. 样本数据的采集问题：采集的样本数据是否真实有效;
2. 算法选择不合理问题：如果选用的算法不能处理样本数据，那么数据再好也无济于事。

因此我们需要从数据和算法两方面考虑，提出合适的方案。

## 4.2数据采集

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。另外由于属性值都是连续值，不利于进行决策树构造，因此我们还需要对数据进行必要的离散化操作，减小决策树构建的复杂度。

## 4.3数据预处理

4.3.1 缺失值填充

在某些情况下，可提供使用的数据可能缺少某些属性的值。例如，在医学领域我们希 望根据多项化验指标预测患者的结果，然而可能仅有部分患者具有验血结果。在这种情 况下，经常需要根据此属性值已知的其他实例，来估计这个缺少的属性值。

这里我采用平均值估算的方法处理缺省值，若某个样本数据的某一属性值缺失，我们计算其对应标签的所有数据该属性值的平均值，用平均值来进行估算。具体计算方法如下：

设样本集为S，

## 4.4算法改进

4.4.1 数据离散优化

作为ID3算法的改进算法，C4.5在算法功能上增强了不少，但也存在一些不足之处：C4.5算法采用信息增益率作为最佳分裂属性的评判标准，计算描述属性的信息增益，并选择其中信息增益率最大的描述属性作为分裂属性。在连续属性离散化时，由于C4.5算法连续属性离散化方法可知，算法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。因此我们将针对连续属性最佳分割阈值选择算法的计算复杂度问题提出了一些改进，以提高决策树的效率。

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。

### 4.4.2 熵值计算优化

C4.5 算法由于大量使用了对数函数进行熵值运算，增加了计算机的运算时间，降低了 每一次属性选择时算法的运算效率，所以为了解决这个问题，引入泰勒中值定理和麦克劳林 展开式，对熵值中的对数运算进行变换，优化 熵值运算，缩短其运算时间。

**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

### 其中（）

1. 当n=0时，泰勒公式给出拉格朗日中值定理

1. 当n=1时，泰勒公式变为

可见

了解完泰勒中值定理后，我们需要将log2(m)运算按照如上进行展开，

### 令

不同的展开点，对于特定的预测任务所需要的计算复杂度不同。当需要估计的点位于收敛域内，且距离展开点较近时，所需要展开的级数就比较小。当需要估计的点位于收敛域外，泰勒级数将无法进行预测。当需要顾及的点位于收敛域内且距离展开点较远时，所需要的计算量就比较大。因此，我们通过泰勒中值定理对其进行展开并估计，这样可以大大减小计算复杂度，提高算法的性能。

### 4.4.3 特征选择优化

引入一个平衡度系数λ（0<λ<1），该平衡系数是一个模糊的概念，其大小由决策者根据先验知识领域知识来确定。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。

如果指定某一属性的平衡系数为λ，引入平衡系数后

选择最高修正信息增益的属性为当前节点的测试属性，该属性的分枝对应于未合并样本子集和复合样本子集。

## 4.5 本章小结

本章提出了基于决策树处理的方案，分析并介绍了数据的采集和预处理问题，同时在C4.5算法的基础上进行算法改进，使得该算法更适合应用于空气质量评估数据中。

# 第五章. 实验过程与结论分析

## 5.1 算法仿真

## 5.2 模型评价

## 5.3 实验结果

## 5.4 实验分析

## 5.5 本章小结

# 第六章. 总结与展望

## 6.1 本文工作总结

## 6.2 下一步工作