华录杯政府开放数据创意赛复赛

天津市空气质量指数 (AQI) 预测

团队：VincentZhang0cb49
队长: 张贺

俄勒冈州立大学 百度美国研究院

October 15, 2018

Contents

1 引言

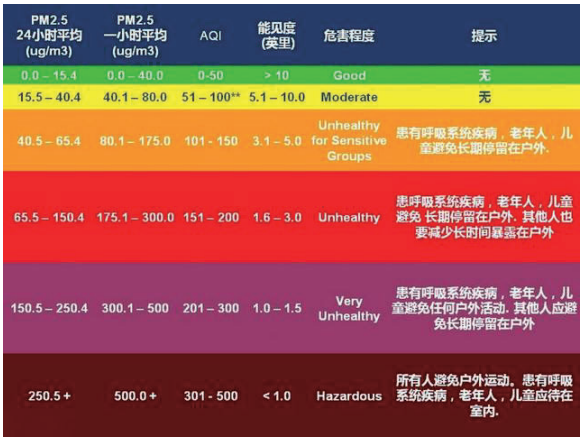
空气污染是当前世界最主要的环境问题之一，对人类健康、工农业生产、动植物生长和全球环境等都会造成很大的伤害。空气污染物侵入人体主要有三条途径：表面接触、摄入含污染物的食物和水、吸入被污染的空气，其中以第三条途径最为重要。大气污染对人体健康的危害主要表现为引起呼吸道疾病。在高浓度污染物的突然作用下，人体可发生急性中毒，甚至在短时间内死亡。长期接触低浓度污染物，会引起支气管炎、支气管哮喘、肺气肿和肺癌等病症。

为了对存在于大气、空气中的污染物质进行定点、连续或者定时的采样、测量和分析，国内大中城市均广泛建立了空气质量监测站。站内安装多参数自动监测仪器作连续自动监测，将监测结果实时存储并加以分析后得到相关的数据。其中待监测因子包括：污染极细颗粒物（PM2.5，PM10），臭氧，二氧化硫，一氧化碳，硫化氢，氮氧化物，挥发性有机污染物，总悬浮颗粒物，铅，苯，气象参数，能见度等。综合空气质量监测站提供的数据，当前环境的空气质量可以通过空气质量指数（AQI）来衡量。空气质量监测站是空气质量控制和对空气质量进行合理评估的基础平台，是一个城市空气环境保护的基础设施。

监测空气质量非常重要，但是同时，预测未来一段时间的空气质量也已经成为国家发展和人民生活的一个亟待解决的问题。比如，为了确保未来几天大型活动的顺利进行，政府可能会采取交通管制、节能减排等一系列行政措施保证天气质量。但这也不可避免的会对生产生活产生理想。如果我们能够预测到未来几天空气质量会随着天气等因素的变化而自动好转，这就可以少进行或者不进行交通管制等行政措施，从而减少其对生产生活造成的影响。同样，对于市民来说，很多情况下希望预知未来的空气情况来安排自己的生活，尤其像出行、锻炼、郊游这些活动，更是需要安排在良好的空气质量的天气条件下。因此虽然空气质量检测站虽然能够提供详实的空气监测数据，却难以准确的预测未来一段时间的空气质量。这主要是因为影响空气质量的因素众多。单纯的历史过往空气数据，很难为准确的预计提供支持，还要考虑其他一些方面的因素。比如，大风能够迅速吹走空气中的污染物质，冬天华北平原经常会出现数日空气污染严重，大家“等

风来”的情况。降雨、降雪也能够明显过滤掉空气中的可吸入颗粒物等污染物质。同时，不同的季节空气质量通常呈现周期性的变化。如北京、天津等华北城市秋天大多秋高气爽，空气质量良好；而冬天总会有较多时间雾霾严重，空气质量差。除了天气、气候等因素外，地理环境也对空气质量起着至关重要的作用。盆地往往污染物难以扩散；而沿海地区因为海洋的净化作用，空气质量大多优于内陆地区。只有将这些因素综合考虑，才能较为准确的对空气质量进行预测。

建立的空气质量预测系统对社会有重大意义，可以防止危害事件发生，让有关部门有时间进行预防措施。我们的模型对健全这一预测体系进行了有益的尝试，并取得了不错的效果。针对空气质量难预测、需预测这一问题，我们团队提出了一套综合考虑近期天气、气候特征、季节影响和地域因素的空气质量预测方案。该方案利用提供的天津市 2017-2018 年气象数据和空气质量监测数据，并综合在外部获取的地理位置、气候特征等信息，通过数据挖掘和机器学习的方法，建立了天津市各区域空气质量预测模型。改模型预测性能优异，能够达到 85% 以上的预测准确率。并且该模型易于训练，有很好的适用性，能够推广到多个城市和地区。



PM2.5 24小时平均 (ug/m3)	PM2.5 一小时平均 (ug/m3)	AQI	能见度 (英里)	危害程度	提示
0.0 – 15.4	0.0 – 40.0	0-50	> 10	Good	无
15.5 – 40.4	40.1 – 80.0	51 – 100**	5.1 – 10.0	Moderate	无
40.5 – 65.4	80.1 – 175.0	101 - 150	3.1 – 5.0	Unhealthy for Sensitive Groups	患有呼吸系统疾病，老年人，儿童避免长期停留在户外.
65.5 – 150.4	175.1 – 300.0	151 – 200	1.6 – 3.0	Unhealthy	患呼吸系统疾病，老年人，儿童避免 长期停留在户外. 其他人也要减少长时间暴露在户外
150.5 – 250.4	300.1 – 500	201 – 300	1.0 – 1.5	Very Unhealthy	患有呼吸系统疾病，老年人，儿童避免任何户外活动. 其他人应避免长期停留在户外
250.5 +	500.0 +	301 - 500	< 1.0	Hazardous	所有人避免户外运动。患有呼吸系统疾病，老年人，儿童应待在室内.

Figure 1: 空气质量等级

2 空气质量评价体系与影响因素

2.1 空气质量评价体系

国际上衡量空气质量优劣通常使用空气质量指数，即 AQI 评价体系。这一体系第一步是对照各项污染物的分级浓度限值，以细颗粒物（PM2.5）、可吸入颗粒物（PM10）、二氧化硫（SO2）、二氧化氮（NO2）、臭氧（O3）、一氧化碳（CO）等各项污染物的实测浓度值（其中 PM2.5、PM10 为 24 小时平均浓度）分别计算得出空气质量分指数（Individual Air Quality Index，简称 IAQI）

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_P - BP_{Lo}) + IAQI_{Lo} \tag{1}$$

其中：

- $IAQI_P$——污染物项目 P 的空气质量分指数；
- C_P——污染物项目 P 的质量浓度值；
- BP_{Hi}——相应地区的空气质量分指数及对应的污染物项目浓度指数表中与 C_P 相近的污染物浓度限值的高位值；
- BP_{Lo}——相应地区的空气质量分指数及对应的污染物项目浓度指数表中与 C_P 相近的污染物浓度限值的低位值；

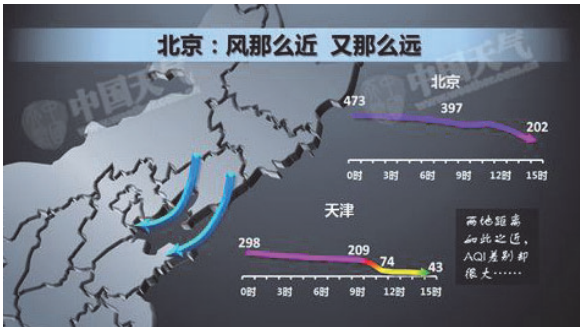


Figure 2: 气象条件对空气质量的重要影响

- $IAQI_{Hi}$——相应地区的空气质量分指数及对应的污染物项目浓度指数表中与 BP_{Hi} 对应的空气质量分指数；
- $IAQI_{Lo}$——相应地区的空气质量分指数及对应的污染物项目浓度指数表中与 BP_{Lo} 对应的空气质量分指数。

第二步是从各项污染物的 IAQI 中选择最大值确定为 AQI，当 AQI 大于 50 时将 IAQI 最大的污染物确定为首要污染物:

$$AQI = \max\{IAQI_1, IAQI_2, IAQI_3, ..., IAQI_n\} \tag{2}$$

其中：

- $IAQI_i$——空气质量分指数；
- n——污染物项目。

第三步是对照 AQI 分级标准，确定空气质量级别、类别及表示颜色、健康影响与建议采取的措施。图 1 是较为常用的 AQI 分级标准。

2.2 空气质量影响因素

国内外学者对空气质量影响因素进行了深入的研究。其中一个主要观点和研究方向是气象等因素对空气质量的影响，即从气象角度探讨空气污染物的形成和不易扩散的原因。国内的学者如王郁，侯青等通过研究背景空气持续污染特征发现沙尘暴天气是影响空气质量的主要原因。张记刚，

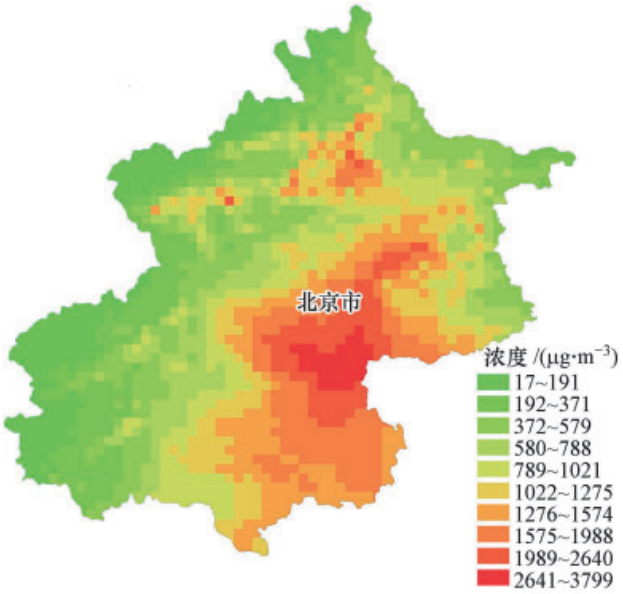


Figure 3: 地理地形对空气质量的重要影响

周涛，郭淼淼等认为天气因素，如温度，湿度，风速，日照时间均对大气颗粒物的浓度产生影响。李小飞，张明军等学者通过研究中国空气污染指数变化特，指出了诸多气象因素。包括降水量、风速风向、逆温、地面气压、地面温度、相对湿度、云量等对空气质量的作用，以及地形因素所涉及的影响。虽然这些学者研究方法和数据不尽相同，但他们均认为 AQI 与气象条件有极大关系，不同气象条件下污染物产生和扩散的条件不同，排入相同数量的污染物，空气中的污染物浓度也会有不同。比如静稳天气条件下风力微弱，容易出现逆温层，不利于颗粒物的扩散，重污染天气易发，而对于风力大，对流强的地区和时段，大气扩散稀释能力强，此时空气质量相对污染物排放量不会像静稳天气下那么敏感。所以，对空气质量进行预测需要结合气象条件。图 2 描述了空气质量与天气的密切联系。

同时，地理地形和区域功能也对空气质量产生不可忽视的作用。图 3 显示了北京不同区域空气污染物浓度随地形和功能区域的变化。我们可以看到在西部北部山区，空气污染物少，而在东部南部的平原人口稠密区空气质量较差。

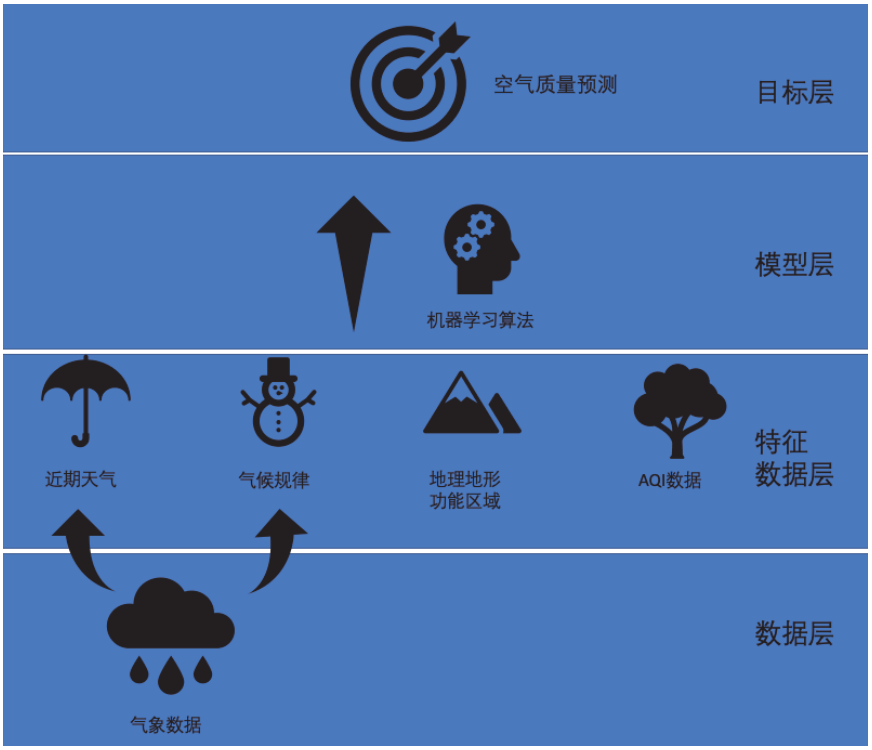


Figure 4: 预测模型

3 预测模型

对于空气污染物的预测问题，国内外大多采用传统的数值预报。数值预报模型的准确率很大程度依赖于污染源的排放数据，该类数据获取的复杂性和不确定性使得数值预报在实际应用中受到限制。针对数值预报模型的弊端，我们采用了机器学习中的多种分类和回归模型，通过分析建立影响因子与污染物浓度之间复杂的线性或非线性关系来建立模型。利用机器学习，可以对大量的空气质量和气象历史数据进行智能分析和归纳总结，通过解读复杂非结构性数据，挖掘出空气质量指数与各污染物因子以及温度、湿度、风速等气象条件之间的内在关系，并建立起 AQI 与各影响因子之间的复杂计算模型，从而训练一个有效的学习模型来对空气质量进行预测。并且该模型可以通过改变训练样本，方便快捷的移植到其他城市的空气质量预测项目中，而不用针对特定的城市、气象、空气质量条件建立特定的模型。

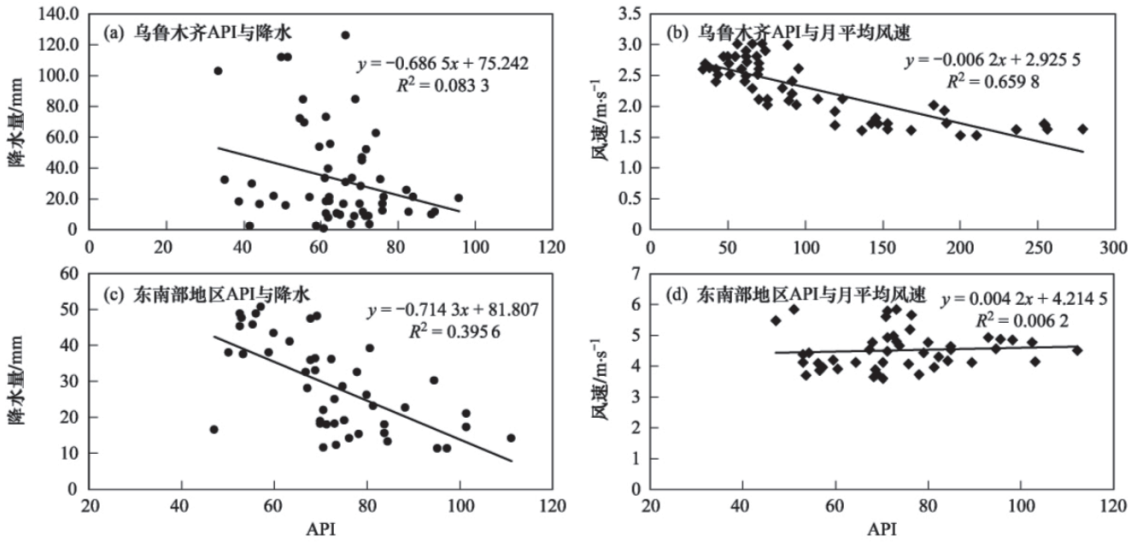


Figure 5: 降水和风速与空气质量的相关关系

3.1 数据收集

为了使模型具有较高的准确性，我们收集和整理了多项与空气质量相关的数据。这些数据包括:

- 2017 年 4 月 1 日至 2018 年 3 月 31 日天津市各空气质量监测站发布的空气质量数据。其中有 18 个不同区域的监测站数据，数据频率为每小时记录一次。当次数据包括测量日期、测量时刻、监测站地理位置、空气质量指数、空气质量等级、主要污染物、PM2.5 浓度，PM10 浓度，CO 浓度，NO2 浓度，一小时臭氧浓度，八小时臭氧浓度，SO2 浓度等。其中空气质量指数，PM2.5 浓度，PM10 浓度，CO 浓度，NO2 浓度，一小时臭氧浓度，八小时臭氧浓度，SO2 浓度为数值型数据，其余为描述型数据。

- 2017 年 4 月 1 日至 2018 年 3 月 31 日天津市各气象站发布的气象数据。这类数据分为两种，一种为市内六区日数据，数据内容包括：站号，站名，监测日期，日平均气压，日平均气温，日降水量，日平均 10 分钟风速，日最多风向；另一种为市内六区小时数据，数据内容包括：站号，站名，监测日期，观测时刻，气压，气温，小时降水量，2 分钟

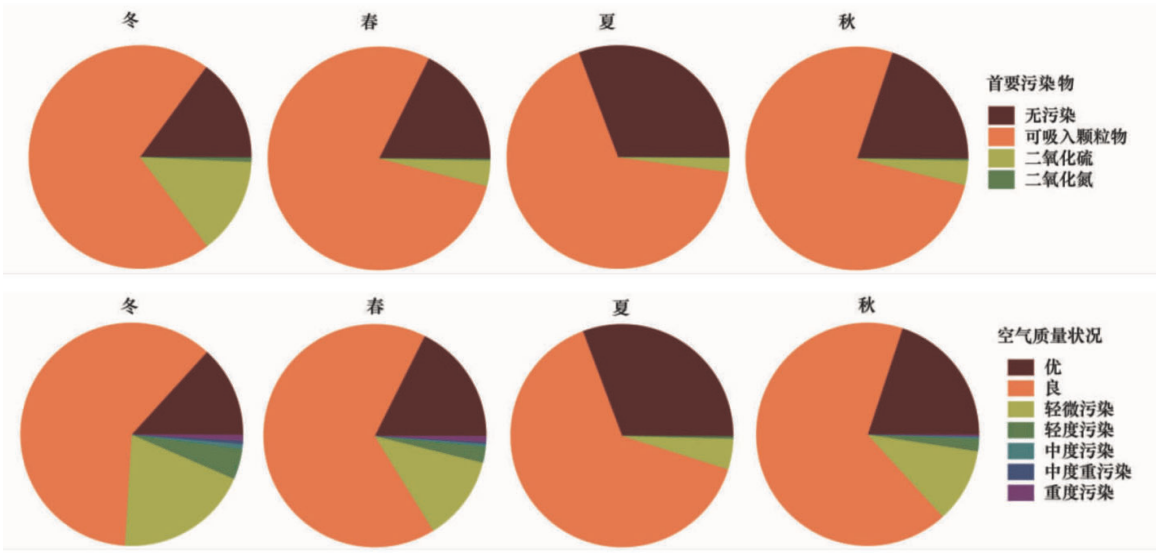


Figure 6: 大气污染随季节变化

风向，2 分钟平均风速，2 分钟风向，2 分钟平均风速。这套数据能够与空气质量数据在时间上吻合起来，进而挖掘天气情况对空气质量的影响。同时两种不同精度的数据我们都加以利用。日数据用于 24 小时粗颗粒度的空气质量预测，小时数据用于下一小时的精细空气质量预测。图 4 给出了相关研究中气象要素降水和风速与空气质量的相关性。

- 天津市地理地形数据。
- 2010 年 1 月 1 日至 2018 年 3 月 31 日天津市天气数据。这一数据用来挖掘不同季节的气候特征。图 4 给出了相关研究中大气污染随季节变化的规律。

3.2 特征数据提取

我们将空气质量预测分为空气质量等级预测和空气质量指数（AQI）预测。前者是一个分类问题，而后者是一个回归问题，但是两者可以共享特征数据。

气象规律特征从 2010-2018 年天津气象数据中按照季节（春季 3-5 月，夏季 6-8 月，秋季 9-11 月，冬季 12-2 月）提取九年中该季节气候特征分量加和值。例如，将所有 9 年中春季降水量数据加和，得到的数据即位春季降水特征值。

针对 24 小时（长时段）空气质量预测，其时间尺度长，预测颗粒度低，因此选取预测当天之前一周的天气数据作为近期天气原始数据，并从中提取出降水，风向风力，气压等特征值；选取预测当天之前一周的空气质量数据作为近期空气质量原始数据，并从中提取出相应污染物浓度作为特征值，提取 AQI 值和 AQI 等级作为回归（分类）值。针对 1 小时（段时段）精准空气质量预测，选取预测时刻之前 24 小时的天气数据作为近期天气原始数据，并从中提取出降水，风向风力，气压等特征值；选取预测时刻之前 24 小时的空气质量数据作为近期空气质量原始数据，并从中提取出相应污染物浓度作为特征值，提取 AQI 值和 AQI 等级作为回归（分类）值。

我们将 18 个空气质量监测站点定位于天津地形图上，并将它们按照功能区划（如商业区、科教区、工业区、郊区等）和地形（平原、丘陵、水网等）两个维度分类。这两个维度一个反应了该区域的人流和经济活动情况（这一情况通常对空气污染物的生成有重要影响），另一个反应了该区域的地理地形情况（这一情况通常对空气污染物的扩散有重要影响）。标注出 18 个空气质量监测站的地形分类和功能区域分类，并以此分类作为地理特征值。

我们注意到，这次比赛提供的天气数据和 AQI 数据并非完善，例如在小时天气数据就会缺少某些时刻的值。因此我们假设天气条件在小时尺度上是连续变化的，并用缺省数据周边的数据（如当前数据 +1 或-1 小时）作为假象数据来填充空缺。

3.3 模型选择和实现

针对空气质量等级预测，我们选取了多个常用的分类模型，包括梯度提升（Gradient Boosting）分类器, 随机森林（Random Forest）分类器，提升（adaboost）分类器，决策树（Decision Tree）分类器，最近邻（KNN）分类器，多层感知（Multilayer Perceptron）分类器以及随机梯度下降（SGD）分类器。

梯度提升（Gradient Boosting）分类器可以将一系列弱学习因子（weak learners）相结合来提升总体模型的预测准确度。在任意时间 t，根据 t-1 时刻得到的结果我们给当前结果赋予一个权重。之前正确预测的结果获得较小权重，错误分类的结果得到较大权重。其算法可以归纳为：

- 初始分类目标的参数值
- 对所有的分类树进行迭代：
  - 根据前一轮分类树的结果更新分类目标的权重值（被错误分类的有更高的权重）
  - 用训练的子样本建模
  - 用所得模型对所有的样本进行预测
  - 再次根据分类结果更新权重值
- 返回最终结果

决策树算法容易理解与解释，并且是非参数的，所以不需要担心离群点和数据是否线性可分的问题但决策树算法容易过拟合，这也正是随机森林等集成学习算法优于决策树算法的地方。随机森林算法因为有许多的决策树组成，可以很好的处理连续型的特征，且速度快可扩展，也不像支持向量机算法那样需要调整大量的参数。

KNN 算法首先贮藏所有的训练样本，然后通过分析（包括选举，计算加权和等方式）一个新样本周围 K 个最近邻，然后把新样本标记为在 K 近邻点中频率最高的类。这种方法有时候被称作“基于样本的学习”，即为了预测，我们对于给定的输入搜索最近的已知其相应的特征向量。KNN 算法简单有效，但因为需要存储所有的训练集，占用很大内存，速度比较慢。

多层感知器（MLP）是一种前向结构的人工神经网络，映射一组输入向量到一组输出向量。除了输入输出层，它中间可以有多个隐层，最简单的 MLP 只含一个隐层，即三层的结构。MLP 可以被看作是一个有向图，由多个的节点层所组成，每一层都全连接到下一层。除了输入节点，每个节点都是一个带有非线性激活函数的神经元（或称处理单元）。一种被称为反向传播算法的监督学习方法常被用来训练 MLP。MLP 是感知器的推广，克

服了感知器不能对线性不可分数据进行识别的弱点。MLP 用公式总结起来就是:

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x))) \tag{3}$$

其中 G 是 softmax 函数。

AdaBoost 方法对于噪声数据和异常数据很敏感。但在一些问题中,AdaBoost 方法相对于大多数其它学习算法而言,不会很容易出现过拟合现象。AdaBoost 方法中使用的分类器可能很弱（比如出现很大错误率），但只要它的分类效果比随机好一点（比如两类问题分类错误率略小于 0.5），就能够改善最终得到的模型。而错误率高于随机分类器的弱分类器也是有用的,因为在最终得到的多个分类器的线性组合中，可以给它们赋予负系数，同样也能提升分类效果。AdaBoost 训练中，在每一轮中加入一个新的弱分类器，直到达到某个预定的足够小的错误率。每一个训练样本都被赋予一个权重，表明它被某个分类器选入训练集的概率。如果某个样本点已经被准确地分类，那么在构造下一个训练集中，它被选中的概率就被降低；相反,如果某个样本点没有被准确地分类，那么它的权重就得到提高。通过这样的方式，AdaBoost 方法能“聚焦于”那些较难分（更富信息）的样本上。

SGDClassifier 是一个用随机梯度下降算法训练的线性分类器的集合。默认情况下是一个线性（软间隔）支持向量机分类器。其在很多问题上都有优异的表现。

针对空气质量指数（AQI）值预测，我们选取了多个常用的回归模型,包括线性回归, 随机森林回归模型，提升（adaboost）方法回归模型，决策树回归模型，最近邻回归模型，多层感知回归模型以及随机梯度下降回归模型。

这些模型各有各自的特点，直观上很难判别哪个模型更适合空气质量预测这一问题，因此我们对这些模型都进行了尝试，并从中选出最优的模型。

因为数据集并没有天然分为训练集合和测试集合，因此我们采取了 K-fold cross-validation 验证法。这一方法将训练集分割成 K 个子样本，一个单独的子样本被保留作为验证模型的数据，其他 K-1 个样本用来训练。交叉验证重复 K 次，每个子样本验证一次，平均 K 次的结果或者使用其它结合方式，最终得到一个单一估测。这个方法的优势在于，同时重复运用随

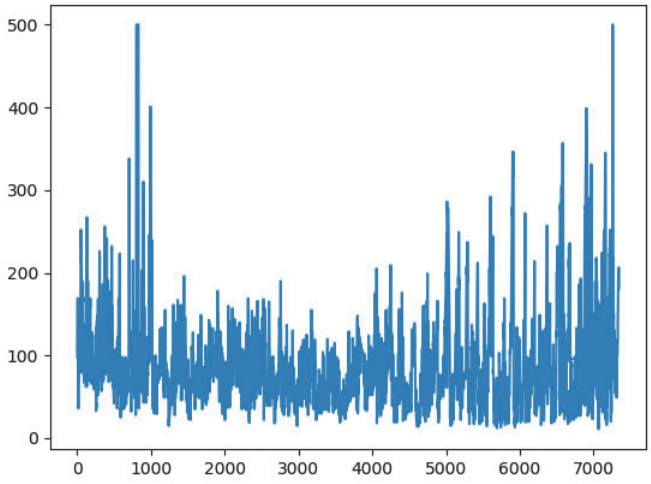


Figure 7: 中山北路站点全年空气质量指数变化

机产生的子样本进行训练和验证，每次的结果验证一次，10 次交叉验证是最常用的。其公式为：

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \tag{4}$$

我们利用 python 中机器学习包 sklearn 实现了上述模型。具体代码请参见 train.py。

3.4 预测精度

我们选取了天津市中山北路空气质量监测站的数据用于验证模型精度。图 7 展示了中山北路站点全年 AQI 变化情况，我们可以发现其全年 AQI 变化波动大，规律性弱，要想精确预测难度很大。

图 8 展示了 8 种空气质量等级预测模型的精度。其中 GB 模型得到了超过 85% 的精度，表现最为优异。除此之外，RF 模型、Adaboost 模型也得到了高于 80% 的预测精度。而通常客气质量预测精度在 70% 以下。我们的模型在预测精度上超过了当前绝大多数模型。

我们也注意到高斯朴素贝叶斯模型（GausianNB）精度明显低于其他模型。这可能是因为朴素贝叶斯模型假设属性之间相互独立，这个假设在实

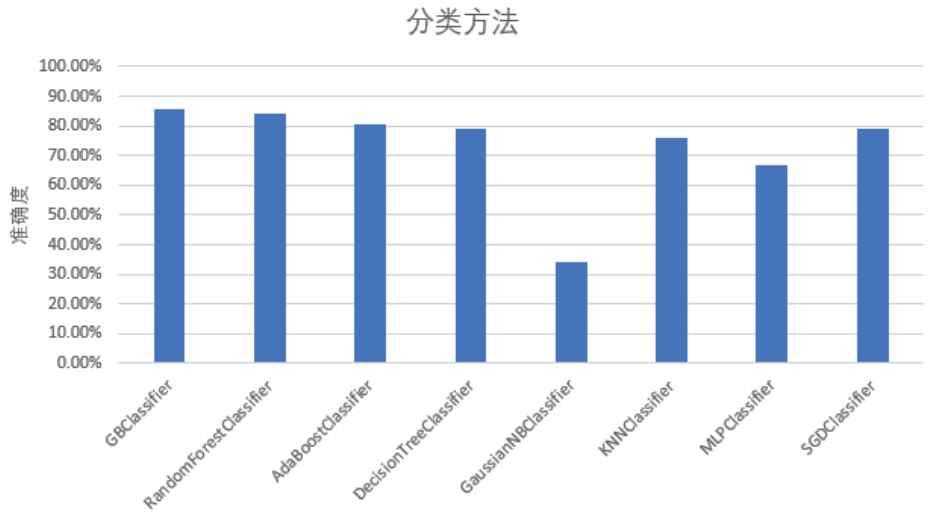


Figure 8: 空气质量等级预测结果

际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好。而在属性相关性较小时，朴素贝叶斯性能最为良好。而我们的输入特征向量中具有明显的相关性，不能看成为独立分量。

图 9 展示了 6 种空气质量指数预测模型的精度。其中 RF 模型 MSE 最低（315），表现最为优异。除此之外，线性回归模型也有不错的表现。

4 总结与展望

准确的预测空气质量对于社会有很积极的意义。当前的预测方法主要是基于检测到的各项污染物浓度，往往没有综合考虑多种影响因素，比如大风和雨雪天气会对空气中的污染物扩散有利，有些地势较低的地区则污染物不容易扩散。另一方面，当前的空气质量预测服务大多只能粗略预测未来一天的空气质量，并不能精确到小时。为了解决上述问题，提供更加精确的空气质量预测，在此项目中，我们通过利用 2017 年 4 月 1 日至 2018 年 3 月 31 日的天津市空气质量数据和气象数据来训练模型，综合考虑了各项空气污染物浓度和气象数据，使其能够精确预测未来一小时的空气质量等级和空气质量指数。同时，只需要将训练数据替换成当地的气象数据，我们的模型就可以很容易的推广到不同的地区，提供快速而准确的空气质

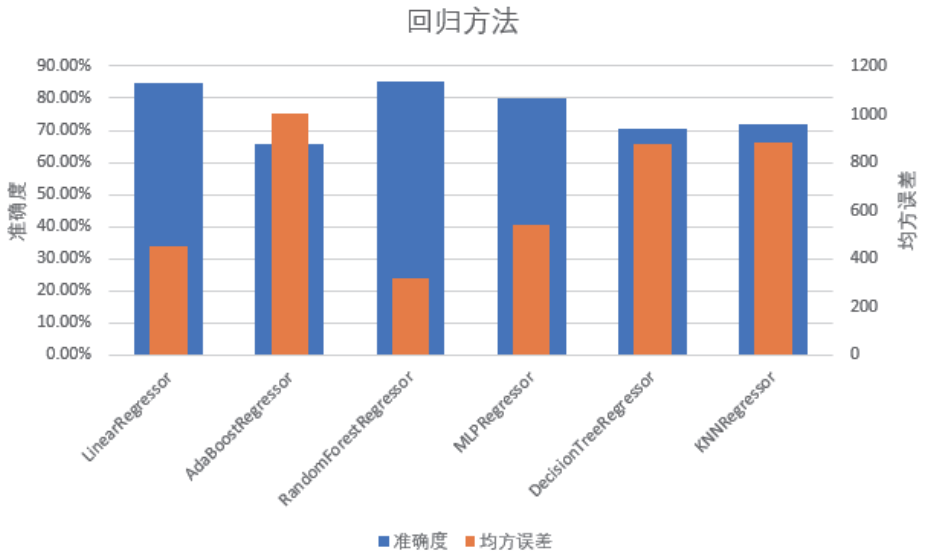


Figure 9: 空气质量指数预测结果

量预测。此外，如果有需求，该模型可以不限于只预测未来一小时的空气质量，而是扩展为以小时为单位的未来 24 小时空气质量预测，这比起当前的预测数据在精度上会有很大提高。

同时下一步我们计划探索使用深度学习中的 Seq2Seq 模型来建立新的模型。Seq2Seq 模型是一种机器学习模型，它使用解码器和编码器从数据中学习序列化的特征模式。Seq2Seq 模型适用于许多机器学习应用场景中，尤其是机器翻译等自然语言处理相关领域。由于 Seq2Seq 模型非常适合描述具有时序特点的输入输出关系，因此能够很好地描述随时间变化的空气质量指数和天气条件。