2019 年（第六届）全国大学生统计建模大赛

基于特征工程的 SVM
北京雾霾成因分析及预测

参赛单位	北京科技大学
参赛者姓名	1. 王顺钢
	2. 李文蝶
	3. 杨佳亦

基于特征工程的SVM北京雾霾成因分析及预测

目录

表格和插图清单

基于特征工程的SVM北京雾霾成因分析及预测

摘要

自2013的“雾霾中国”起，“雾霾”成为以北京为首的大中型城市发展中的“拦路虎”，危害人身体健康，影响生态环境的同时极大降低了居民生活幸福指数；而PM2.5指数是目前衡量雾霾污染最普及最重要的指标，对其进行有效的监控预测对实现雾霾污染治理具有重要的理论和现实意义。

由此，本文以UCI网站所得的2010年-2014年的气象数据作为训练集，根据相关气象影响因子的数据建立基于特征工程的SVM多分类模型，取名为FEMC-SVM，预测PM2.5的等级（优,良,中,差），FEMC-SVM模型具体建立，优化过程及结果如下：

首先对预处理后的数据进行相关性分析，选出与PM2.5相关性较强的5个气象因素并将其和PM2.5的数据为训练集建立基于决策树的多分类支持向量机模型，但经初步人工指定参数模型验证结果的最佳准确率仅为0.5。

为充分挖掘数据信息，引入特征工程(Feature Engineering)，对数据集进行标准化和PCA优化，获得彼此无相关关系的主成分数据，提高数据和特征所决定的机器学习“上限”；并使用Grid Search暴风搜索遍历所有的参数组合，以模型准确率为标准进行参数优化，得到最好的参数组合为：[核函数：rbf，gamma=2.08，惩罚因子C=2]，对应的FEMC-SVM模型最高准确率为62.8%--对所建立四分类问题，表明FEMC-SVM模型的拟合效果较好。

同时，为了展现所建立FEMC-SVM模型的优越性，分别以原数据和引入特征工程对数据处理后的数据为训练集，将机器学习常见的分类器：KNN，随机森林，决策树等与所建立的FEMC-SVM模型进行对比分析，证明了所建立的FEMC-SVM模型的错误率更低，且对特征工程处理后的数据进行拟合所得的准确率明显优于对原数据的拟合。

最后，就FEMC-SVM模型应用，从目标结果的可行性和实用性考虑，通过绘制各影响因子的6维联合分布图获得人们日常可以感知的不同的典型气象情况的参数定量范围，然后采用情景分析法根据不同情景下FEMC-SVM模型应用结果，对人们的日常出行提出合理化建议为：低温干燥且风速较高时空气质量较“优”，适宜户外活动；高湿，温度稍高且风速适宜时空气质量为良，可进行正常的户外活动；高温高湿风速大时空气质量中等，不适宜敏感人群外出；而寒冷干燥无风时空气质量较差，建议所有人尽量避免外出。

关键词： PM2.5分析预测，特征工程，相关性分析，多分类支持向量机，情景分析法

一，绪论

（一）选题背景及意义

雾霾，是雾和霾的统称。它由空气中大量的灰尘和污染气体构成，二氧化硫、氮氧化物和可吸入颗粒物是雾霾组成的三大主要物质。雾霾常见于城市。中国不少地区将雾并入霾一起作为灾害性天气现象进行预警预报，统称为“雾霾天气”。雾霾是特定气候条件与人类活动相互作用的结果。高密度人口的经济及社会活动必然会排放大量细颗粒物（PM 2.5），一旦排放超过大气循环能力和承载度，细颗粒物浓度将持续积聚，此时如果受静稳天气等影响，极易出现大范围的雾霾。

2013 年，“雾霾”成为年度关键词。在短短的一年中，雾霾肆虐了整个中国——它影响范围广：笼罩超过 30 个省市；持续时间长：就北京而言，该年内仅有 5 天不是雾霾天；危害程度大：就年底而言，上海多地多次出现 PM2.5 数据超过 500（重度污染等级）。这一年，中国较大的 500 个城市中，仅有 1%的城市空气达到世界卫生组织标准。此后的数年，雾霾问题一直成为中国老百姓所关心的问题。2016 年 9 月，世界卫生组织指出，目前仍有 92%的人口在低标准空气质量环境中生存，每年有近三百万人死于空气污染。近年来，尽管大规模的雾霾状况未有出现，但不时升高的 PM2.5 粒子浓度仍时刻让人警惕。而且，雾霾带来的灾害，不仅仅包括空气质量的降低，更涉及生态环境的破坏--它的大范围肆虐会同时导致人们交通出行的不便、身体健康的损害、对动植物生存的威胁、自然气候的改变以及经济发展的受阻。因此，对雾霾状况的预测、监管与治理刻不容缓。

身处于如今的大数据时代，数据中蕴含着丰富的信息。机器学习已应用于人们生活的方方面面，如人脸识别、预测天气、垃圾邮件过滤等等，远超出大多数人的想象。随着各种数据以指数级增长，需要使用良好的工具、优化的算法对数据进行分析，并通过这些数据挖掘及分析掌握其内涵，从中获取有效知识信息并应用以提高人类应对未知世界的能力。如果我们能通过对海量数据进行处理、分析、挖掘，建立恰当的模型来拟合雾霾状况，那么数据中的有用信息则能为我们所用，为雾霾的预测、监管与治理助力。

（二）研究现状

近年来，不同于以往的气象学预测方法，气象部门及众多学者基于数据对短期雾霾情况预测进行了大量的研究。经查阅资料可知，短期的雾霾状况预测主要是利用统计学方法和动力学方法从经验统计逐步走向数理统计和动力理论。在目前的研究中，我们发现，主要有以下三种方法（模型），即灰色理论模型、指数

平滑模型理论、神经网络模型，具备较好地拟合雾霾浓度随时间、空间变化的特征的能力;用这三类模型对雾霾浓度进行预测,多数情况下都有较好的预测精度。

灰色理论模型利用微分方程建立预测模型的依据，它能够将无规律的历史数据经累加后，变为有指数增长规律的上升形状数列，而一阶微分方程解的形式恰为指数增长形式；通过灰数的不同生成方式、不同级别的 GM 模型来调整、修正、提高预测精度。应用在雾霾预测中，灰色理论模型[1-5]能够为 PM2.5 浓度在时间维度上进行拟合，并尝试建立预测模型。如王江洪等[1]（2018）应用郑州市 2008 年至 2017 年间的空气质量监测数据结合综合污染指数法对该市的空气环境质量进行基础评价，同时利用 GM(1，1) 灰色数学模型对该市的空气环境质量进行预测；

指数平滑模型能够对不同时间的观察值赋予不同的权值，因此通过加大近期观察值的权数，可以加强观察期近期观察值对预测值的作用，使预测能够迅速反映实际的变化。在对未来雾霾浓度进行预测时，尤其适合使用该模型，原因在于相比灰色理论模型，它有选择地使用时间序列数据，强调了近期的观测值对未来雾霾浓度的影响，使得预测结果更加贴合生活实际，提高了预测精度[6-10]。侯琼煌等[6]（2014）通过建立 3 次指数平滑模型对未来 3 年的雾霾天气状况进行预测，得出未来三年内我国雾霾天气仍会频发的结论，并就其原因进行了讨论。

神经网络模型则通过一系列的“神经元”（输入数据与权数相乘，加上偏置项）的计算，挖掘输入数据的深层信息，得到输入与输出间特有特征的自适应系统。对于包含时空尺度下的雾霾浓度数据，具有非线性等复杂关系，用普通的线性模型无法准确拟合，因此，人们就利用深度学习的挖掘非线性等复杂关系能力强的优点来寻找雾霾浓度在不同时间、空间下的分布规律 [11-15]。黄伟政[11]实现了卷积-回归神经网络对雾霾浓度的预测，并进一步多卷积联合神经网络，对雾霾时空数据进行了细分尺度的训练，对雾霾的空间变化做了 LISA 集聚图的分析；宋利红等[12]通过小波变换实现了 PM2.5 浓度序列不同时间尺度的周期变化及其在空间中的分布，后构建了深度置信-BP 神经网络雾霾预测模型 ，最后搭建了预测结果更优的 RNN 深度循环神经网络模型。

但是现在看来，还没有一个比较好的方法能够通过气象数据，如温度、气压、累计风速等因素来准确预报未来雾霾状况，并根据 PM2.5 浓度与影响因子相关性图来预测短期雾霾浓度--这便成为我们的建模目标。

而在我们选择对 PM2.5 进行分析预测模型时，因为线性分类器（Logistic regression）受制于特征与目标间的线性假设，要求分类问题必须线性可分，忽略交互效应和非线性因果关系；朴素贝叶斯给定目标值时设定属性之间相互条件独立，而 PM2.5 的气象影响因子间的相关性不可忽略；决策树往往只能达到局部最优结果，因含有随机错误或噪声，容易过拟合；KNN 不适合影响因子较多的高

维空间且对较小的训练集容易过拟合；灰色模型 GM（1,1）主要应用于对含有不确定因素的系统进行预测，且模型输出值为未来某一时刻的特征量，或达到某一特征量的时间，无法达到输出雾霾等级的期望...而 SVM 可以使用核函数可以向高维空间进行映射从而解决非线性的分类且分类思想简单（样本与决策面的间隔最大化）--就目前来说，SVM 是针对小样本数据高维分类效果较好的分类器--所以我们选择以基于核技术的非线性 SVM 模型为基础，建立结合特征工程和基于决策树的 SVM 多分类模型，取名为 FEMC-SVM[①]模型，实现此次对 PM2.5 的分析预测。

二，模型构建思路及创新

（一）FEMC-SVM 模型构建思路

本文《基于特征工程的 SVM 北京雾霾成因分析预测》主要基于气象因素对雾霾进行分析；

第一部分是数据预处理和相关性分析：选取 UCI 网站上 2010 年-2014 年的气象数据作为训练集，首先通过 SPSS 和 Python 进行各气象因子与 PM2.5 浓度的相关性分析；其次，出于模型的实用性考虑----由于查阅当天的空气质量时，比起 PM2.5 的具体浓度，人们往往更加关心空气污染的等级程度，所以我们按照 PM2.5 检测网的空气质量新标准将 PM2.5 的浓度分为“优”，“良”，“中”，“差”四个等级，模型预测结果输出为当天的空气质量等级。

第二部分是初步建立 SVM 模型预测：通过由相关性分析选出的影响因子和 PM2.5 初步建立基于决策树的多分类 SVM 模型，化二分类算法为四分类算法，分析模型训练结果的同时与其它模型比较，展现 SVM 模型的优越性；

第三部分为 SVM 模型的优化：为了更好进行结果优化，引入特征工程处理数据的思想，对数据进行标准化和 PCA 处理，并进一步通过 Grid search 穷举算法调参实现对模型参数的优化，并将优化后的模型取名为 FEMC-SVM 模型；

最后为 FEMC-SVM 模型的应用：采用情景分析法，根据所得数据的 6 维分布图获得情景参数定量范围，并根据应用结果对人们在几种不同气象条件的出行提出合理的建议。

（二）FEMC-SVM 模型创新

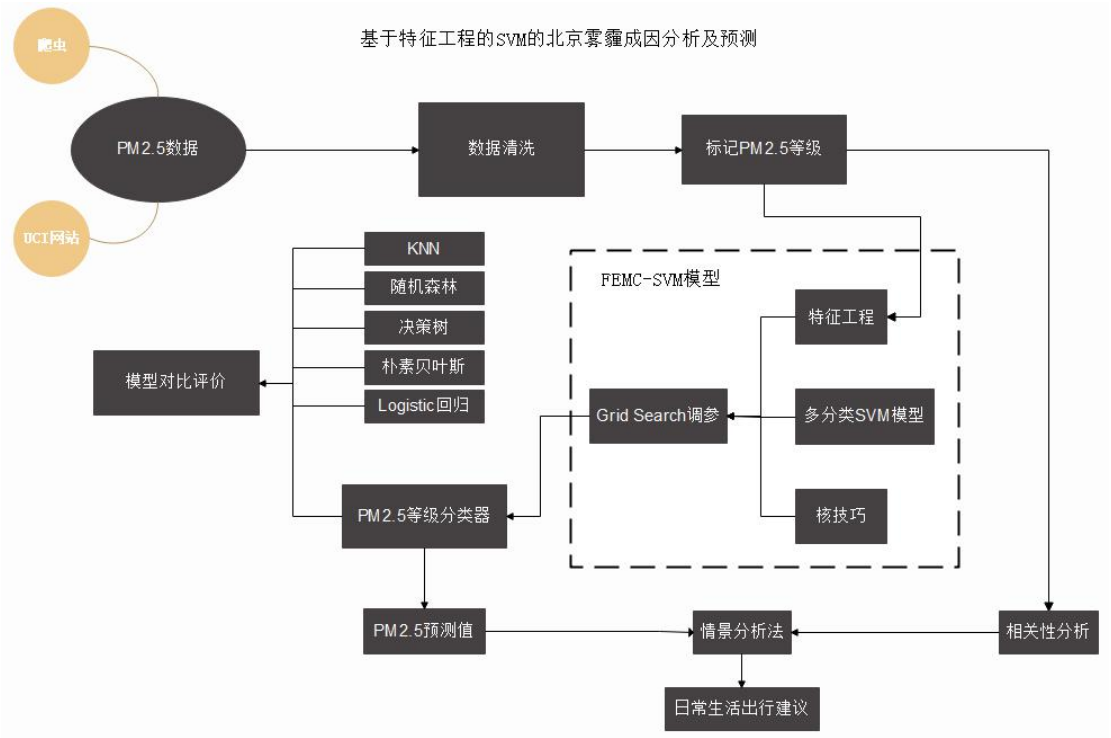
[①] Feature Engineering Multiple Classifiers:SVM

1， 将对气象热词“雾霾”的分析化为通过人们可以感知的气象影响因子预测“雾霾”典型指标--PM2.5 等级问题，并基于决策树的 OVR SVM 用原 SVM 二分类器实现判别 PM2.5 污染等级的多分类目标；

2，引入特征工程(Feature Engineering)优化数据集的思想，并将其与 Grid search 穷举算法调参结合，建立优化后的 FEMC-SVM 模型，并在优化前后与不同的分类器进行比较，验证 FEMC-SVM 模型更好的分类预测结果；

3，模型应用时，利用 Python 语言绘制 6D 分布图展现不同等级的 PM2.5 的各影响因子联合分布情况，并根据 6D 图所得信息采用情景分析法对人们在不同气象因子数据组合条件下的出现提出合理的现实建议。

（三）FEMC-SVM 模型构建流程图



三，变量描述及数据预处理

（一）变量描述

雾霾是漂浮大气中的 PM2.5 等尺寸微粒、粉尘、气溶胶等粒子，在一定的湿度、温度等天气条件相对稳定状态下产生的天气现象。由其定义可知，影响雾霾的因素众多----限于自身所能获得的数据有限，本模型预测时仅考虑影响雾霾的气象因子；同时，为了定量的分析处理雾霾相关的数据，我们选取了目前广受

大众认同的 PM2.5（环境空气中空气动力学当量直径小于等于 2.5 微米的颗粒物）为评价指标，查阅相关资料后选取的影响因子如下表 1：

影响因子	释义
DEWP	露点：空气中水气含达到饱和的气温(a, f),表征湿度
TEMP	观测时间点对应的温度(a,f)
PRES	观测时间点对应的压强(h*Pa)
Iws	累积风速，观测时间点对应的累积风速(m/s)
Is	累计降雪，到观测时间点为止累计降雪的时长(小时)
Ir	累计降雨，到观测时间点为止累计降雨的时长(小时)

表 1

（二）数据来源

数据文件夹中的 data_pro.csv 数据为 UCI 网站上所找的训练数据，因为影响因子较为全面，用于构建模型；data_new.csv 为爬虫所得数据，作为后期的预测数据集预测未来数据；

UCI 网站：

【UCI Machine Learning Repository: Beijing PM2.5 Data Data Set】

http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data#

爬虫网站：URL:

http://www.tianqihoubao.com/aqi/beijing-201703.html

（三）数据预处理

1，数据整理

鉴于从 UCI 网站上下载的数据为 2010 年 01 月 02 日 0 时--2014 年 12 月 31 日 23 时每小时各气象因子和 PM2.5 的数据，为了使雾霾（PM2.5）的预测更具现实意义，通过 Python 进行数据预处理，取 24 小时的数据平均值为该天的气象数据值，同时对缺失的日数据（包括一天内没有 24 小时数据值的）进行了剔除整理。

2，SPSS 软件描述变量分布

为了检查检验数据集的完整性和数据的大致分布，用 SPSS 软件的“描述统计”对各影响因子及 PM2.5 的数据进行分析，结果如下表 2：

因子/统计量	样本数	极差	最小值	最大值	平均值	标准差
Pm2.5	1788	549.52	2.96	552.48	98.70	77.55

DEWP	1788	59.54	33.33	26.21	1.85	14.16
TEMP	1788	47.33	14.46	32.88	12.50	11.52
Iws	1788	462.00	1.19	463.19	23.81	40.98
PRES	1788	49.42	994.04	1043.46	1016.40	10.07
Is	1788	14.17	0.00	14.17	0.05	0.55
Ir	1788	17.58	0.00	17.58	0.20	1.01

表 2

由表 2 可知，处理后的 data_pro_aday.csv 没有缺失数据，故无需进行缺失值计算。

3，Python 绘图展示变量分布

为了更好的表示各类数据的分布，绘制各个影响因子和 PM2.5 的数值分布如下图 1：

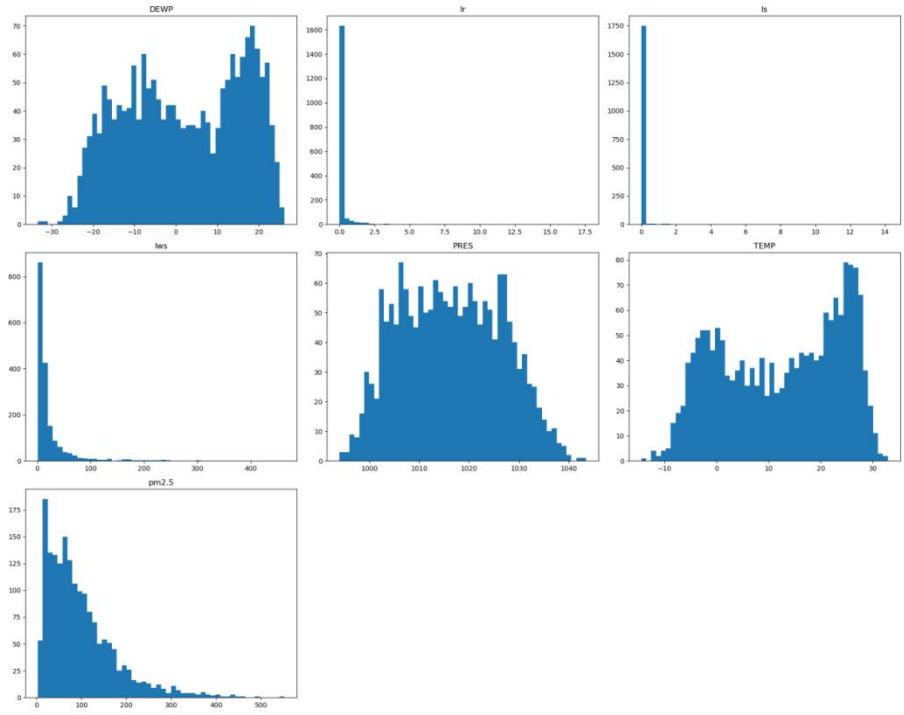


图 1

图 1 中，横轴为各变量取值范围，纵轴为该变量在对应值的数量分布；由图可知，各变量的分布大不相同，其中，DEWP,TEMP,PRES 的分布图大致近似为正态分布；而 Is,Ir（累计降雪，降雨量）的分布图像，近似一条横坐标在（0,2）间的竖直直线集中在 0 到 0.2 之间；Iws 分布处于常数分布于对数正态分布间，三者结合反映了北京干燥多风的现实，而 PM2.5 的分布近似对数正态分布。

四，PM2.5 与气象影响因子的相关性研究

（一）数据可视化分析

为了探究 PM2.5 数值和气象数据因子的相关性强弱，首先对数据进行可视化处理--以各影响因子数据为横轴，PM2.5 数值为纵轴分别绘制散点图，观察 PM2.5 与各影响因子间有无明显的趋势和联系。各图绘制结果如下图 2：

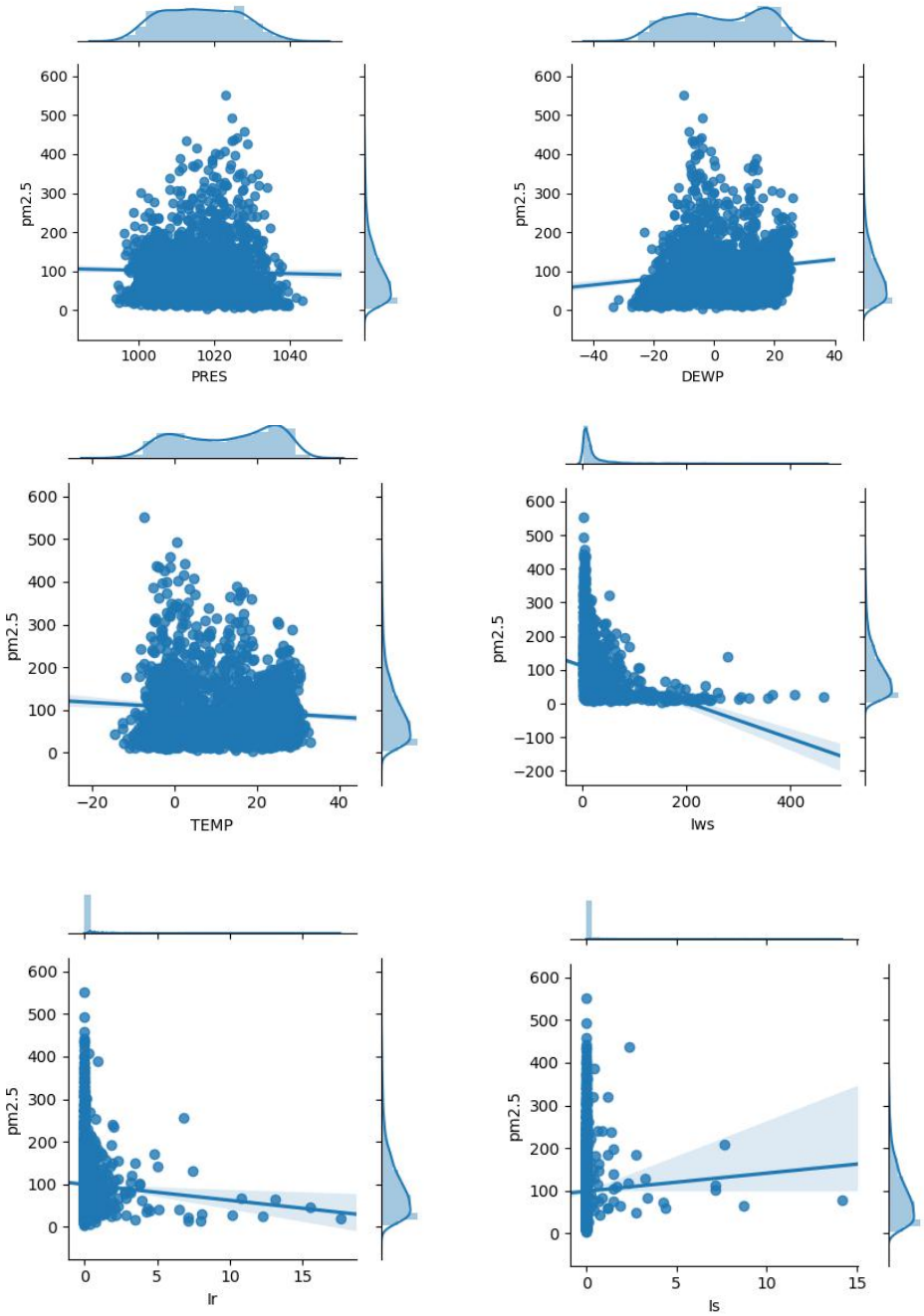


图 2

由各散点图可看出 Iws, Is, Ir 不为 0 时，与其散点图的分布大致沿趋势线分布，说明三者与 PM2.5 的浓度有一定的影响；由于数据较多，而 DEWP, TEMP，PRES 的数据分布较为集中，故无法直接由散点图判断它们与因变量 PM2.5 是否具有一定的相关性，值得一提的是，PRES 的散点图趋势线近似为水平直线，故粗略预

判其可能不是 PM2.5 的有效影响因子（对 PM2.5 数值的影响不大，相较其他气象因子可以忽略）。

（二）相关系数计算分析

虽然散点图能较为直观地展示各影响因子和 PM2.5 间的独立分布关系，但它无法对相关关系进行准确的度量，缺乏说服力。故我们采用统计软件 SPSS 计算 data_pro_aday.csv 中各影响因子和 PM2.5 的统计学三大相关性系数--Pearson 系数，Kendall 系数，Spearman 系数来比较各气象数据与 PM2.5 的相关性大小。

总的来说，三个相关性系数反应的都是两个变量之间变化趋势的方向以及程度，其取值范围均为[-1，1]，0 表示两个变量不相关，正值正相关，负值负相关，相关系数绝对值越大相关性越强。

1，相关系数基础

①，Pearson 相关系数

最常见的相关系数，也成积差相关系数，用来衡量两个变量线性相关程度的指标；计算公式为：（X,Y）为两个随机变量，x_i, y_i为两变量可能的取值；

$$\rho_{xy}=\frac{\operatorname{cov}(X,Y)}{\sqrt{Var(X)Var(Y)}}=\frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{E(X^2)-E(X)^2}\sqrt{E(Y^2)-E(Y)^2}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

由公式可知，Pearson 相关系数是协方差与标准差的比值，用该系数准确判断相关性时一般假设数据成对的来自正态分布的总体，相关性系数受极端值影响较大，且该系数只能衡量两变量间是否存在线性相关关系。

②，Kendall 相关系数

Kendall 相关系数是一种秩相关系数，用于反映分类变量的相关性，适用于对两个有序变量进行非参数的相关性检验，由于其相关性系数的计算公式随具体数据情况变化，可直接通过 SPSS 求解，故不再列出具体的繁琐公式。

③，Spearman 秩相关系数

“秩”，即为一种顺序或排序，因此该相关系数是根据原始数据的排序位置进行求解，计算公式为：$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$；

公式说明：对两变量(X,Y)的数据进行排序，记下排序后的位置-秩次：(X’,Y’)，秩次差值即为d_i，变量（X,Y）的数据个数即为n。

由于Spearman系数没有对变量数据数量，分布和极端值的限制，故适用范围较广，尤其在分析多组交叉数据间的相关性时，使用频率很高。

2，SPSS软件相关系数计算结果

根据SPSS软件计算三大相关性系数的结果汇总如下表3：

影响因子	Pearson	P1值	Kendall	P2值	Spearman	P3值
DEWP	0.15	0.00	0.17	0.00	0.25	0.00
TEMP	-0.09	0.00	0.01	0.63	0.01	0.66
PRES	-0.03	0.25	-0.07	0.00	-0.11	0.00
Iws	-0.28	0.00	-0.29	0.00	-0.42	0.00
Is	0.03	0.20	0.10	0.00	0.13	0.00
Ir	-0.05	0.04	0.06	0.00	0.08	0.00

表3

表中，第2，4，6列为各影响因子不同相关系数的具体数值，其后列的P_1，P_2，P_3分别为各相关系数的显著性水平，P值越小，两变量间相关性的判别可信度越高。

其中，假设检验验证变量相关性时，设H0：两变量无关，P值即为H0成立的概率；当P大于某一值α（选定的显著性水平，一般为0.05或0.01）时，接受原假设，即认为两变量无关；当P小于α时，拒绝原假设，即认为两变量存在相关性。特别地，当α=0.05时，若P<α，则表明二者相关性显著；当α=0.01时，若P<α,则表明二者相关性非常显著；

因此，挑选建立模型的影响因子时应在P值小于0.05的前提下，比较各因素与PM2.5间的相关系数大小：由Pearson相关系数，选出DEWP,Iws,Ir三个影响因子；又由Kendall，Spearman相关系数，将Is补充选入建立模型的影响因子中。

3，Python绘图

为直接观察验证 SPSS 的特征选择，用 Python 将各影响因子的三大相关系数数值绘制成图，如下图 3：

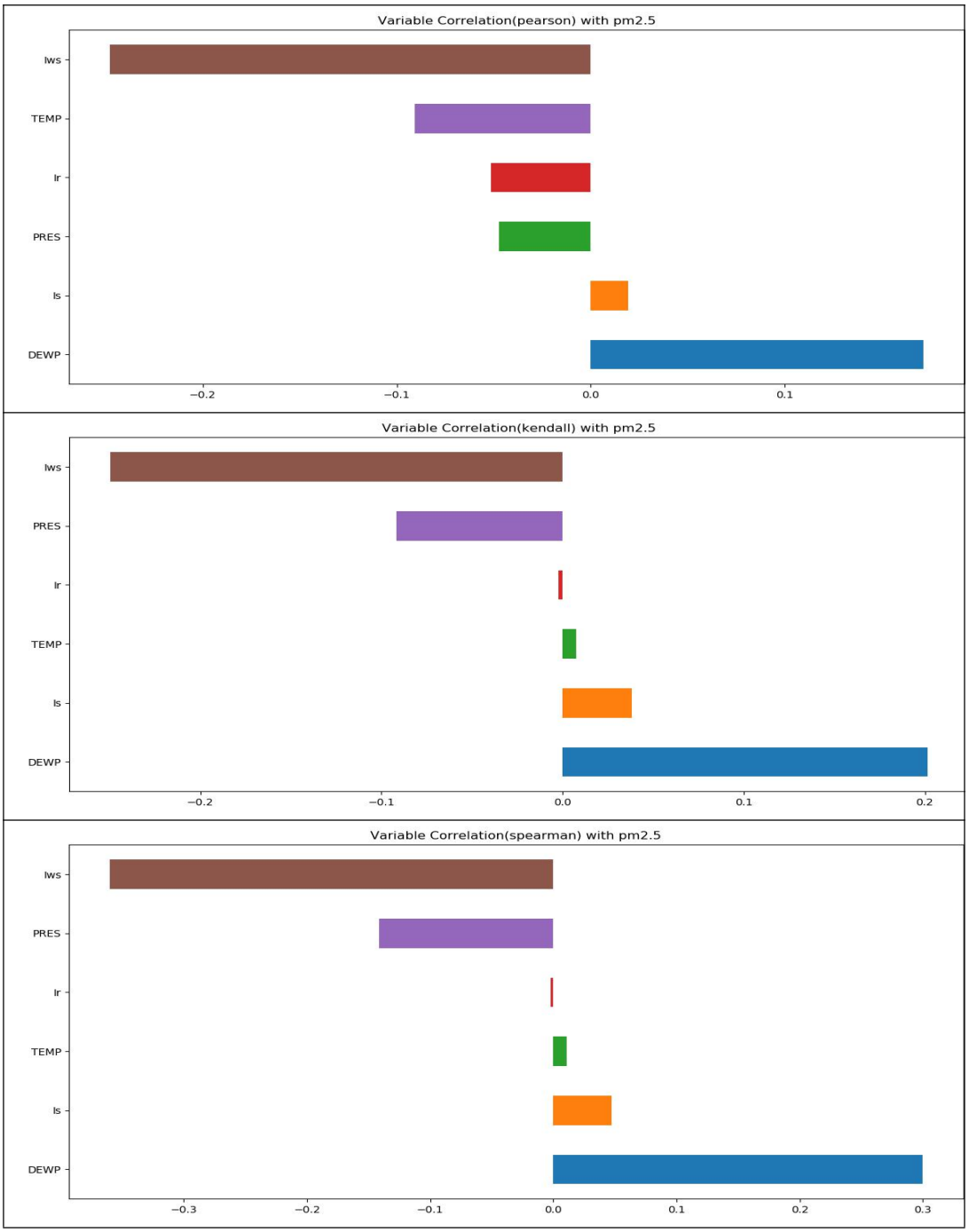


图 3

如图 3，从上至下分别为各个变量 Pearson 系数，Kendall 系数，Spearman 系数的大小比较，而无论何种相关系数，DEWP，Iws 都是相对相关系数最大的两项，故其对 PM2.5 的数值影响力也很大。

另外，为了更好地展示所有变量间的相互关系，绘制了 PM2.5 和各影响因子的热力图，即交叉填充表，用图形展示各离散变量间的组合关系，如下图 4：

图 4

由图 4，第一列数据（除第一行外）展示了所有的影响因子与 PM2.5 的组合关系系数，热力图的颜色越接近橙黄色代表两变量间的关系系数越高，正相关性越强，黑色模块较多一部分由于各变量间的关系系数较小，更大程度上源自变量间负相关而导致的组合关系系数为负；

此外，由热力图可知，各影响因子间并不独立--DEWP 和 TEMP 间的组合关系系数高达 0.82，故建立模型时不能将其视为理想的独立变量。

五，SVM 模型初步构建与评价

（一）支持向量机（support vector machines, SVM）

支持向量机是一种二类分类模型。它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；支持向量机还包括核技巧，这使它成为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化，可形式化伟一个求解凸二次规划（convex quadratic programming）的问题，也等价于正则化的合页损失函数的最小化问题。支持向量机的学习算法是求解凸二次规划的最优化算法。

1，线性可分支持向量机

假设给定特征空间上的训练的数据集：$T=\{(x_1,y_1),(x_2,y_2),...(x_N,y_N)\}$，其中，$x_i\in\chi=R^n, y_i\in Y=\{+1,-1\}, i=1,2,,,N$；$x_i$ 为第 i 个特征向量，也称为实例，y_i 为实例的类标记；（x_i，y_i）称为样本点。学习的目标是找到一个分离超平面，能将实例分到不同的类。分离超平面对应于方程 $w\cdot x+b=0$，它由法向量 w 和截距

b 决定，可用 (w,b) 表示。分离超平面将空间分为两部分，一部分为正类，一部分为负类。给定线性可分数据集，通过间隔最大化或等价求解相应的凸二次规划问题，学习的超平面为 $w\cdot x+b=0$，以及相对应的分类决策函数 $f(x)=sign(w\cdot x+b)$。

支持向量机的基本思想为求解能够正确划分数据集并且几何间隔最大的分离超平面。考虑到空间间隔和几何间隔的关系式后，最大间隔分离超平面的求法等价于求下述约束最优化问题

$$\max_{w,b}\frac{\hat{\gamma}}{\|w\|}$$

$$s.t.y_i(w\cdot x_i+b)\geq\hat{\gamma},i=1,2,....N$$

又因最大化 $\frac{1}{\|w\|}$ 和最小化 $\frac{1}{2}\|w\|^2$ 是等价的，故上式可等价为：

$$\min_{w,b}\frac{1}{2}\|w\|^2$$
$$s.t.y_i(w\cdot x_i+b)-1\geq 0,i=1,2,....N \qquad (*)$$

将（*）作为原始最优化问题，应用拉格朗日对偶性，通过求对偶问题的最优解，得到其最优解。这样做的优点是，对偶问题求解往往更加简便；而且引入核函数，可过渡到非线性分类问题。

首先构建拉格朗日函数，为（*）式，引进拉格朗日乘子 α_i ，定义拉格朗日函数：

$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i y_i(w\cdot x_i+b)+\sum_{i=1}^{N}\alpha_i$ ，其中， $\alpha=(\alpha_1,\alpha_2,...\alpha_N)^T$

根据拉格朗日对偶性，原问题的对偶问题是极大极小问题： $\max_{\alpha}\min_{w,b}L(w,b,\alpha)$

（1）求 $\min_{w,b}L(w,b,\alpha)$

可得： $\min_{w,b}L(w,b,\alpha)=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i$

（2）求 $\min_{w,b}L(w,b,\alpha)$ 对 α 的极大，转化为下列等价的对偶最优化问题：

$$\min_{w,b}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$
$$s.t.\sum_{i=1}^{N}\alpha_i y_i=0 \qquad (**)$$
$$\alpha_i\geq 0,i=1,2,....N$$

设$\alpha=(\alpha_1,\alpha_2,...\alpha_N)^T$为对偶最优化问题（**）的解，因为满足 KKT 条件，故存在下标 j，使得α_j^*>0，则可由：

$$w^*=\sum_{i=1}^{N}\alpha_i y_i x_i$$

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i\cdot y_i(x_i.y_j)=0$$

得到原始最优化问题（*）的最优解w^*、b^*：

则分离超平面可写为$\sum_{i=1}^{N}\alpha_i y_i(x\cdot x_i)+b^*=0$；

分类决策函数可写为$f(x)=sign(\sum_{i=1}^{N}\alpha_i y_i(x\cdot x_i)+b^*)$。

这种算法即为线性可分支持向量机的对偶学习算法。

2，非线性支持向量机与核函数（kernal function）

当分类为非线性时，可将核技巧应用与 SVM,基本思想：将输入空间（欧氏空间 R^n或离散集合）对应于一个特征空间（希尔伯特空间H），使输入空间的超曲面模型对应于H中的超平面模型，即将问题转化为在H中求解线性支持向量机。

定义 1（核函数）

设χ是输入空间，H为特征空间，若映射：ϕ（x）:$\chi\to H$，s.t.对所有$x,z\in\chi$，函数$K(x,z)$满足：$K(x,z)=\phi(x)*\phi(z)$，则$K(x,z)$为核函数，ϕ（x）为映射函数。

如（一）中（**）所示，线性支持向量机的对偶问题中，目标函数与决策函数（分离超平面）均只涉及输入实例与实例间的内积，故可用$K(x,z)=\phi(x)*\phi(z)$代替（**）中的x_i*x_j，对偶问题的目标函数化为：

$$W(\alpha)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i$$

相应地，分类决策函数式为：

$$f(x)=sign(\sum_{i=1}^{N_s}\alpha_i y_i\varphi(x_i)\cdot\varphi(x)+b^*)=sign(\sum_{i=1}^{N_s}\alpha_i y_i K(x_i\cdot x)+b^*)$$

这等价于通过φ将输入空间的内积 $x_i * x_j$ 变换到特征空间的内积 $\varphi(x)\cdot\varphi(z)$，隐式地在新的特征空间里从训练样本中学习线性支持向量机，当φ为非线性函数时，学习到的含有核函数的 SVM 是非线性分类模型。

而就我们的模型而言，选取核函数时，由于相关性分析展示出 PM2.5 与各影响因子间不为简单的线性关系，（各影响因子的 Person 相关系数较小且相关性并不全都显著），所以我们不考虑 linear 核函数;在代码运行时，Poly 核函数的计算量过大且结果并不理想（选择准确率作模型评价标准，当用 poly 核函数时，准确率多为 0.3,0.4，相较另两种核函数偏低）；又鉴于数据处理时，标准化前后数据集的数据分布没有太大变化--所得数据大致服从正态分布,所以初步认为“rbf”为最合适的核函数。后经过分别使用三种核函数的结果（见附录（二））验证了核函数选择的正确性。

“rbf”--高斯核函数的定义如下：

定义 2 （高斯核函数）（Gaussian kernel function）

K（x，z）$=\exp(-\frac{\|x-z\|^2}{2\sigma^2})$， 对应的 SVM 为高斯径向基函数（radial basis function,rbf）分类器；

对应的分类决策函数为：$f(x)=sign(\sum_{i=1}^{N}a_i\cdot y_i\exp(-\frac{\|x-x_i\|^2}{2\sigma^2})+b^*)$

综上，非线性 SVM 的学习算法可总结为：

（1）选取适当的核函数 $K(x,z)$ 和参数 C,求解最优化问题：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i$$

$$s.t.\sum_{i=1}^{N}\alpha_i y_i=0$$

$$0\le\alpha_i\le C, i=1,2,....N \qquad (***)$$

（2）选择α^*的一个正分量$0<\alpha^*<C$,计算$b^*=y_j-\sum_{i=1}^{N}\alpha_i\cdot y_i K(x_i,x_j)$；

（3）构造决策函数：$f(x)=sign(\sum_{i=1}^{N}\alpha_i\cdot y_i K(x_i\cdot x)+b^*)$

当$K(x,z)$为正定核时，（***）为凸二次规划问题，解存在。

由于我们要解决的是线性不可分的多分类问题，即以打上标签的 pm2.5 为分类类别，DEWP，TEMP 等因素为影响分类类别的特征。因此引入核技巧是必要的。选取适当的核函数（即 rbf）可以让我们的分类在更高的维度变得可分，以便得到更好地分类结果。我们构造了$(x_i,y_i)\quad i=1,2,\cdots,n$的分类问题，其中

$x_i=(\alpha_{i1},\alpha_{i2},\cdots,\alpha_{im})\quad m=1,2,\cdots,\quad \alpha_k,k\in 1\sim m$为第$i$条样本的第$k$个特征，

$y_i\in\{0,1,2,3\}$为分类类别。我们以 $(x_i,y_i)\quad i=1,2,\cdots,n$来训练 SVM，而在测试阶段，向训练好的 SVM 输入测试集中的x^*，来得到一个预测结果y，并且与实际的y^*进行对比，通过准确率，召回率等指标进行模型的评价，以便后续进行改进。

（二）运用多个支持向量机解决多分类问题

SVM 是一种典型的二类分类模型，但我们日常生活中要解决的问题，往往是多类的。查阅相关文献后，决定使用王正海等[5](2014)提出的基于决策树的多分类支持向量机算法--构造多个 SVM 的二分类器实现多分类的需求，如下图 5：

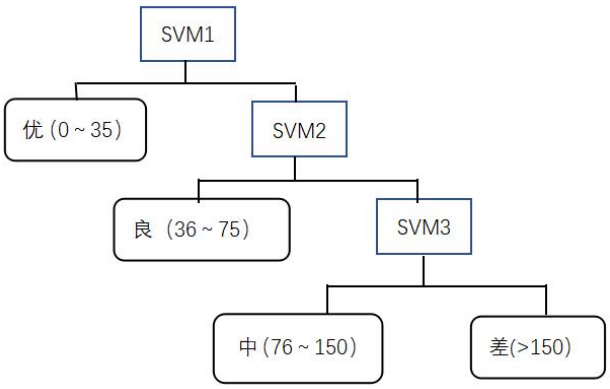


图 5

具体来说，首先将 PM2.5 数据按其浓度高低（如图 6，单位为 ug/m^3）划分为“优”，“良”，“中”，“差”四个等级，分别标记为“0”，“1”，“2”，“3”，模型分类结果即为类别标签的数值；

其次，通过对参考[6]的学习,我们采用由 OVR SVM(“一对多法”)衍生出的基于决策树的分类：首先用 SVM1 将所有数据分为两个类别，再将子类用 SVM2 进一步划分为两个次级子类，从而通过 3 个二分类器实现了最初的四分类问题，达到较精确判断 PM2.5 浓度等级的目标；

另外，用 Python 进行模型训练时，将处理好的数据集进行随机划分，选择其中的 70%进行训练，30%用于测试，从而有利于对模型进行进一步改进。

（三）模型初步训练结果及对比分析（可视化）

首先将 1788 条数据按照 6：2：2 的比例从原始数据集中随机切分为训练集、验证集和测试集，将训练集用于模型拟合的数据样本，将验证集用于调整模型的超参数，将测试集用于模型泛化能力的评估。根据参考[6]，初步建立了基于决策树的 SVM 四分类模型，最初人为指定惩罚因子 C=1，得到三种应用不同核函数时，四分类 SVM 模型的准确率随搜索半径的变化如下图 7：

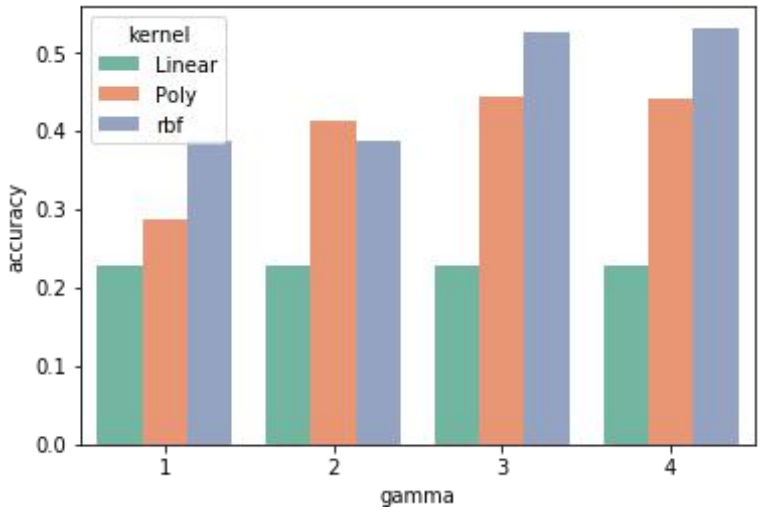


图 6

由图 6 易得，无论参数 gamma 为何值，准确率与不同核函数的关系：rbf>Poly>Linear;且三种核函数的准确率均随 gamma(搜索半径)的增大而增大，但初步调参所得的最高准确率仅为 0.5。

又因为 Ploy（多项式核函数）占据内存过大且准确率明显低于 rbf（高斯核函数）,故接下来仅选取了 Linear（线性核函数）和 rbf 进行不同 C（惩罚因子）,gamma 下模型准确率变化如图 7：

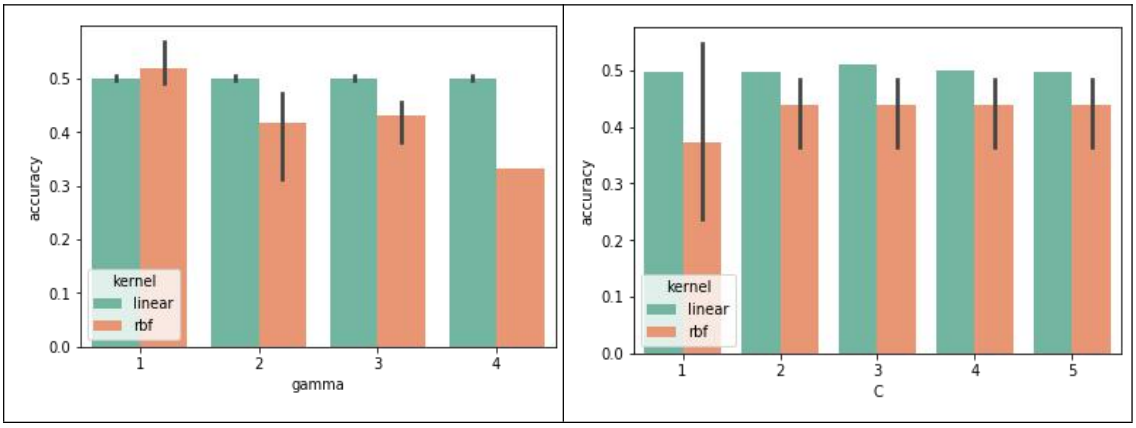


图 7

由图 7，单独考虑某一参数（C,gamma）所得的四分类 SVM 模型准确率没有明显规律，仅可初步估测不同模型参数对模型结果有较大影响；

但所得不同的 C,gamma 数值下，所得模型准确率均未明显超过 0.5，最高数值约为 0.509，故需要对所建模型进行进一步的参数优化。

六，模型优化--FEMC-SVM 模型

（一）特征工程

由机器学习广为流传的话：“数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已”。

特征工程是利用数据领域的相关知识来创建能够使机器学习算法达到最佳性能的特征的过程。而目前认为，特征工程的主要内容如图 8 所示。简而言之，特征工程就是一个把原始数据转变成特征的过程,这些特征可以很好的描述这些数据，并且利用它们建立的模型在未知数据上的表现性能可以达到最优(或者接近最佳性能)。从数学的角度来看，特征工程就是人工地去设计输入变量 X'。

而为了实现这一目标，特征处理便是其中最重要的部分。Python 的 sklearn 库提供了较为完整的特征处理方法，在此，我们选取了其中的两种方法，对各个特征的数据进行标准化并对所有的特征进行 PCA 降维处理。

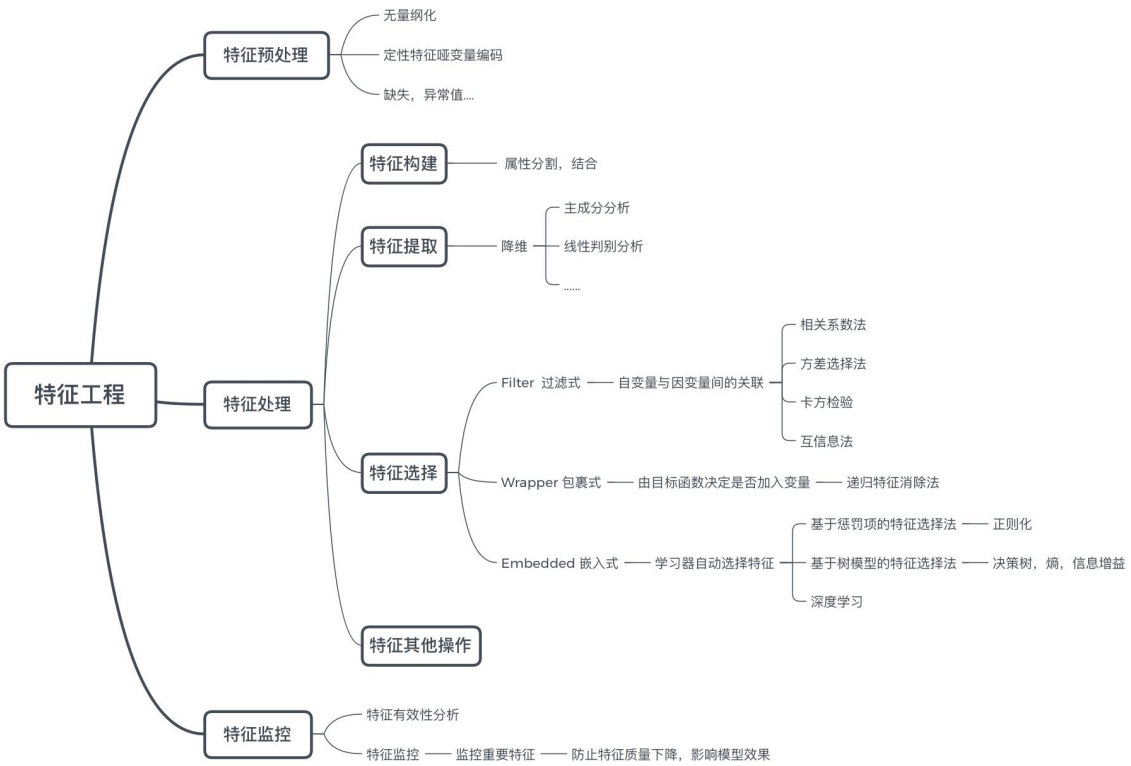


图 8

1，数据的标准化

大型数据分析项目中，数据来源不同，量纲及量纲单位不同，为了让它们具备可比性，需要采用标准化方法消除由此带来的偏差。原始数据经过数据标准化处理后，各指标处于同一数量级，适合进行综合对比评价。这就是数据标准化。即对样本矩阵作如下变换：

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_i}{\sqrt{s_{ii}}},i=1,2,...,m;j=1,2,...,n$$

其中：

$$\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n}x_{ij},\quad i=1,2,...,m$$

$$s_{ii}=\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\bar{x}_j)^2,\quad i=1,2,...,m$$

标准化的意义在于取消由于量纲不同、自身变异或者数值相差较大所引起的误差。图 9 是我们对数据集进行标准化后的数据分布图：

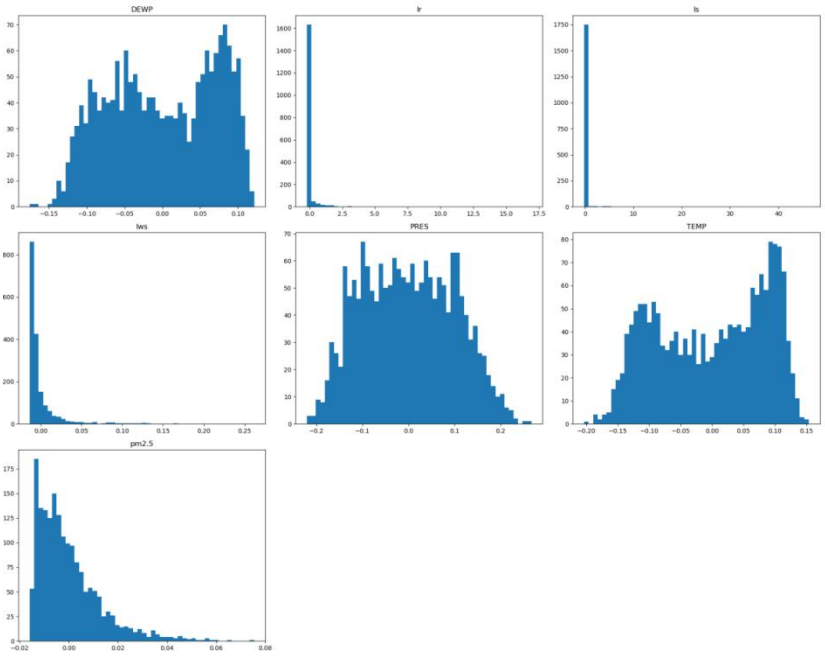


图 9

部分因素与 pm2.5 的密度估计图:

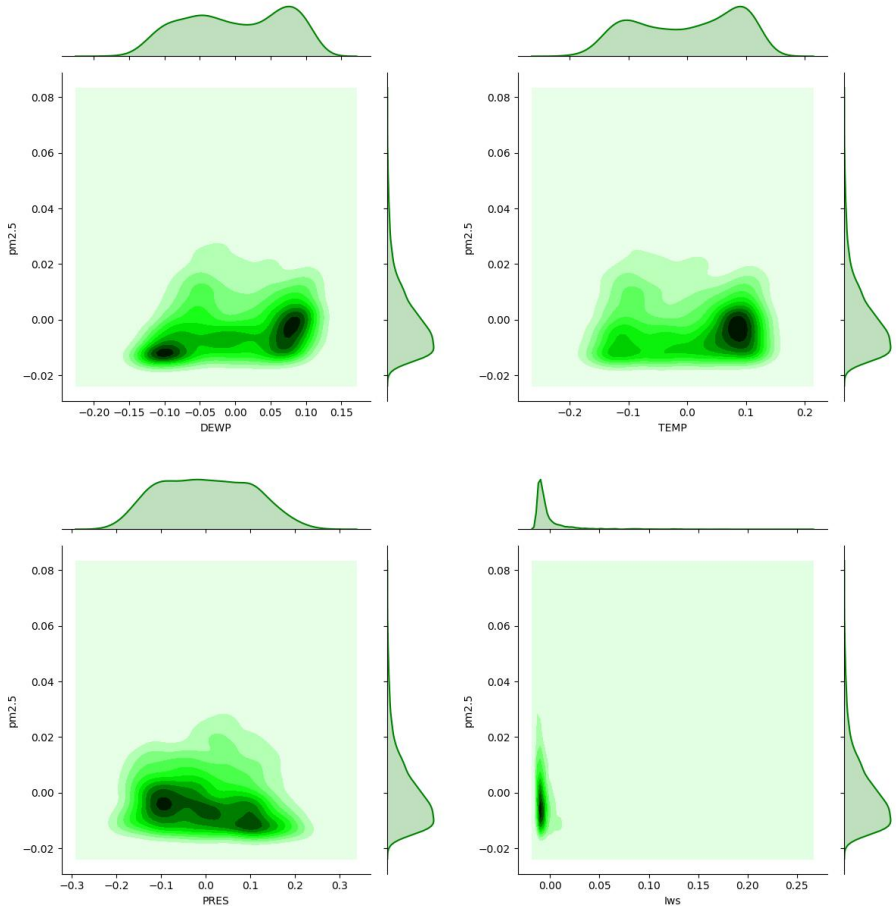


图 10

观察图 10，我们可以看出 DEWP 在-0.10 和 0.07 处颜色最深，TEMP 在 0.1 处颜色最深，PRES 在-0.1~0.1 颜色均略深，而 Iws 在-0.01 处颜色较深。据此也可以推断出，经过标准化处理的数据对分布的提取有了一定的帮助。

综合分析图 9,10 可得，标准化后的数据是以 0 为均值，方差为 1 的分布，消除了由于原始数据数量级等因素造成的分布“削弱”现象。由于分布的统一性，这为我们后来对数据进行进一步处理提供了合理性。但考虑到不同特征之间可能存在一定的相关性，这导致我们假设不同特征之间相互独立有一定出入，从而对模型的训练也带来一定的影响，因此我们对不同特征之间的相关关系进行处理。也就是以下的主成分分析法。

2，主成分分析（PCA）

主成分分析法（PCA）利用正交变换把由线性相关变量表示的观测数据转换为少数几个由线性无关变量表示的数据，线性无关的变量称为主成分。

①，主成分分析的基本思想：

统计分析中，数据的变量之间可能存在相关性，以致增加了分析的难度。于是，考虑由少数不相关的变量来代替相关的变量，用来表示数据，并且要求能够保留数据中的大部分信息。主成分分析中，首先对给定数据进行规范化，使得数据每一变量的平均值为 0，方差为 1。之后对数据进行正交变换，原来由线性相关变量表示的数据，通过正交变换变成由若干个线性无关的新变量表示的数据。新变量是可能的正交变换中变量的方差的和(信息保存)最大的，方差表示在新变量上信息的大小。将新变量依次称为第一主成分、第二主成分等。

以下我们给出样本主成分的定义以及主成分分析的几点性质：

定义 1.（样本主成分）给定样本矩阵 X 。样本的第一主成分 $y_1 = a_1{}^T x$ 是在 $a_1^T a_1 = 1$ 条件下，使得 $a_1^T x_j (j = 1,2,...,n)$ 的样本方差 $a_1^T S a_1$ 最大的 x 的线性变换；样本第二主成分 $y_2 = a_2^T x$ 是在 $a_2^T a_2 = 1$ 和 $a_2^T x_j$ 与 $a_1^T x_j (j = 1,2,...,n)$ 的样本协方差 $a_1^T S a_2 = 0$ 条件下，使得

$a_2^T x_j (j = 1,2,...,n)$ 的样本方差 $a_2^T S a_2$ 最大的 x 的线性变换；一般地，样本第 i 主成分 $y_i = a_i^T x$ 是在 $a_i^T a_i = 1$ 和 $a_i^T x_j$ 与 $a_k^T x_j (k < j, j = 1,2,...,n)$ 的样本协方差 $a_k^T S a_i = 0$ 条件下，使得 $a_i^T x_j (j = 1,2,...,n)$ 的样本方差 $a_i^T S a_i$ 最大的 x 的线性变换。

主要性质：

定理 1. 设 x 是 m 维随机变量, Σ 是 x 的协方差矩阵, Σ 的特征值分别是 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$，特征值对应的单位特征向量分别是 $\alpha_1, \alpha_2, \cdots, \alpha_m$, 则 x 的第 k 主成分是：

$$y_k = \alpha_k^T x = \alpha_{1k} x_1 + \alpha_{2k} x_2 + \cdots + \alpha_{mk} x_m, \quad k = 1,2,...,m$$

x 的第 k 主成分的方差是：

$$\mathrm{var}(y_k) = \alpha_k^T \Sigma \alpha_k = \lambda_k \quad k = 1,2,\cdots,m$$

即协方差矩阵Σ的第 k 个特征值。

定理 2. 对任意正整数 q，$1 \le q \le m$，考虑正交线性变换 $y=B^T x$，其中 y 是 q 维向量，B^T 是 $q \times m$ 矩阵，令 y 的协方差矩阵为 $\Sigma_y = B^T \Sigma B$，则 Σ_y 的迹 $tr(\Sigma_y)$ 在 $B=A_q$ 时取得最大值，其中矩阵 A_q 由正交矩阵 A 的前 q 列组成。

在使用样本主成分时，一般假设样本数据是规范化的。主成分分析的主要目的是降维，所以一般选择 $k(k<m)$ 个主成分(线性无关变量)来代替 m 个原有变量(线性相关变量)，使问题得以简化，并能保留原有变量的大部分信息，而由以上定理可以保证选择 k 个主成分是最优选择。

②，主成分分析应用于模型优化

为了最大程度地保留原数据的信息，调用 PCA 函数对未进行相关性分析选取应用与 SVM 模型前的数据（含 6 个影响因子(DEWP，TEMP，PRES，Iws，Is，Ir)和 PM2.5 的数据） 进行处理，得到正交变换后未筛选特征的所有成分各自重要性百分比（从大到小排序）：

[0.47544879 0.16634176 0.16556205 0.14814249 0.03265048 0.01185442]；

由于所得数据数据量较小，人工选取主成分个数时，即使只是去掉最后一个重要性不足 1.2%的成分，模型验证结果的准确度也大打折扣。据此，我们保留下 6 个影响因子正交变化后的所有特征作为选取的主成分，即放弃了 PCA 的降维优势，主要利用其对数据的前半部分处理--通过特征分解将原有数据投射至高维空间得到新的无相关关系的主成分数据，并在此基础上进一步进行参数优化。

③，主成分分析前后的可视化对比结果：

我们从数据集中的特征挑选几组进行 PCA 前后特征之间的绘图：

对比因素	PCA 前	对比因素	PCA 后

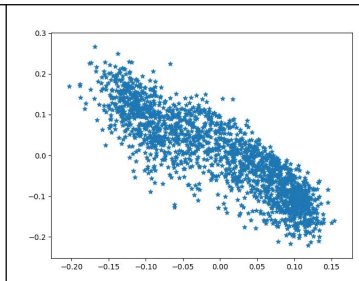
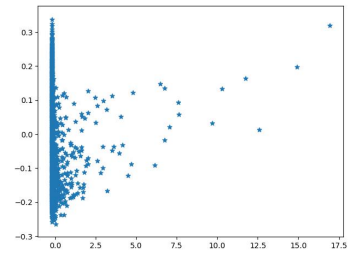
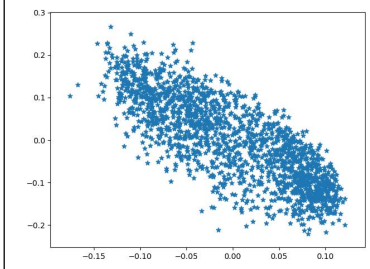
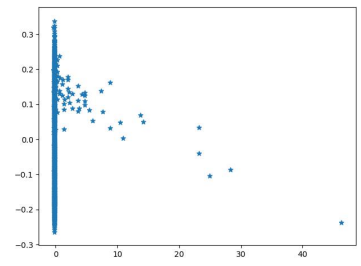
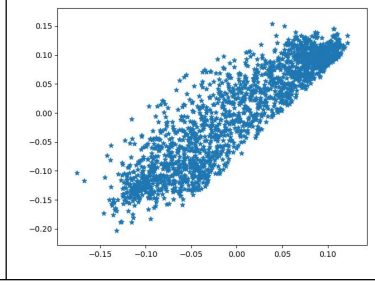
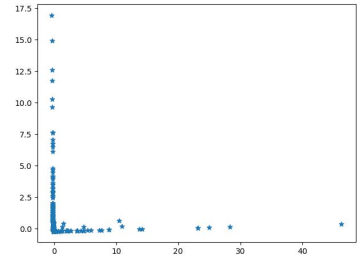
(TEMP, PRES)		(主成分 1，主成分 2)	
(DEWP, PRES)		(主成分 3，主成分 2)	
(DEWP, TEMP)		(主成分 3，主成分 1)	

表 4

由表 4 对比可知，在对标准化数据未作 PCA 处理之前，散点图中的点围绕在近似椭圆的区域内，各个特征之间存在一定的关系，并不是相互独立的。观察 PCA 处理后的图像，各个主成分之间的相关性削弱，这对后期 pm2.5 的预测有了一定的帮助。

同时，由图 11（PCA 处理后生成的 6 个主成分的热力图），可清晰的观察出 PCA 生成的 6 个主成分间的组合系数均近似为 0 ，故可基本认为 PCA 生成的 6 个主特征无相关关系。

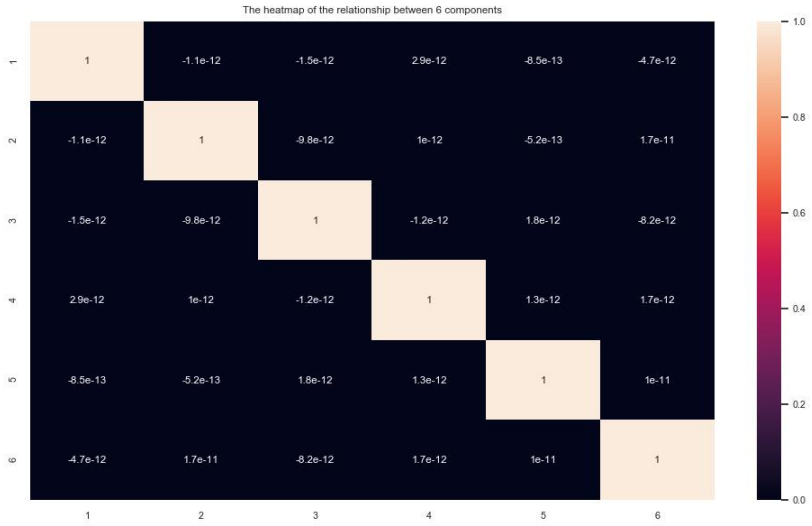


图 11

（二）FEMC-SVM 最优参数选择：Grid search

1，主要思想

grid search 是一种暴风搜索法，它首先为不同的超参数设定一个值列表，然后计算机会遍历每个超参数的组合进行性能评估，选出性能最佳的参数组合。用于超参数优化，通过优化超参数之间的最优组合来改善模型性能。

2，具体实现

从 sklearn 库中导入 GridsearchCv.首先,我们的参数设置如下:parameters:[{'kernel': ['rbf',' poly' ,' linear'], 'gamma': [0.001, 0.01, 0.1, 1, 10, 100], 'C': [1, 10, 100, 1000]}。

程序运行后，我们发现在程序运行的前期准确率较高，而越往后，准确率越低。于是我们将参数组合做如下修改：核函数为 rbf，合适的 gamma 参数选择为 0.001、0.01、0.1、1，合适的 C 为 1 或 10，继续实验。

第二次实验结果显示，准确率有所波动，但基本趋势是先有所上升（达到最高准确率约 0.55）而后下降，我们将参数组合继续细分：核函数仍为 rbf，gamma 参数为 1 到 3 以 0.05 为间隔的所有小数，C 为 1、2、3、4、5。以模型准确率为评判准侧，进一步缩小参数组合范围：核函数为 rbf，gamma 参数为 1.8 到 2.2 以 0.01 为间隔的小数，惩罚因子 C 为 1、2、3。

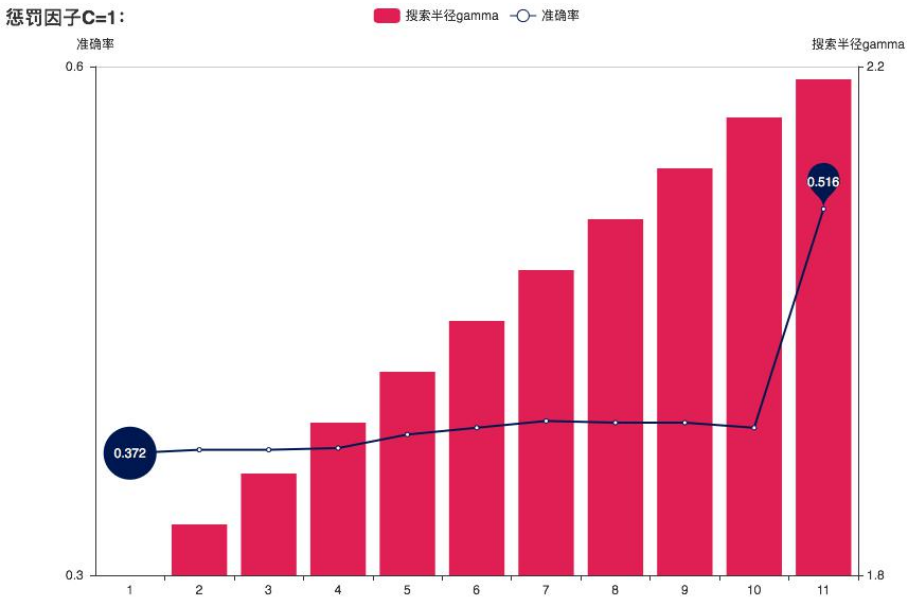


图 12

如图 12 所示，当惩罚因子 C=1 时，随搜索半径 gamma 在（1.8,2.2）间逐渐增大，准确率先平稳波动，在 gamma 接近 2.2 时由 0.372 迅速上升至最高准确率 r_1=0.516，但 r_1 尚未达到初步预调的最高准确度 0.55,说明 C=1 不是最优的惩罚因子选择；

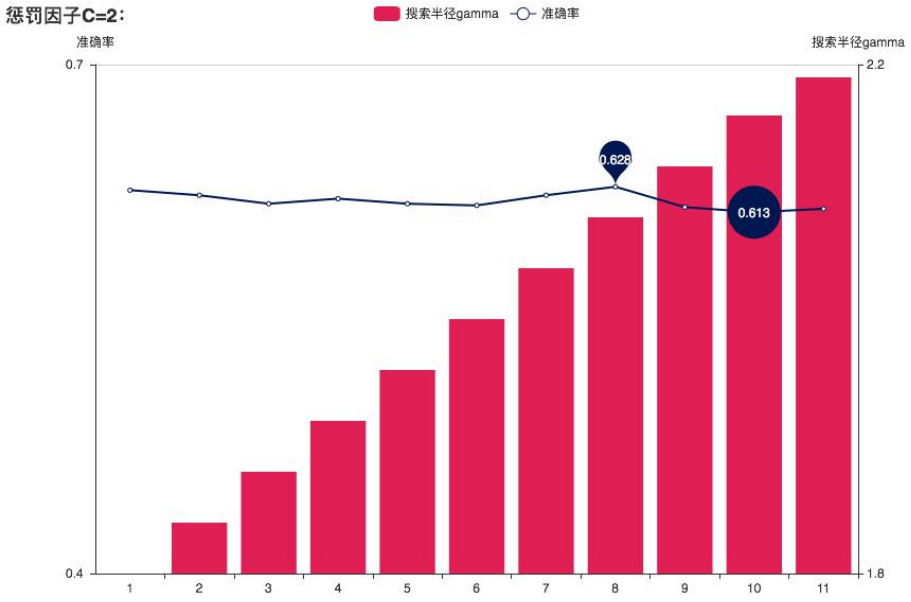


图 13

同理，如图 13 所示，当惩罚因子 C=2，$gamma \in (1.8,2,2)$ 时，准确率随搜索半径的变化程度不大，缓慢上升至最高点 r_2=0.628 后逐渐下降后稍有回升但仍低于 r_2；而模型分类的准确率在该参数范围下始终大于 0.613，故 C=2 为惩罚因子最优值，C=2,gamma=2.08 目前最优的参数组合；

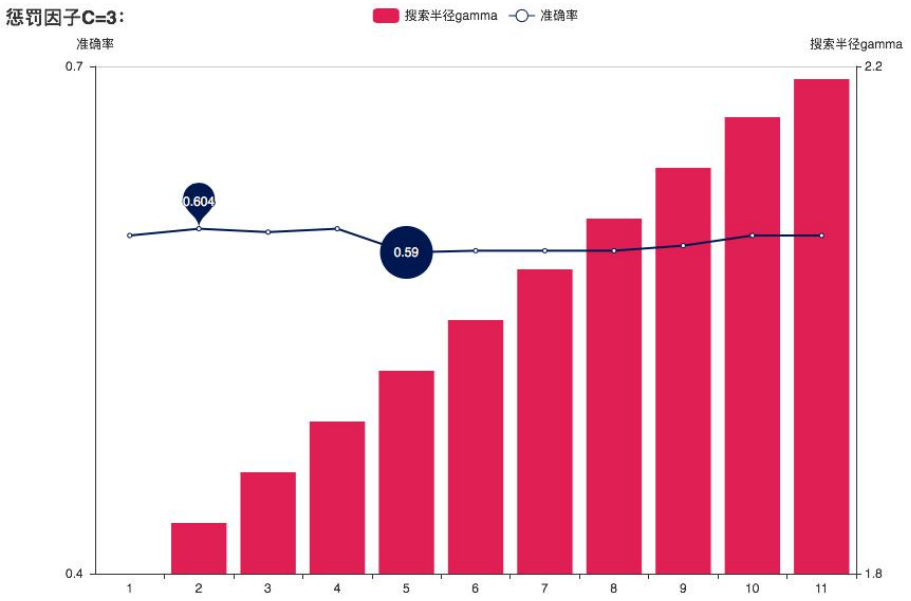


图 14

如图 14 所示，当惩罚因子 C=3 时，分类准确率最高仅为 0.604，而继续增大惩罚因子 C 的数值，准确率明显降低；

故根据上述重复操作，根据当前最高准确率出现的位置不断缩小参数组合范围--最终，所得模型最高准确率为 62.8%，对应的核函数为 rbf，gamma 参数为 2.08，惩罚因子 C=2。

最后，本文利用交叉检验的方法对以上得到的模型进行稳定性检验。首先，将数据集分成十份，轮流将其中 9 份作为训练数据，1 份作为测试数据，进行试验，并计算出每次试验的准确率，结果如图 15。随后又将数据集依次分成 2~10 份，轮流计算出准确率的平均值并绘制图像，如图 16。

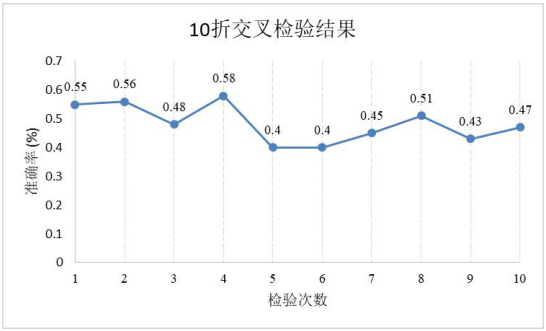


图 15

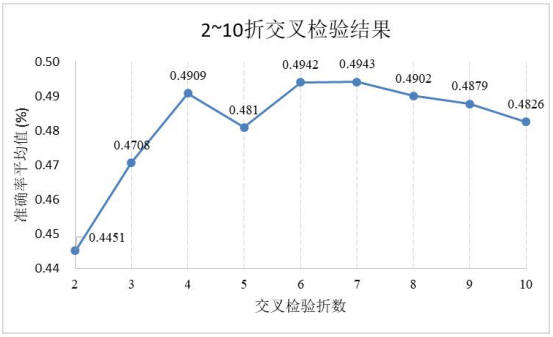


图 16

由图 15,16 可得，所建立的 FEMC-SVM 的模型准确率较为稳定。

（三）优化后模型评价及对比分析

在机器学习领域还有诸多可用来做分类的模型，我们在不同分类器之间以及PCA 降维前后分别对他们进行测试对比，并选择常用的 MSE（均方误差）以及 MAE（平均绝对误差）为指标，结果如下表 5：

常见分类器/指标	MSE_1	MAE_1	MSE_2	MAE_2
KNN	1.11	0.74	0.83	0.62
随机森林	1.06	0.72	0.86	0.62
决策树	1.31	0.83	1.02	0.70
Logistic 回归	**0.93**	**0.65**	1.13	0.78
朴素贝叶斯	2.84	1.31	2.14	1.08
支持向量机	1.40	0.91	**0.81**	**0.61**

表 5

同时，以横坐标表示进行分类的总次数，纵坐标统计分类过程中累积的错误次数绘制各分类器和 FEMC-SVM 模型应用于特征工程处理之后的数据的错误率对比变化曲线图如下：

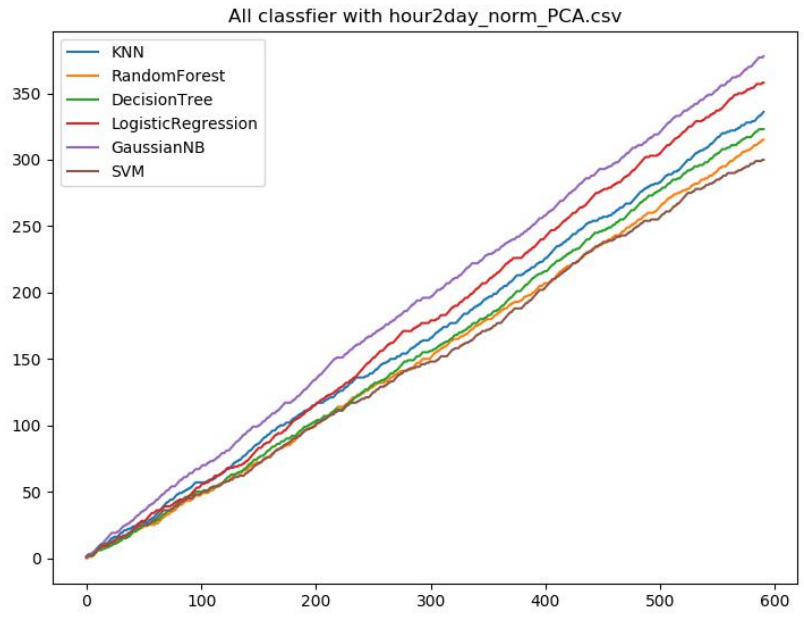


图 17

由表 5 我们可以看出，对于原数据，KNN、随机森林、决策树、逻辑斯蒂回归以及支持向量机效果较好，而传统的朴素贝叶斯对此数据集分类效果较差。而由图 17，就统计的 600 次分类结果来看，FEMC-SVM 模型对应的曲线斜率最低，显著优于 Gaussian NB（先验为高斯分布的朴素贝叶斯），KNN 等算法，随着分类次数的增加，SVM 分类效果较 Random Forest(随机森林)稳定的优良性也逐渐显现出来。

另外，对比数据集标准化及 PCA 降维前后，我们发现对原数据集进行 PCA 降维后，MSE，MAE 均有不同幅度的降低，分类器的分类效果普遍有大幅度提升，

其中 SVM 的结果提升显著且分类效果最好。由此可见，对数据特征的有效提取有利于提高训练结果。

七，FEMC-SVM 模型应用预测

（一）多维数据联合分布分析

由于各影响因子间有复杂的相关关系不便处理，为了更好地观察各影响因子数据下 PM2.5 等级的分布，通过 Python 绘制了不同 PM2.5 等级下各影响因子间的 6 维联合分布图（图 18）：

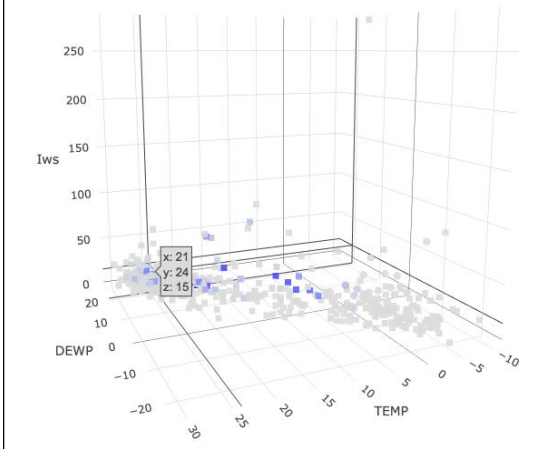
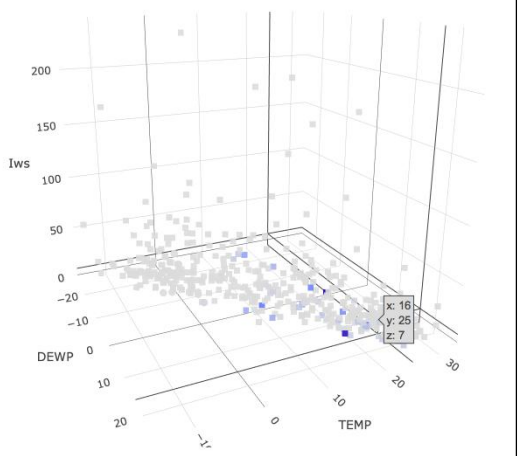
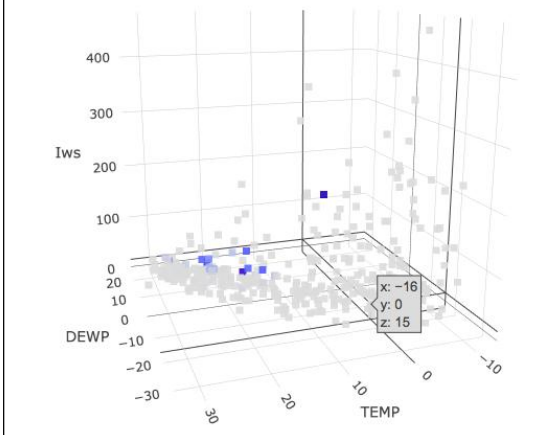
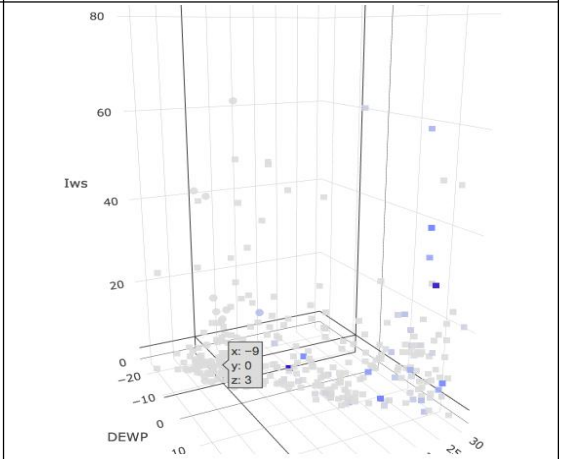
<table>
<tr><td>PM_Label:0,“优”</td><td>PM_Label:1,“良”</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>PM_Label:2,“中”</td><td>PM_Label:3,“差”</td></tr>
<tr><td> 
</td><td> 
</td></tr>
</table>

图 18

其中，x 轴代表 DEWP，y 轴代表 TEMP，z 轴代表 Iws，点的大小代表 PRES，点的形状（圆形、方形）代表 Is，点的颜色（深浅）代表 Ir；

各图中被标记具体数值的点，为该 PM2.5 等级下密集点群处的代表，能够表示该级别 PM2.5 等级下所出现天气状况的典型特征；

故可根据不同 PM2.5 的等级下影响因子的联合分布分别获取各自典型天气状况的影响因子范围，从而建立不同情景下的参数定量范围，并经人工排列分割参数范围获取情景分析所用影响因子验证集--data_Scenarios Analysis.csv。

（二）情景分析法

虽然通过优化后的 FEMC-SVM 模型预测 PM2.5 的等级达到了较高的准确率（62.8%），但由于通过 UCI 网站数据所得的气象因素间关联复杂，仅以个例预测 PM2.5 的等级本身具有极大的不确定性且现实意义不大。而情景分析法[16]是假定某种现象或某种趋势将持续到未来的前提下，对预测对象可能出现的情况或引起的后果作出预测的方法，对从现实角度定量描述不同气象因子数据对 PM2.5 等级的影响具有较大优势。

所以，根据我们选择用情景分析法，将不同情况下人们对天气的直观感受定量描述并输入 FEMC-SVM 模型中得到预测结果，并据此对人们的出行提出合理化的现实建议。

1，情景描述

我们基于人们对常见气象因素（e.g.温度，湿度）的直接感受设定了 4 种情景，情景的具体描述如下：

情景一：“零方案”：此情景中，空气清新，气候寒冷干燥，风速较高；

情景二：“一方案”：此情景中，空气质量较好，污染较少，风速适宜，气温较高；

情景三：“二方案”：此情景中，空气中存在一定污染，降水多，气候炎热湿润，风速较大；

情景四：“三方案”：此情景中，空气中含大量污染物，降水极少，气候极度干燥，刮风天气较少，温度较低，；

2，情景参数定量

基于多维数据联合分布分析中，不同 PM2.5 等级下各影响因子间的 6 维联合分布图，找到密集点群处的代表（如图 15，已标出），划定该点附近的密集点区域，记录下该区域各因子大小的范围，取其为表中各方案下气象因子数据的不同范围。

方案/因子	DEWP	TEMP	Iws	Is	Ir	PRES
方案一	<=-15	<=0	>=10	0	1.5	1200
方案二	10~20	>=20	5~10	0	1	1000
方案三	>=20	>=20	>=15	0	0.5	800
方案四	<=-5	<=0	0~5	0	0	600

表 6

3，FEMC-SVM 模型预测结果

根据情景参数定量表的信息，我们根据各方案下气象因子数据的不同范围取一定间隔建立了的数据集 data_Scenarios_Analysis.csv；

然后将该数据集放入 FEMC-SVM 模型中，取同一方案的各具体参数值的 pm2.5 预测等级结果众数为该方案的预测等级，并记下预测等级数占该方案下所有分类结果总数的百分比，具体结果如下表 7 所示：

方案	百分比
方案一	0.50
方案二	0.60
方案三	0.70
方案四	0.57

表 7

由于选取情景参数时无法获取有用的数据，且各影响因子在不同等级下的分布复杂，可能存在挑选出的典型天气状况（e.g.高温高湿）同时出现在多个方案中的情况，故通过 FEMC-SVM 模型所得的同一方案下的 PM2.5 等级存在差异，但表 7 中预测等级的百分比均不低于 0.5，故可认为所取预测结果可信度较高。

八，评价展望

（一）结论及建议

1，PM2.5 成因分析总结

根据用 SPSS 和 Python 对 UCI 网站下载的气象因子数据和 PM2.5 的相关性分析可得：获取的气象因子中，Iws（累计风速）,DEWP（露点）对 PM2.5 浓度的影响力分别位列第一，二位，且前者为负相关，后者为正相关；又因为 DEWP 为空气中水气含达到饱和的气温，是表征湿度的物理量，查阅资料（附录六）可得，DEWP 与当天空气中的相对湿度正相关，故认为相对湿度是 PM2.5 的第二大影响因素，且与 PM2.5 浓度正相关；

2，基于模型应用预测结果的出行建议

（1）当气候干燥寒冷，风速较高时，适宜参加户外活动呼吸清新空气，且呼吸高质量的空气有助于放松心情、给予人良好的与自然接触的体验；

（2）当气温较高，湿度较大，风速适宜时，能够参加正常的户外活动，正常的呼吸不会影响人体健康；

（3） 当气候炎热湿润，风速较大时，中等的空气质量不利于敏感人群进行户外活动，应尽可能避免出门活动；

（4）当气候寒冷干燥，风速较低时，空气质量比较差，空气中的污染物对人体健康有较大威胁，所有人都应当减少室外活动，不得已外出一定要做好相应防护措施。

（二）模型评价反思

1，数据来源限制

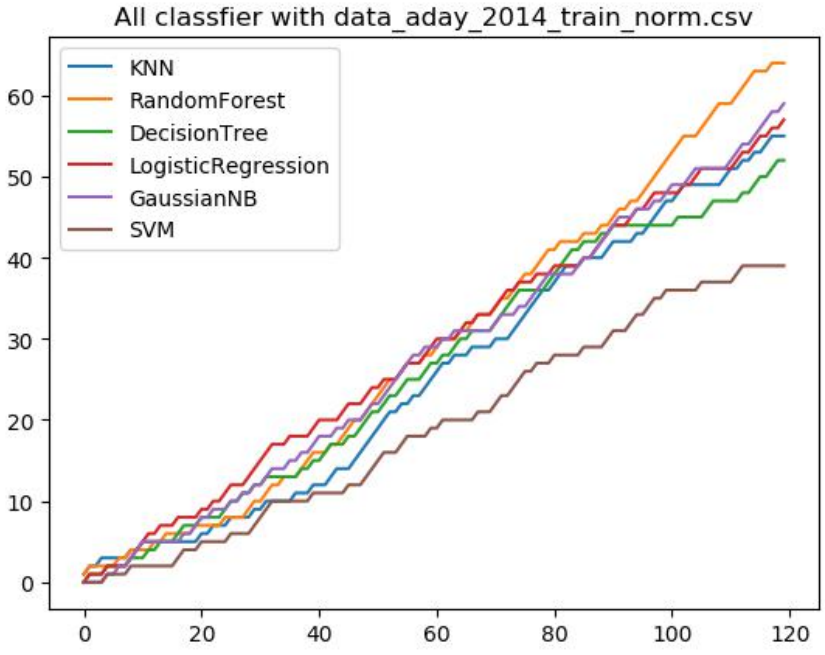


图 19

将模型的数据训练集改为经过 data_aday_2014_train_norm.csv--将爬取的数据和 UCI 重叠部分（2014 年） 重新整合成一个数据集标准化的 2014 年气象数据，所得各个模型的错误次数随其分类总次数的变化如图 19；相较图 17，图 19 中更加明显地展示了 SVM 优于其他分类模型的正确率--对尚未经过 PCA 处理的数据错误率低于 30%且曲线走势越来越平缓。这也再一次验证了“数据和特征决定了机器学习的上限”，若我们能将更加完整精确的相关气象数据代入所建模型，相信会提高模型的精确度，取得更好的预测效果。

2，模型评价

针对 PM2.5 预测分析的数据所建立的 FEMC-SVM 模型是实现小，中样本数据多分类的较优选择，相比其他模型，它综合了特征工程充分挖掘数据信息的优势和基于核技巧的非线性 SVM 实现高维数据分类的优良性，具有良好的推广应用能力；

当然，Grid Search 遍历实现参数最优化的方法对数据较大的样本太过耗时，可结合仿生学习算法（例如人工鱼群算法）按一定路径搜索出最优的参数组合。

（三）可进一步提升的方向

1，PM2.5 影响因子扩充

由于雾霾是特定气候条件与人类活动相互作用的结果，除了此次建模所考虑的气象因素外，影响雾霾的因素还有：化石燃料燃烧和尾气排放；城市“热岛效应”；风向；地理位置；季节时间等，故仅仅从 6 个气象影响因子的角度预测 PM2.5 的浓度显然是不够精确的，例如参考[10],运用文献分析方法归纳总结出京津冀地区雾霾发生的原因主要在于自然和社会经济原因，且两者相较，后者所起的作用更主要更深层；

但由于社会经济条件的变化终究会反映为自然气象条件的变化，故根据某天的易感气象数据预测当天的 PM2.5 指标仍是具有理论意义的；另外，通过对时间序列数据的分析研究，根据前一天或前几天的各影响因子数据和 PM2.5 浓度（或等级），较为准确地预测未来时间的 PM2.5 也是不错的研究方向。

2，模型改进

①，Kpca 的使用

由于用于建模的数据之间的关系未知，我们利用 PCA，即通过线性映射将输入空间投射到高维空间实现对数据主成分的分析，得到的可能不是最优的主成分结果，查阅资料提出可能的改进：使用 KPCA，即非线性主成分分析方法，能够更加客观、全面地找到数据间的特征。

②， 神经网络的使用

由于影响 Pm2.5 浓度的影响因子众多，针对同一因子又有海量数据可供选择以挖掘，而对于量大的数据集而言，神经网络模型无疑是最佳选择。在数据量大的前提下，选择合适的网络模型，结合各优化算法调参，根据误差反向传播算法原理优化参数，得到准确率相对较高的 pm2.5 浓度与各影响因子关系的模型，是此模型进一步优化的方向之一。

参考文献

引文文献

[1] 王江洪,贾盼盼,王晓琼.使用灰色理论模型对郑州雾霾情况进行预测[J].江西化工, 2018(3)

[2] 周银明. 基于灰色理论和支持向量机的 PM2.5 浓度预测[D]. 浙江农林大学.
[3] 龚明 叶春明. 基于修正灰色马尔科夫链的上海市 PM2.5 浓度预测[J]. 自然灾害学报(25):104.
[4] 熊萍萍[1,2,3,4], 李军[1], 张倩[1],等. 基于核与灰半径序列的 GM（1,N）预测模型及其在雾霾中的应用[J]. 山西大学学报（自然科学版）(2):280.
[5] 韩丽洁. 基于灰色关联分析和最小二乘支持向量机的光伏功率预测算法的研究[D]. 天津大学, 2014.
[6] 侯琼煌,杨航.基于三次指数平滑模型的雾霾天气分析与预测[J].环境保护科学, 2014(06):77-81
[7] 李慧敏. 基于三次指数平滑模型的雾霾天气预测[J]. 中国环境管理干部学院学报(3).
[8] 方天舒. 陕西雾霾天气预测[J]. 合作经济与科技, 2016(11):184-185.
[9] 石明珠. 基于时空混合模型的 PM2.5 浓度预测[D]. 燕山大学.
[10] 谢心庆. 乌鲁木齐市 PM_(2.5)浓度的动态分析[D]. 新疆财经大学.
[11] 黄伟政.基于卷积神经网络的雾霾时空演化预测方法研究[D]
[12] 宋利红. 基于深度学习的雾霾预测方法研究[D]
[13] 张平华, 盛凯. 基于神经网络的雾霾预测模型实证研究[J]. 菏泽学院学报, 2018, 40(05):24-28.
[14] 赵智. 基于遗传算法优化 BP 神经网络雾霾预测模型的研究[J]. 科技展望, 2015(27).
[15] 郑国威, 王腾军, 杨友森, et al. 基于遗传小波神经网络的 PM2.5 浓度预测模型[J]. 测绘与空间地理信息, 2018(9):248-250,256.
[16] 曾忠禄,张冬梅.不确定环境下解读未来的方法:情景分析法[J].情报杂志, 2005(05):15-17

阅读型文献

[17] 李航.统计学习方法.北京:清华大学出版社, 2012 年
[18] 刘顺祥.从零开始学 Python--数据分析与挖掘.北京:清华大学出版社, 2018 年
[19] 杜二玲,卢秀丽,窦林立.支持向量机在雾霾天气预测中的应用[J].内蒙古科技与经济, 2017(17):57-58
[20] 康春婷,李卫东.大数据技术在雾霾治理中的应用[J].中国经贸导刊, 2016(32):28-30
[21] 李洋.情景分析法在美国预见情报中的运用分析[J].信息记录材料, 2017 年 18(2):196
[22] 李晓燕.京津冀地区雾霾影响因素实证分析[J].生态经济,2016,v.32;No.303(03):146-152
[23] 沈聪.基于情景分析法的盆地城市 NO2 排放预测研究[J].魅力中国,2017(31)
[24] 周勇,周念彤.基于大数据的北京雾霾成因分析与 2017 年 PM2.5 浓度预测[J].技术与创新管理,2017,38(06):573-581

[25] 曾忠禄,张冬梅.不确定环境下解读未来的方法:情景分析法[J].情报杂志,2005(05):15-17
[26] Qin S ,Liu F ,Wang J etal.Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models[J]. Atmospheric Environment, 2014, 98:665-675
[27] Sun W ,Sun J.Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm[J].Journal of Environmental Management,2016,188:144-152

附录

（一）SPSS 相关系数计算结果

1，Pearson 相关系数计算结果如下：

相关性

		pm2.5	DEWP
pm2.5	皮尔逊相关性	1	.149**
	显著性（双尾）		.000
	个案数	1788	1788
DEWP	皮尔逊相关性	.149**	1
	显著性（双尾）	.000	
	个案数	1788	1788

**. 在 0.01 级别（双尾），相关性显著。

相关性

		pm2.5	TEMP
pm2.5	皮尔逊相关性	1	-.085**
	显著性（双尾）		.000
	个案数	1788	1788
TEMP	皮尔逊相关性	-.085**	1
	显著性（双尾）	.000	
	个案数	1788	1788

**. 在 0.01 级别（双尾），相关性显著。

相关性

		pm2.5	PRES
pm2.5	皮尔逊相关性	1	-.027
	显著性（双尾）		.246
	个案数	1788	1788
PRES	皮尔逊相关性	-.027	1
	显著性（双尾）	.246	
	个案数	1788	1788

相关性

		pm2.5	Iws
pm2.5	皮尔逊相关性	1	-.284**
	显著性（双尾）		.000
	个案数	1788	1788
Iws	皮尔逊相关性	-.284**	1
	显著性（双尾）	.000	
	个案数	1788	1788

**. 在 0.01 级别（双尾），相关性显著。

相关性

		pm2.5	ls
pm2.5	皮尔逊相关性	1	.030
	显著性（双尾）		.201
	个案数	1788	1788
ls	皮尔逊相关性	.030	1
	显著性（双尾）	.201	
	个案数	1788	1788

相关性

		pm2.5	lr
pm2.5	皮尔逊相关性	1	-.049*
	显著性（双尾）		.038
	个案数	1788	1788
lr	皮尔逊相关性	-.049*	1
	显著性（双尾）	.038	
	个案数	1788	1788

*. 在 0.05 级别（双尾），相关性显著。

3，Kendall 相关系数和 Spearman 秩相关系数计算结果截图如下：

相关性

			pm2.5	DEWP
肯德尔 tau_b	pm2.5	相关系数	1.000	.169**
		显著性（双尾）	.	.000
		个案数	1788	1788
	DEWP	相关系数	.169**	1.000
		显著性（双尾）	.000	.
		个案数	1788	1788
斯皮尔曼 Rho	pm2.5	相关系数	1.000	.252**
		显著性（双尾）	.	.000
		个案数	1788	1788
	DEWP	相关系数	.252**	1.000
		显著性（双尾）	.000	.
		个案数	1788	1788

**. 在 0.01 级别（双尾），相关性显著。

相关性

			pm2.5	TEMP
肯德尔 tau_b	pm2.5	相关系数	1.000	.008
		显著性（双尾）	.	.633
		个案数	1788	1788
	TEMP	相关系数	.008	1.000
		显著性（双尾）	.633	.
		个案数	1788	1788
斯皮尔曼 Rho	pm2.5	相关系数	1.000	.010
		显著性（双尾）	.	.663
		个案数	1788	1788
	TEMP	相关系数	.010	1.000
		显著性（双尾）	.663	.
		个案数	1788	1788

相关性

			pm2.5	PRES
肯德尔 tau_b	pm2.5	相关系数	1.000	-.072**
		显著性（双尾）	.	.000
		个案数	1788	1788
	PRES	相关系数	-.072**	1.000
		显著性（双尾）	.000	.
		个案数	1788	1788
斯皮尔曼 Rho	pm2.5	相关系数	1.000	-.112**
		显著性（双尾）	.	.000
		个案数	1788	1788
	PRES	相关系数	-.112**	1.000
		显著性（双尾）	.000	.
		个案数	1788	1788

**. 在 0.01 级别（双尾），相关性显著。

相关性

			pm2.5	Iws
肯德尔 tau_b	pm2.5	相关系数	1.000	-.291**
		显著性（双尾）	.	.000
		个案数	1788	1788
	Iws	相关系数	-.291**	1.000
		显著性（双尾）	.000	.
		个案数	1788	1788
斯皮尔曼 Rho	pm2.5	相关系数	1.000	-.420**
		显著性（双尾）	.	.000
		个案数	1788	1788
	Iws	相关系数	-.420**	1.000
		显著性（双尾）	.000	.
		个案数	1788	1788

**. 在 0.01 级别（双尾），相关性显著。

相关性

			pm2.5	ls
肯德尔 tau_b	pm2.5	相关系数	1.000	.102**
		显著性（双尾）	.	.000
		个案数	1788	1788
	ls	相关系数	.102**	1.000
		显著性（双尾）	.000	.
		个案数	1788	1788
斯皮尔曼 Rho	pm2.5	相关系数	1.000	.126**
		显著性（双尾）	.	.000
		个案数	1788	1788
	ls	相关系数	.126**	1.000
		显著性（双尾）	.000	.
		个案数	1788	1788

**. 在 0.01 级别（双尾），相关性显著。

相关性

			pm2.5	lr
肯德尔 tau_b	pm2.5	相关系数	1.000	.058**
		显著性（双尾）	.	.002
		个案数	1788	1788
	lr	相关系数	.058**	1.000
		显著性（双尾）	.002	.
		个案数	1788	1788
斯皮尔曼 Rho	pm2.5	相关系数	1.000	.075**
		显著性（双尾）	.	.001
		个案数	1788	1788
	lr	相关系数	.075**	1.000
		显著性（双尾）	.001	.
		个案数	1788	1788

**. 在 0.01 级别（双尾），相关性显著。

（三）模型初步训练结果

选用 gamma=1，2,3,4 进行实验，比较不同核函数的最佳准确率：

核函数	Gamma	准确率
Linear	1	0.227
Linear	2	0.227
Linear	3	0.227
Linear	4	0.227
Poly	1	0.288
Poly	2	0.413
Poly	3	0.444
Poly	4	0.442
rbf	1	0.387
rbf	2	0.387
rbf	3	0.526
rbf	4	0.532

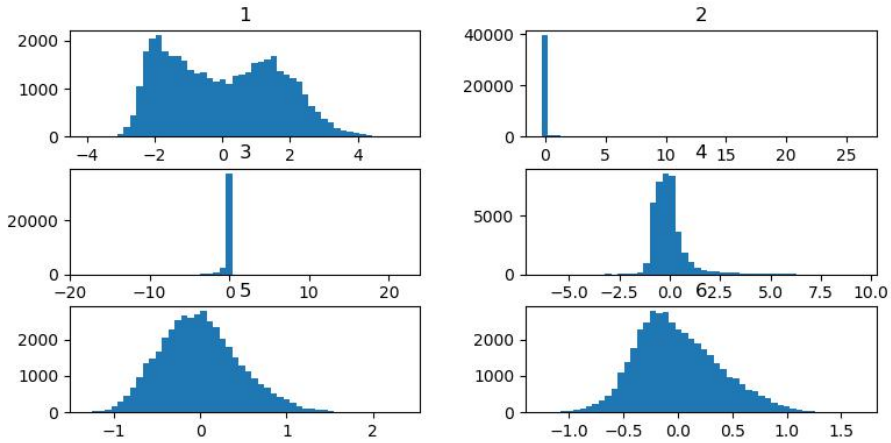
（三）数据集主成分分析计算结果

主成分分析所得数据集：feature_xg1.csv

各个成分方差百分比	7.68542622e-01	2.28465532e-01	2.73590859e-03	1.67505591e-04	8.84313109e-05
	-6.62372739e-03	2.84923614e-04	9.99906642e-01	-1.17292398e-02	-2.27518000e-03

具有最大方差的成分	7.73945920e-03	-1.05324419e-04	1.18080280e-02	9.99821968e-01	1.25176583e-02
	-7.80997859e-01	6.19627443e-02	-6.44174810e-03	1.39063290e-02	-6.21263294e-01
	4.65694114e-01	7.20717569e-01	1.74490942e-03	2.87943421e-03	-5.13501551e-01
	-4.16013899e-01	6.90453990e-01	-1.65411720e-03	-4.09626937e-03	5.91765332e-01

（四）PCA 处理生成的主成分分布图



（五）Grid Search 优化过程

Gamma	C	准确率
1	1.8	0.372
1	1.84	0.374
1	1.88	0.374
1	1.92	0.375
1	1.96	0.383
1	2.00	0.387
1	2.04	0.391
1	2.08	0.390
1	2.12	0.390
1	2.16	0.387
1	2.19	0.516
2	1.8	0.626
2	1.84	0.623
2	1.88	0.618
2	1.92	0.621
2	1.96	0.618

2	2.00	0.617
2	2.04	0.623
2	2.08	0.628
2	2.12	0.616
2	2.16	0.613
2	2.19	0.615
3	1.8	0.6
3	1.84	0.604
3	1.88	0.602
3	1.92	0.604
3	1.96	0.590
3	2.00	0.591
3	2.04	0.591
3	2.08	0.591
3	2.12	0.594
3	2.16	0.600
3	2.19	0.600

（六）模型应用数据处理

由于 UCI 数据集中 DEWP 的数据无法直接通过爬虫获得，经查阅资料，DEWP 露点：空气中水气含达到饱和的气温(a, f),表征湿度；

经上网查阅，露点（DEWP）与相对湿度对应关系表如下：

相对湿度	露点	相对湿度	露点	相对湿度	露点	相对湿度	露点
0.1%	-51.75	4.0%	-17.84	2.1%	-24.49	15.0%	-3.02
0.2%	-46.08	4.1%	-17.58	2.2%	-24.02	16.0%	-2.25
0.3%	-42.62	4.2%	-17.33	2.3%	-23.57	17.0%	-1.15
0.4%	-40.11	4.3%	-17.07	2.4%	-23.14	18.0%	-0.83
0.5%	-38.12	4.4%	-16.83	2.5%	-22.73	19.0%	-0.15
0.6%	-36.47	4.5%	-16.59	2.6%	-22.33	20.0%	0.50
0.7%	-35.06	4.6%	-16.35	2.7%	-21.94	30.0%	6.24
0.8%	-33.82	4.7%	-16.12	2.8%	-21.57	40.0%	10.48
0.9%	-32.72	4.8%	-15.90	2.9%	-21.20	50.0%	13.86
1.0%	-31.73	4.9%	-15.67	3.0%	-20.85	60.0%	16.70
1.1%	-30.82	5.0%	-15.46	3.1%	-20.51	70.0%	19.15
1.2%	-29.99	6.0%	-13.47	3.2%	-20.18	80.0%	21.31
1.3%	-29.22	7.0%	-11.77	3.3%	-19.86	90.0%	23.24
1.4%	-28.50	8.0%	-10.28	3.4%	-19.55		
1.5%	-27.82	9.0%	-8.95	3.5%	-19.25		
1.6%	-27.19	10.0%	-7.75	3.6%	-18.95		
1.7%	-26.59	11.0%	-6.65	3.7%	-18.67		

1.8%	-26.03	12.0%	-5.64	3.8%	-18.39		
1.9%	-25.49	13.0%	-4.71	3.9%	-18.11		
2.0%	-24.98	14.0%	-3.83				

致谢

在此对在建立，优化模型全程认真给予意见，提供参考资料，指引模型应用方向的赵鲁涛老师和学校负责此次建模比赛，耐心为我们答疑解惑的刘秀芹老师和徐伟老师表示深深的感谢。