湖南城市学院

实验名称	实验四　空气质量分类预测
姓　　名	李灵慧
学　　号	2202501-18
学　　院	市政与测绘工程学院
专　　业	地理空间信息工程
指导教师	汤淼

2025 年　4　月　10　日

1 数据预处理

1.1 数据问题

（1）24 小时平均二氧化碳存在负值

```
发现24hCOavg的无效值：
         Ctnb  Ctn Prvn       Date  ...  24hSO2avg 24hNO2avg  24hCOavg  24hO3avg
73779  210200  大连市  辽宁省 2013-11-07  ...         21        23     -0.14        31
88804  210800  营口市  辽宁省 2013-11-07  ...         24        27     -0.05        48
```

图 1.1.1 负值

（2）数据各项类型未确定

1.2 应对方法

（1）修改负值为 0

（2）定义各项数据类型，存为 parquet 格式，提高读写效率

2 分类预测方法

2.1 SVM 支持向量机

（1）介绍

支持向量机（Support Vector Machine，简称 SVM）的核心思想是通过在特征空间中找到一个最优超平面，将不同类别的数据分开。SVM 通过最大化间隔（即超平面到最近数据点的距离）来提高模型的泛化能力。

（2）部分参数

1. C
   - 含义：惩罚参数（Penalty parameter），用于控制分类错误的惩罚力度。
   - 作用：较大的 C 值会使模型更严格地惩罚分类错误，可能导致过拟合；较小的 C 值会使模型更宽容错误，可能导致欠拟合。
   - 默认值：1.0
2. kernel
   - 含义：核函数类型，用于将数据映射到高维空间以实现线性可分。
   - 可选值：
     - "linear"：线性核函数，适用于线性可分的数据。
     - "poly"：多项式核函数，适用于小规模非线性数据。
     - "rbf"（默认）：径向基函数（高斯核），适用于大规模非线性数据。
     - "sigmoid"：Sigmoid 核函数，较少使用。
     - "precomputed"：预计算核矩阵，用于自定义核函数。
   - 默认值："rbf"
3. degree
   - 含义：多项式核函数的度数（仅在 kernel="poly"时有效）。
   - 作用：控制多项式核函数的复杂度。

- 默认值：3

4. gamma

- 含义：核函数的系数（仅在 kernel="rbf"、kernel="poly"和 kernel="sigmoid"时有效）。
- 可选值：
  - "scale"（默认）：gamma = 1 / (n_features * X.var())，其中 X.var()是数据的方差。
  - "auto"：gamma = 1 / n_features。
  - 浮点数：直接指定 gamma 值。
- 作用：较大的 gamma 值会使模型更关注靠近决策边界的点，可能导致过拟合；较小的 gamma 值会使模型更平滑，可能导致欠拟合。
- 默认值："scale"

5. coef0

- 含义：核函数中的独立项（仅在 kernel="poly"和 kernel="sigmoid"时有效）。
- 作用：调整核函数的偏移量。
- 默认值：0.0

（3）各参数影响

1、Kernel

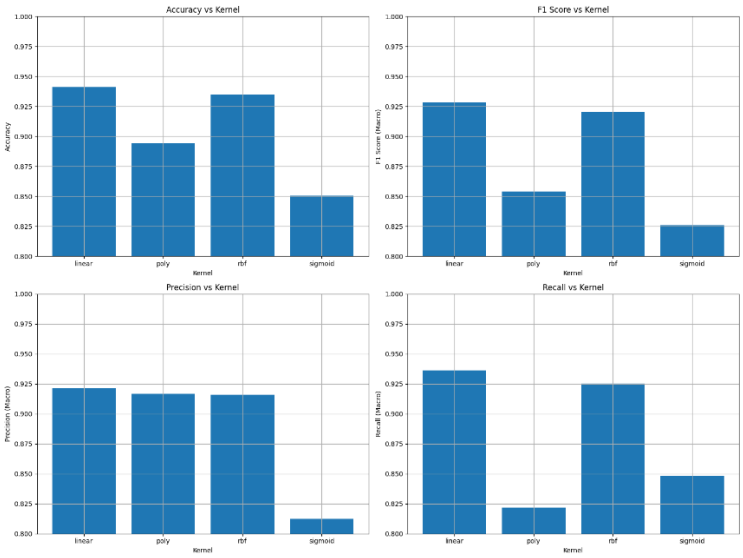


图 2.1.3.1 核函数对分类精度影响

2、C

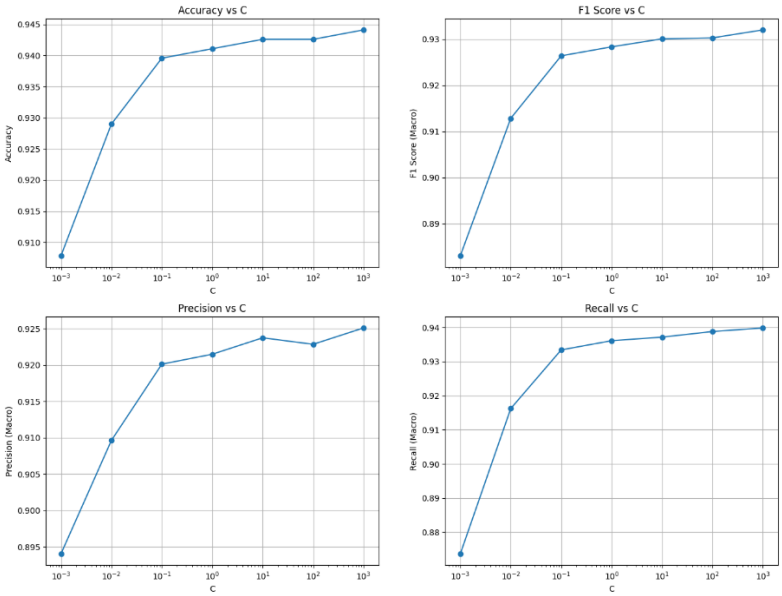


图 2.1.3.2.1 核函数为 linear 时

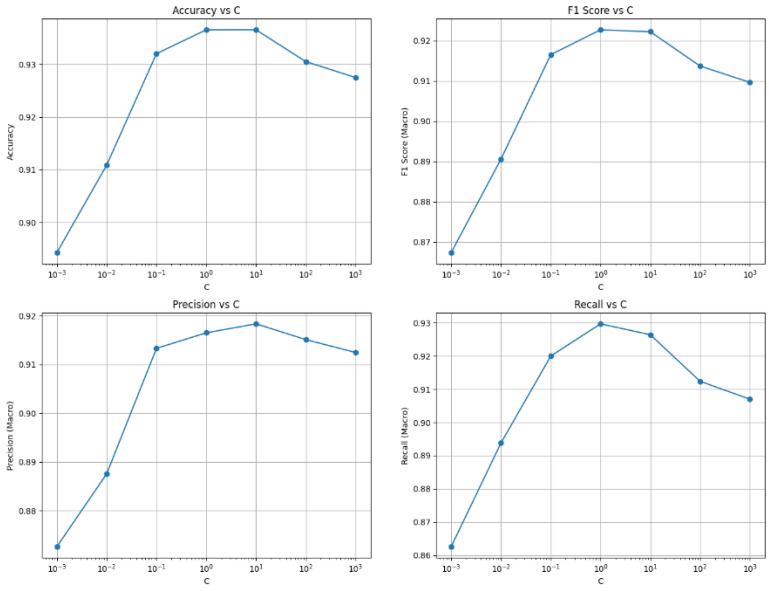


图 2.1.3.2.2 核函数为 rbf 时

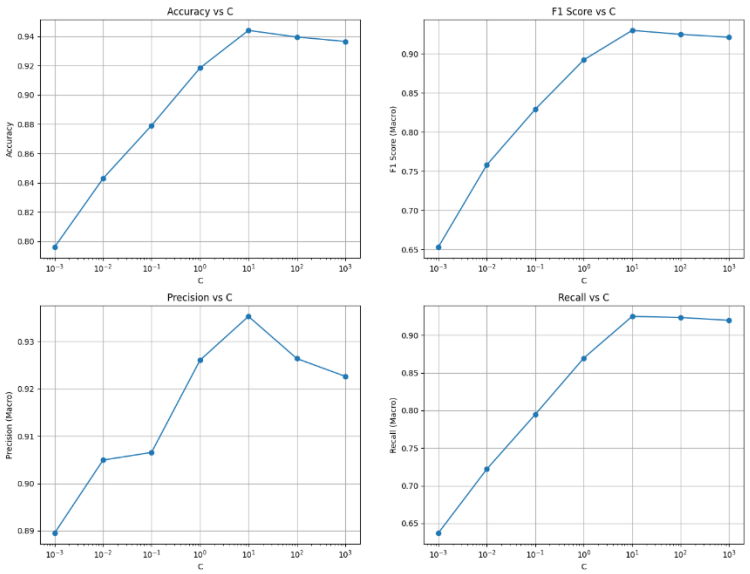


图 2.1.3.2.3 核函数为 poly 时

3、Gamma

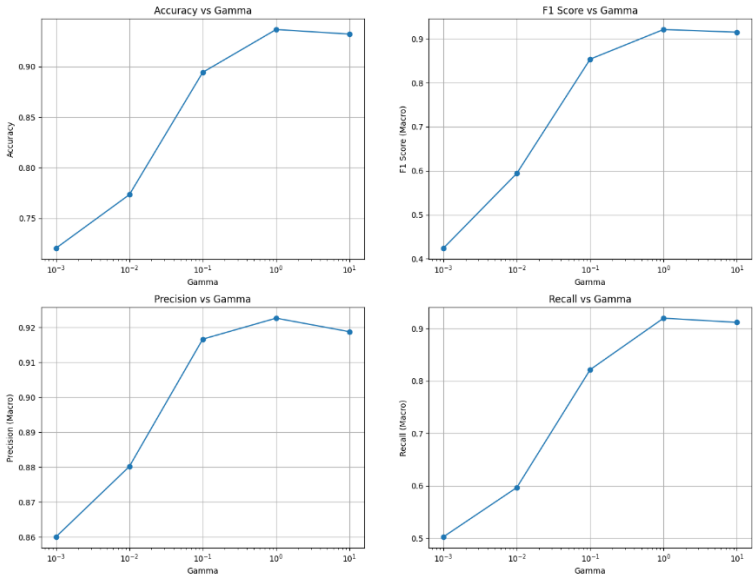


图 2.1.3.3.1 核函数为 poly 时

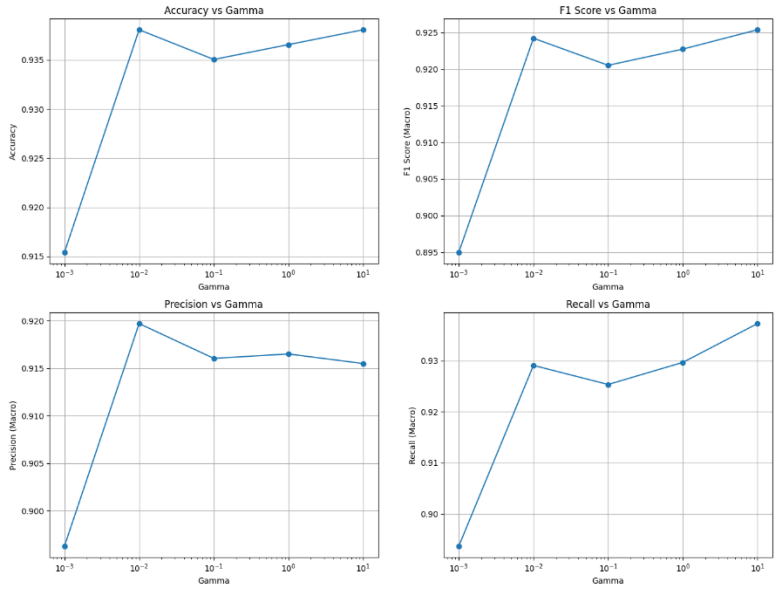


图 2.1.3.3.2 核函数为 rbf 时

4、Degree

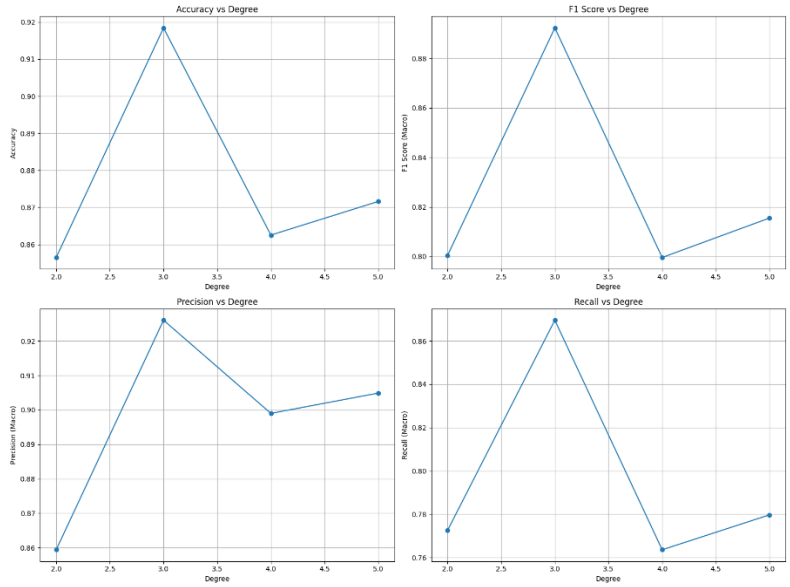


图 2.1.3.4 不同 Poly 次数下精度

从各个图表中可以知道，对于本次实验数据，最佳的核函数是 linear，对于 linear 核函数，C 值为 10 的 3 次时最佳。参数 C 在不同的核函数下对结果精度的影响不同，并非越大越好。参数 Gamma 在核函数为 Poly 时影响较大，最佳值是 1。Degree 最佳值是 3。可以看出，一般情况下，默认参数就够用了。

2.2 AdaBoost 自适应增强算法

（1）简介

AdaBoost 是一种集成学习方法，通过结合多个弱学习器（通常是简单的模型，如决策树桩）来构建一个强大的分类器。AdaBoost 的核心思想是通过迭代地调整数据的权重，使得后续的弱学习器更加关注之前模型错误分类的样本，从而逐步提高整体模型的性能。

（2）参数

1. estimator

- 含义：弱学习器的类型。
- 作用：指定用于构建 AdaBoost 模型的弱学习器。默认情况下，如果未指定，通常使用单层决策树（决策树桩）。
- 默认值：None（表示使用默认的弱学习器，通常是决策树桩）

2. n_estimators

- 含义：弱学习器的数量。
- 作用：指定 AdaBoost 模型中要组合的弱学习器的数量。增加数量可以提高模型的性能，但同时也会增加计算成本。
- 默认值：50

3. learning_rate

- 含义：学习率。
- 作用：控制每个弱学习器在最终模型中的贡献权重。较低的学习率会使模型更谨慎地结合弱学习器，可能需要更多的弱学习器来达到较好的性能。
- 默认值：1.0

4. random_state

- 含义：随机种子。
- 作用：确保结果的可重复性。在训练过程中，随机种子可以控制样本权重的初始化和其他随机操作。
- 可选值：
  - None：不设置随机种子。
  - 整数：指定随机种子。
- 默认值：None

2.3 BPNN 反向传播神经网络

（1）介绍

BPNN 是一种基于人工神经网络的监督学习算法。它通过模拟人脑神经元的连接方式，利用多层结构（输入层、隐藏层和输出层）来学习数据中的复杂模式。

BPNN 的核心思想是通过正向传播计算网络的输出，然后通过反向传播调整网络的权重，以最小化预测值与真实值之间的误差。这种算法能够自动学习数据中的特征表示，适用于各种复杂的非线性问题。

（2）部分参数

1. hidden_layer_sizes

- 含义：隐藏层的神经元数量和层数。
- 作用：定义神经网络的结构，即每个隐藏层有多少个神经元。
- 默认值：(100,)（表示有一个隐藏层，包含 100 个神经元）

2. activation

- 含义：隐藏层的激活函数。
- 可选值：
  - "identity"：线性激活函数。
  - "logistic"：Sigmoid 激活函数。
  - "tanh"：双曲正切激活函数。
  - "relu"（默认）：ReLU 激活函数。
- 作用：为神经元的输出引入非线性，使网络能够学习复杂的模式。
- 默认值："relu"

3. solver

- 含义：优化算法。

- 可选值：
    - "lbfgs"：拟牛顿法，适合小数据集。
    - "sgd"：随机梯度下降，适合大数据集。
    - "adam"（默认）：自适应矩估计，适合大多数情况。
- 作用：用于优化损失函数，更新网络的权重。
- 默认值："adam"

4. alpha
- 含义：L2 正则化参数。
- 作用：通过在损失函数中加入权重的平方和，防止过拟合。
- 默认值：0.0001

5. batch_size
- 含义：每次更新模型参数时使用的样本数量。
- 可选值：
    - "auto"（默认）：自动选择合适的批量大小。
    - 整数：指定批量大小。
- 作用：控制每次迭代的计算量，影响模型的收敛速度。
- 默认值："auto"

6. learning_rate
- 含义：学习率的调度方式。
- 可选值：
    - "constant"（默认）：固定学习率。
    - "invscaling"：学习率随迭代次数的增加而减小。
    - "adaptive"：学习率根据训练进度动态调整。
- 作用：控制权重更新的步长。
- 默认值："constant"

7. max_iter
- **含义**：最大迭代次数。
- **作用**：限制训练的最大迭代次数，避免长时间运行。
- **默认值**：200

8. random_state
- **含义**：随机种子。
- **作用**：确保结果的可重复性。
- **可选值**：
    - None：不设置随机种子。
    - 整数：指定随机种子。
- **默认值**：None

2.4 RF 随机森林

（1）简介

随机森林是一种基于决策树的集成学习算法，通过构建多个决策树并将它们的预测结果进行组合，从而提高模型的稳定性和准确性。随机森林的核心思想是利用“随机性”来减少单个决策树的过拟合问题，并通过集成多个决策树来提高整体模型的性能。

（2）部分参数

1. n_estimators
- **含义**：决策树的数量。

- **作用**：指定随机森林中要构建的决策树的数量。
- **默认值**：100

2. criterion

- **含义**：分裂节点的评估标准。
- **可选值**：
  - "gini"（默认）：Gini 不纯度，用于分类任务。
  - "entropy"：信息增益，用于分类任务。
  - "squared_error"：均方误差，用于回归任务。
  - "absolute_error"：绝对误差，用于回归任务。
  - "poisson"：泊松分布误差，用于回归任务。
- **作用**：决定如何评估分裂节点的优劣。
- **默认值**："gini"

3. max_depth

- **含义**：决策树的最大深度。
- **作用**：限制决策树的深度，防止过拟合。
- **默认值**：None（表示不限制深度）

4. min_samples_split

- **含义**：分裂内部节点所需的最小样本数。
- **作用**：防止决策树过于复杂。
- **默认值**：2

5. min_samples_leaf

- **含义**：叶子节点所需的最小样本数。
- **作用**：防止决策树过于复杂。
- **默认值**：1

6. random_state

- **含义**：随机种子。
- **作用**：确保结果的可重复性。
- **可选值**：
  - None：不设置随机种子。
  - 整数：指定随机种子。
- **默认值**：None

3 模型结果评估与讨论

3.1 ROC 曲线对比（除 SVM 核函数外，所有参数默认）

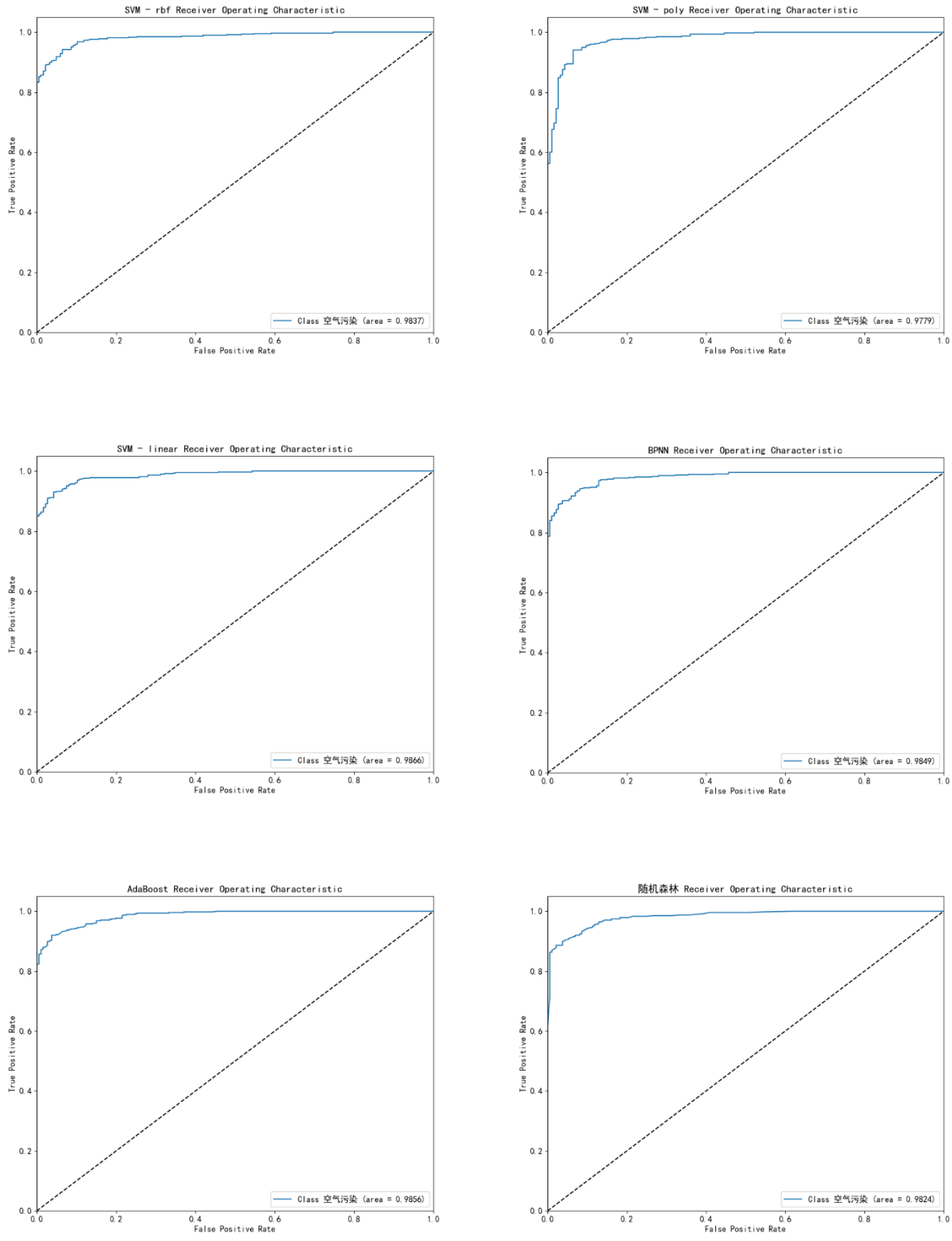


图 3.1 ROC 曲线对比

3.2 PR 曲线对比（除 SVM 核函数外，所有参数默认）

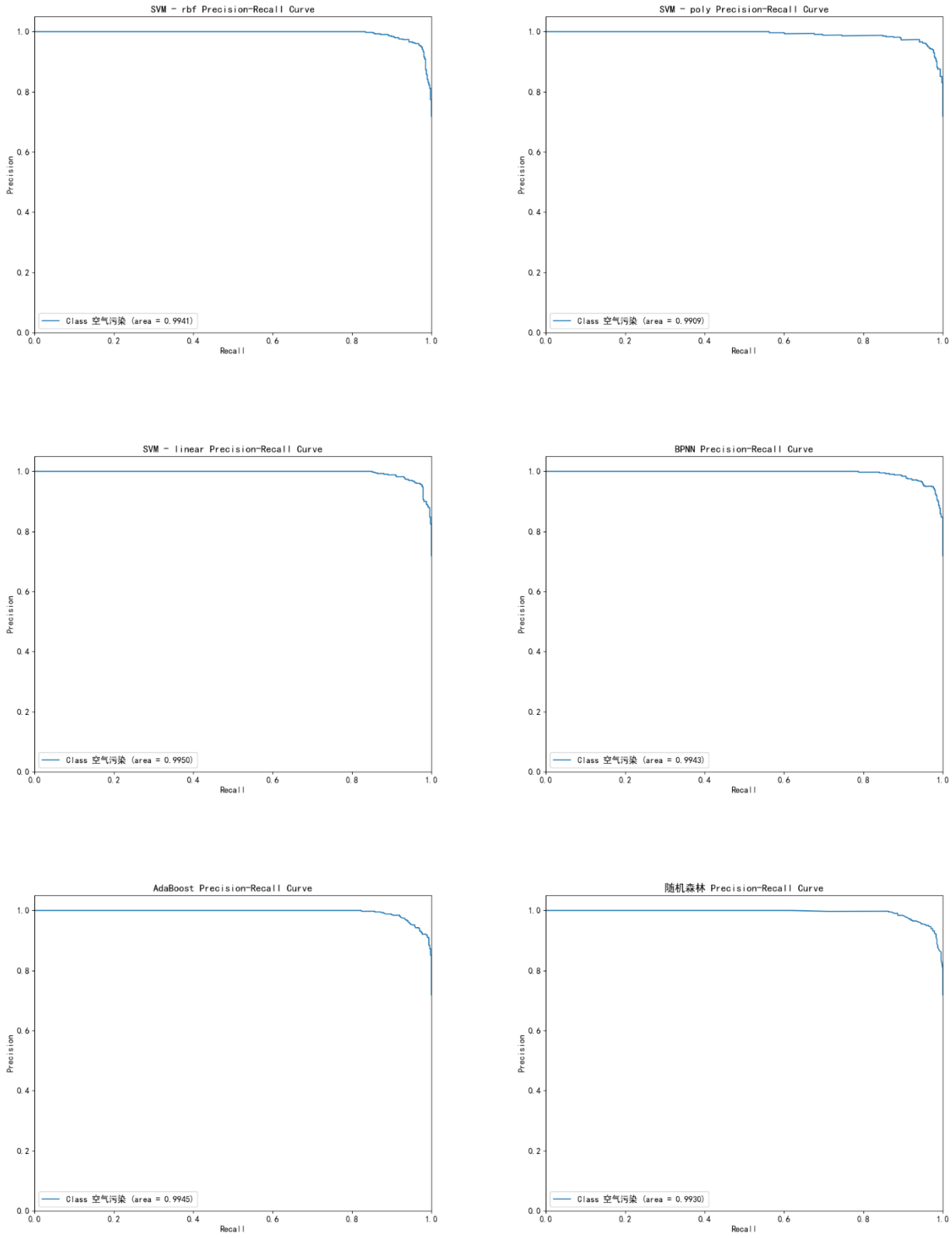


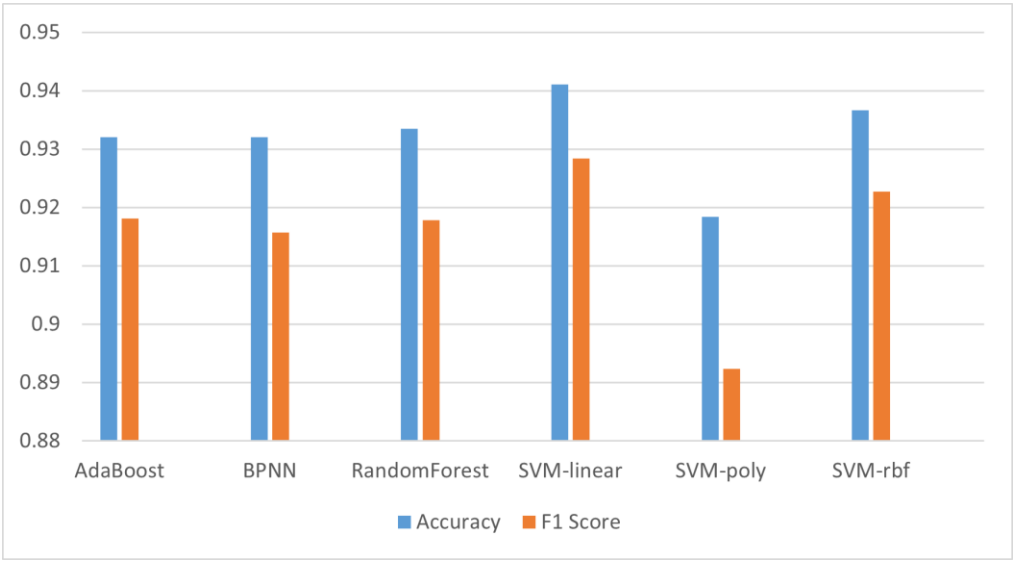
图 3.1 PR 曲线对比

3.4 分类结果汇总表（除 SVM 核函数外，所有参数默认）

Classification Reports						Confusion Matrices			Summary			
Model Name	Class	Precision	Recall	F1-Score	Support	Y Predicted	N Predicted		Accuracy	F1 Score	ROC	PR
AdaBoost	Y	0.97	0.93	0.95	476	445	31	Y Actral	0.932	0.9181	0.9856	0.9945
	N	0.85	0.92	0.88	186	14	172	N Actral				
BPNN	Y	0.95	0.95	0.95	476	454	22	Y Actral	0.932	0.9157	0.9849	0.9943
	N	0.88	0.88	0.88	186	23	163	N Actral				
RandomForest	Y	0.95	0.95	0.95	476	454	22	Y Actral	0.9335	0.9178	0.9824	0.993
	N	0.88	0.88	0.88	186	22	164	N Actral				
SVM-linear	Y	0.97	0.95	0.96	476	451	25	Y Actral	0.9411	0.9284	0.9866	0.995
	N	0.87	0.92	0.9	186	14	172	N Actral				
SVM-poly	Y	0.91	0.98	0.95	476	467	9	Y Actral	0.9184	0.8923	0.9779	0.9909
	N	0.94	0.76	0.84	186	45	141	N Actral				
SVM-rbf	Y	0.97	0.95	0.96	476	450	26	Y Actral	0.9366	0.9227	0.9837	0.9941
	N	0.87	0.91	0.89	186	16	170	N Actral				

注：Y="空气无污染" N="空气无污染"

图 3.4.1 汇总表



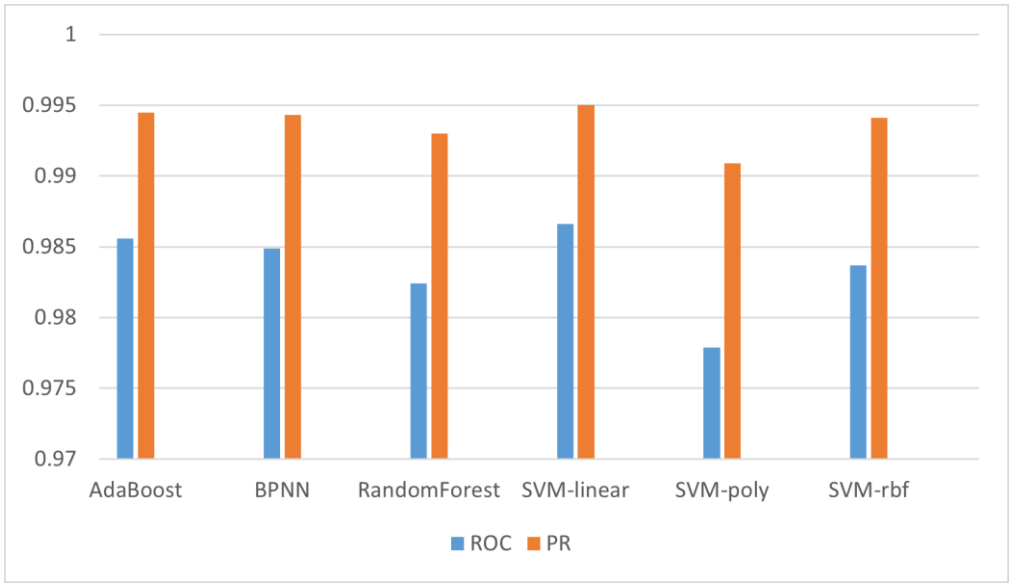


图 3.4.2 柱状统计图

3.3 讨论

（1）模型性能比较

SVM-linear 和 SVM-rbf 在综合指标上表现最优。SVM-linear 的准确率（Accuracy）为 94.11%，F1 分数为 92.84%，ROC 和 PR 曲线下面积分别达到 0.9866 和 0.995，表明其在区分正负类（“污染”与“无污染”）时具有高鲁棒性和稳定性。

AdaBoost 和 RandomForest 在正类（无污染）的召回率（Recall）上表现突出（均为 93%-95%），说明它们能有效减少漏报，但负类（污染）的精确率（Precision）较低（AdaBoost 为 85%），可能导致更多误判。

SVM-poly 表现相对较弱，其负类召回率仅为 76%，且整体准确率（91.84%）和 F1 分数（89.23%）最低，可能是由于多项式核函数对数据特征的非线性拟合能力不足。

（2）实际意义分析

高召回率的实际价值：在空气质量预警场景中，漏报（如未识别出污染事件）可能比误报（如错误预警）后果更严重。因此，AdaBoost 和 RandomForest 的高召回率（Y 类 Recall ≥93%）具有重要应用价值，尤其适合对污染事件敏感的场景。

精确率与误报成本的权衡：SVM-linear 在正负类上均保持较高精确率（Y 类 Precision 97%，N 类 Precision 87%），适合对误报容忍度较低的场景（如避免不必要的应急响应）。

模型复杂性与效率：SVM-linear 和 SVM-rbf 的性能接近，但线性核（SVM-linear）通常训练速度更快，更适合实时预测需求。

（3）总结与建议

推荐模型：若需平衡准确率与稳定性，SVM-linear 是最优选择；若需最大限度减少漏报，可优先考虑 AdaBoost 或 RandomForest。

改进方向：对于负类（N）识别较弱的模型（如 AdaBoost），可通过调整类别权重或引入更多负类样本优化性能。

实际部署考量：需结合硬件资源与实时性需求，例如轻量级模型（如 SVM-linear）更适合资源受限环境，而集成模型（如 RandomForest）可能更适合离线分析。