

文章编号: 1674-8085 (2022) 06-0071-05

基于数据挖掘的天气因素研究

——以吉安县为例

吴玉春,*曾 寰,李金忠,杨 治,刘 华

(井冈山大学电子与信息工程学院, 江西, 吉安 343009)

摘 要: 影响天气的因素有降雨量、气压、气温、风速及风向等, 研究天气影响因素对人们的生产生活有一定的指导作用。本研究提出使用数据挖掘技术探究天气因素之间的相关性。使用 K-S 检验方法对数据进行正态性检验, 使用 Pearson 相关系数, 对天气逐日数据降雨量、气压、气温、风速进行相关性检验, 使用 Eta 相关系数对天气逐日数据降雨量、气压、气温、风速与风向进行相关性检验, 使用一元线性回归对天气因素气压和气温进行分析。在中国气象网站的吉安县 1980-2019 年天气数据上进行实验, 实验结果表明: 数据集中各数据项不服从正态性分布, 天气当中的气压与气温具有强负关联线性关系, 风向类型与气压、气温有强关联关系。

关键词: 天气分析; 数据挖掘; 线性回归; Pearson 相关系数; Eta 相关系数

中图分类号: TP311.13

文献标识码: A

DOI: 10.3969/j.issn.1674-8085.2022.06.011

RESEARCH ON WEATHER FACTORS BASED ON DATA MINING – TAKING JI'AN AS AN EXAMPLE

WU Yu-chun, *ZENG Huan, LI Jin-zhong, YANG Zhi, LIU Hua

(School of Electronics and Information Engineering, Jinggangshan University, Ji'an, Jiangxi 343009, China)

Abstract: This paper proposed to use data mining technology to explore the correlation among weather factors. Pearson correlation coefficient and Eta correlation coefficient tests were performed on daily data of rainfall, air pressure, air temperature, wind speed and wind direction. The linear equation was used to fit the air pressure and air temperature with high correlation. Experiments were carried out on the weather data of Ji'an city from China meteorological website. The experimental results showed that Pearson correlation coefficient analysis suggested air pressure was strongly negative correlated with air temperature, but Eta correlation coefficient analysis suggested wind direction type is strongly correlated with air pressure and air temperature.

Key words: data mining; weather analysis; linear regression; Pearson correlation coefficient; Eta correlation coefficient

数据挖掘定义为从数据集合当中自动抽取隐藏在数据中的有用信息的非平凡过程, 这些信息的表现形式为: 规则、概念、规律和模式^[1]。决策者可以通过数据挖掘技术来分析特定类型的历史数

据和当前数据, 并从中发掘关联关系, 预测未来可能发生行为。随着气象信息化程度的提高, 气象部门积累了海量的气象数据, 据有关文献显示, 气象部门每天业务的应用的数据高达 PB 数量级^[1]。如

收稿日期: 2022-05-19; 修改日期: 2022-06-24

基金项目: 国家自然科学基金项目(62141203); 江西省教育厅科技计划项目(GJJ180574); 吉安市指导性科技计划项目(吉市科计字 2021[8]号基础 11); 井冈山大学科研基金项目(JZ2004)

作者简介: 吴玉春(1975-), 女, 江西吉安人, 高级实验师, 硕士, 主要从事数据挖掘研究(E-mail: 1322391726@qq.com);

*曾 寰(1990-), 男, 江西吉安人, 实验师, 硕士, 主要从事数据挖掘研究(E-mail: 584251395@qq.com)。

何利用好这些数据用以提高预报准确性和灾害天气预警能力成为重要关键问题。

目前,应用于气象研究领域的数据挖掘技术主要包含两个方面:采用数理统计方法如统计分析、相关分析方法和回归分析在内的统计学模型;采用机器学习及数值分析软计算方法应用于空间数据研究和气象模式建模等^[2]。

文献[3]采用了 C4.5 分类方法、k-means 聚类方法以及关联规 Apriori 算法用于挖掘 PM2.5 与能见度、风速、风向、露点和湿度的相关关系。文献[4]采用 GIS 空间分析技术及 Mann-Kendall 检验应用于华东区域暴雨的时空特征进行研究得到暴雨的发生表现出较强的空间差异性。文献[5]使用了数理统计方法用于评估干旱检测模型构建当中,得出采用 SVM 方法用于构建的模型对高程表现出较高的敏感性。文献[6]选取地表温度、大气压强等 5 个影响降水的气象要素,分别采用 C4.5、随机森林对降水进行预测。文献[7]采用 k-means 对经过一定规则提取的序列化特征进行符号化,并最终得到降雨天气预测模型。文献[8]将 KNN 算法应用霾等级的预报应用当中。文献[9]利用神经网络、朴素贝叶斯、随机森林和 K-近邻算法构建天气预报预测模型,并在 Kaggle 网站的天气数据对各模型算法进行训练和测试得出随机深林算法取得最佳性能。文献[10]运用数据挖掘技术中数据预处理技术来对从加拿大政府网站上的爬取天气的天气数据和海洋表面温度数据进行筛选、无关数据剔除、数据归一化等来完善数据,接着使用 SPSS 软件进行时间序列分析,以及使用三维散点图和直观图得到温度在地域上的变化趋势;同时建立连续复小波变换模型来对海洋表面温度进行分析,得到海洋表面温度存在周期性波动变化。综合以上文献,可以看出,天气数据的分析常采用数理统计分析、相关性分析等统计学模型,其中数理统计分析方法主要包含:散点图、直方图等频率分析技术;相关性分析用于对数据进行筛选,可用于对数据进行降维,去除无关数据的影响等。若需对天气中某个因变量随自变量的变化则采用回归分析、机器学习等技术来对收集到的天

气数据进行建模分析。

本文从中国气象网站上获取了吉安市吉安县 1980 年至 2019 年的天气数据,利用数据挖掘技术中的相关性分析和一元线性回归方法对该天气数据进行分析与实验并得出了一些相关结论。

1 研究方法

1.1 相关性系数

对不同数据类型之间的变量采用不同相关性分析方法,数据类型可以分为三类:标度型(Scale)、有序型(Ordinal)和类别型(Nominal),它们之间的定义分别为:标度型数据为连续性的数值如身高、体重、温度等;有序型的数据为具有高低次序的阶段性离散数值如年龄、学历等;类别型的数据为没有高低次序和大小区分如性别、民族等。

相关性分析方法一般采用四种:卡方检验, Spearman 系数、Pearson 系数、Eta 系数。检验不同类型的变量的相关性方法如表 1 所示。

表 1 不同类型变量的相关性方法

Table 1 Correlation methods for different types of variables

	类别型	有序型	标度型
类别型	卡方检验	卡方检验	Eta 相关系数
有序型		Spearman 相关系数	Spearman 相关系数
标度型			Pearson 相关系数

由于本研究涉及的数据类型为类别型数据和标度型数据,因此,接下来将只介绍 Eta 相关系数和 Pearson 相关系数。

Pearson 相关系数适用于两个连续变量的线性相关性描述,其计算方法为公式(1)。

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \cdot \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}} \quad (1)$$

Pearson 相关系数仅仅是变量之间线性关系的一个度量,它不能用于描述非线性关系。当 $|r| \geq 0.8$ 时,可视为高度相关; $0.5 \leq |r| \leq 0.8$ 时,可视为中度相关; $0.3 \leq |r| \leq 0.5$ 时,视为低度相关;当 $|r| \leq 0.3$,说明两个变量之间的相关程度极弱,可视为不相关。

当求取类别型变量与连续型(标度型)变量的关联强度时,可利用 η (Eta)系数。其计算方法如公式(2)所示。

$$Eta = \sqrt{\frac{\sum (x - \bar{x})^2 - \sum (x - \bar{x}_k)^2}{\sum (x - \bar{x})^2}} = \sqrt{1 - \frac{\sum (x - \bar{x}_k)^2}{\sum (x - \bar{x})^2}} \quad (2)$$

其中 $\bar{x}$ 为连续型变量的平均数, $\bar{x}_k$ 为类别型各组下的连续变量的平均数。Eta系数值介于0到1之间。当Eta平方小于0.06时,表明两个变量之间相关程度微弱,可视为不相关;当Eta平方大于0.06小于0.16时,表明两个变量之间有中度的相关关系;当Eta平方大于0.16时,表明变量之间存在强相关。

1.2 一元线性回归

回归分析当中,线性回归可以用来定量的分析两种或两种以上变量间的相互依赖关系,当只包括一个自变量和一个因变量且二者的关系可用一条直线拟合时,称为一元线性回归分析。其线性方程为公式(3)。

$$y = ax + b \quad (3)$$

其中 x 为自变量, y 为因变量。拟合求解时采用最小二乘法,其方程为公式(4),(5)。

$$a = \frac{\overline{xy} - n\bar{x}\bar{y}}{\overline{x^2} - n(\bar{x})^2} \quad (4)$$

$$b = \bar{y} - a\bar{x} \quad (5)$$

2 结果

2.1 数据源

文中分析的数据集来自中国气象网,收集到的为吉安市吉安县1980年1月1日至2019年12月31日逐日天气数据,数据条数为14609条,数据格式为表2。

表中2019年12月31日的数据表示为 $10237 \times 0.1\text{hPa} = 1023.7\text{hPa}$ 的平均气压, $98 \times 0.1^\circ\text{C} = 9.8^\circ\text{C}$ 的平均气温, $38 \times 0.1\text{mm} = 3.8\text{mm}$ 累计一天的降水量,平均风速为 $17 \times 0.1\text{m/s} = 1.7\text{m/s}$,最大风速风向的编码为2。表中各列的数据类型有标度型和类

别型,其中平均气压,平均气温,20-20时累计降水量(一天累计降雨量),平均风速为标度型,最大风速风向为类别型。

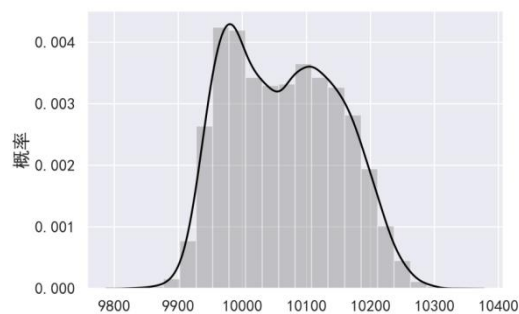
表2 天气数据的原始格式

Table 2 Raw format of weather data.

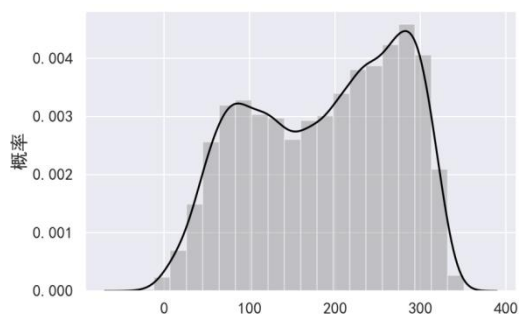
日期	平均本站 气压 0.1hPa	平均 气温 0.1℃	20-20时累 计降水量 0.1mm	平均 风速 0.1m/s	最大风 速风向
1980年01月01日	10085	86	0	30	2
⋮	⋮	⋮	⋮	⋮	⋮
2019年12月31日	10237	98	38	17	2

2.2 数据正态性分布检验

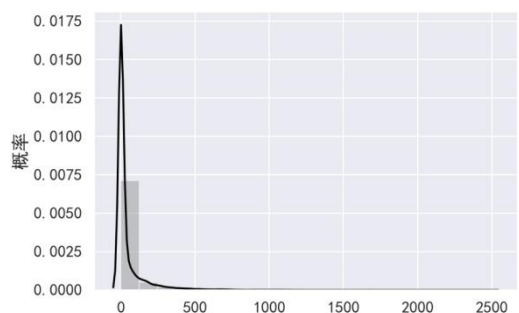
本研究采用概率密度直方图的形式对天气各特征数据进行直观显示,结果如图1所示:



(a) 平均本站气压



(b) 平均气温



(c) 20-20时累计降水量

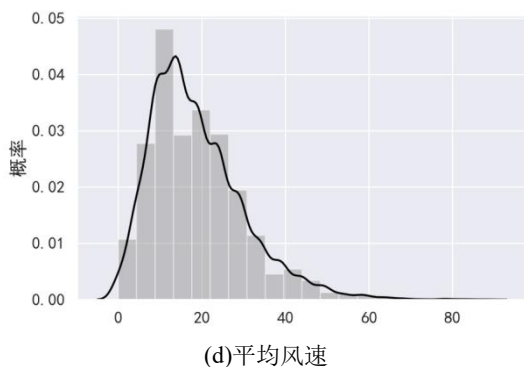


图1 1980-2019 吉安县不同天气因素数据值分布情况

Fig.1 Data distribution of different weather factors in Ji'an from 1980 to 2019

由图1可得出数据集中各特征数据正态分布不明显。因为样本量大小 $14609 > 5000$, 所以采用 Kolmogorov - Smirnov(K-S)检验方法对以上各列天气数据进行进一步正态分布检验, 各列数据项在显著性水平为 0.01 的情况下, 其结果如表3所示, K-S 检验结果表明各数据项均不符合正态分布。

表3 采用 K-S 正态性检验结果

Table 3 The results of k-S normality test.

变量名	中位数	平均值	标准差	偏度	峰度	K-S 检验
平均本站气压	10062	10064.79	85.959	0.177	-0.921	0.069(0.000***)
平均气温	199	188.584	87.051	-0.226	-1.143	0.077(0.000***)
20-20 时累计降水量	0	43.628	115.06	5.346	47.936	0.352(0.000***)
平均风速	17	18.709	11.909	-0.097	11.82	0.103(0.000***)

注: *代表 1% 的显著性水平

结合两种方法对数据进行正态性检验, 得出平均本站气压、平均气温、一天累计降水量、平均风速这些数据项均不符合正态分布的分布规律。如果样本数据正态分布, 则样本数据大概率可以代表总体数据, 样本数据若非正态分布, 则样本数据和总体数据之间关联弱。论文中获取的 1980-2019 年的天气样本数据不符合正态分布, 不能代替总体天气数据规律, 采用相关性分析和一元线性回归的分析结果不作为总体预测结果。

2.3 相关性分析

针对吉安市吉安县天气数据进行了相关性分

析, 其中平均本站气压、平均气温、20-20 时累计降水量, 平均风速各列由于是标度型数据, 因此可以采用 Pearson 相关系数对它们之间的线性相关性进行分析; 由最大风速风向属于类别型数据, 而其他各列数据类型为标量型数据, 因此, 最大风速风向列与其他列相关性分析采用 Eta 相关系数。结果如表4和表5所示。

表4 基于 Pearson 系数的相关性分析结果

Table 4 Correlation analysis results based on pearson coefficient

	平均本站气压	平均气温	20-20 时累计降水量	平均风速
平均本站气压	1	-0.889(0.000**)	-0.15(0.000**)	-0.129(0.000**)
平均气温	-	1	0.004(0.000**)	0.098(0.000**)
20-20 时累计降水量	-	-	1	0.064(0.000**)
平均风速	-	-	-	1

注: **在 0.01 级别 (双尾), 相关性显著。

表5 基于 Eta 系数的相关性分析结果

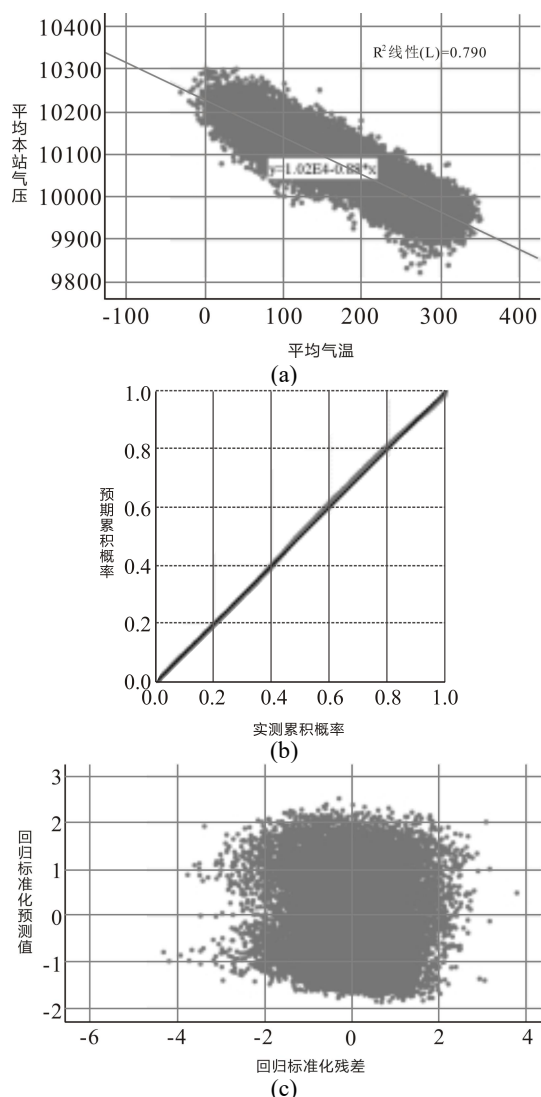
Table 5 Correlation analysis results based on Eta coefficient

	平均本站气压	平均气温	20-20 时累计降水量	平均风速
最大风速的风向	0.451	0.460	0.104	0.308

由表4和表5所展示的结果可以得出, 平均本站气压与平均气温具有较高的负线性相关性, 最大风速的风向分别与平均本站气压, 平均气温具有较强相关性。

2.4 一元线性回归分析

采用一元线性回归方法对平均本站气压与平均气温作进一步分析, 通过数据拟合得到如图3(a)所示线性方程 $y=1.02E4-0.88x$, 其中 x 为平均气温, y 为平均本站气压, 线性置信度 R^2 为 0.790, 表明具有较高的线性拟合度。从图3(b)所示的回归标准化残差的正态 P-P 图和图3(c)所示的回归标准化残差的散点图也可以得出其拟合性好。



(a)平均本站气压与平均气温线性回归图; (b)回归标准化残差的正态 P-P 图; (c)回归标准化残差的散点图

图2 平均本站气压与平均气温的线性回归图

Fig.2 Plot of linear regression between mean atmospheric pressure and mean temperature

2.5 极端天气分析

由于天气状况中暴雨天气危害极大,而在气象部门发布的天气预报中小雨、中雨、暴雨等术语,它们之间的区别是:小雨是指 24 h 内降水量不超过 10 mm 的雨,小到中雨为 5 ~ 16.9 mm,中雨为 10 ~ 24.9 mm,中到大雨为 17 ~ 37.9 mm,大雨为 25 ~ 49.9 mm,大到暴雨为 38 ~ 74.9 mm。24 h 内雨量超过 50 mm 的称为暴雨,超过 100 mm 的称为大暴雨,超过 250 mm 的称为特大暴雨。我们统计了吉安市吉安县降雨量 > 38 mm 的随月份变化的历史统计结果,结果如图 4 所示,吉安市吉安县

暴雨发生频率比较高的时段为 4-6 月份。

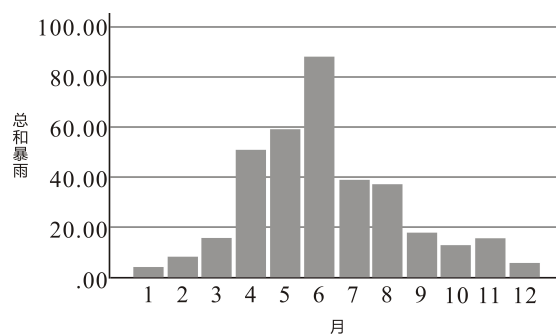


图3 1980-2019 吉安县暴雨天气随月份变化的历史统计

Fig.3 The history statistics of storm varied with month at Ji'an from 1980 to 2019

3 结论

本研究采用数据挖掘相关技术对吉安县 1980 年至 2019 年的天气的逐日数据进行分析,各特征项不服从正态分布,吉安县平均本站气压与平均气温具有较高的负线性相关性,最大风速的风向分别与平均本站气压、平均气温具有强相关性,吉安县暴雨发生频率比较高的时段为 4-6 月份。通过对实验结果的分析,有利于指导人们生产生活。

参考文献:

- [1] 彭霞云,裘薇,李文娟,等.数据挖掘技术用于降水相态判别的尝试[J].科技通报,2018,34(1):44-47.
- [2] 贾志明,王东峰,程智.数据挖掘技术在气象预报研究中的应用[J].黑龙江科学,2020,11(8):34-35.
- [3] 危蓉,徐伟.基于大数据的天气状况研究[J].信息通信,2015(12):122-123.
- [4] 姜宁.基于时空数据挖掘技术的华东区域暴雨时空特征[J].应用生态学报,2017,28(12):4043-4050.
- [5] 张婧娴,沈润平,郭佳.不同数据挖掘方法在综合干旱监测模型构建中的应用研究[J].江西农业大学学报,2017,39(5):1047-1056.
- [6] 张佳华,姚宜斌,曹娜.基于决策树对有无降水进行预测[J].测绘地理信息,2017,42(5):107-109.
- [7] 陈晓云,吴本昌,韩海涛.基于多维时间序列挖掘的降雨天气模型研究[J].计算机工程与设计,2010,31(4):898-902.
- [8] 熊亚军,廖晓农,李梓铭,等.KNN 数据挖掘算法在北京地区霾等级预报中的应用[J].气象,2015,41(1):98-104.
- [9] Fairoz Q Kareem,Adnan Mohsin Abdulazeez,Dathar A Hasan. Predicting weather forecasting state based on data mining classification algorithms[J]. Asian Journal of Research in Computer Science,2021.
- [10] 陈波,姬家昌,多俊杰.关于全球气候变化和极端天气数据的量化分析[J].科学技术创新,2020(24):56-57.