Advances in Applied Mathematics 应用数学进展, 2023, 12(4), 1971-1980
Published Online April 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.124202

Hans汉斯

基于贝叶斯网络的梅雨季节降水诊断分析

黄梦婷，杨卫华*

太原理工大学数学学院，山西 晋中

收稿日期：2023年3月26日；录用日期：2023年4月21日；发布日期：2023年4月29日

摘　要

江南地区由于地理气候因素，每年夏季都会进入梅雨季节，降水频繁。为了能够更加准确的对梅雨季节的降水情况进行分析，从而能够更好的预测梅雨季节可能到来的洪涝灾害，以防造成较大的经济损失，本文以南京市梅雨季节的降水状况为例，利用南京市2011~2022年每年6~7月的气象观测资料，选取发生降水天气的要素资料，基于贝叶斯网络模型和决策树模型，对南京市降水发生情况进行诊断分析。结果表明：贝叶斯网络模型准确率为83.61%，决策树模型的准确率为81.97%，可见，贝叶斯网络模型的准确率更高，效果也更好。

关键词

贝叶斯网络，决策树，降水量，诊断分析

Diagnostic Analysis of Precipitation in Plum Rain Season Based on Bayesian Network

Mengting Huang, Weihua Yang*

School of Mathematics, Taiyuan University of Technology, Jinzhong Shanxi

Received: Mar. 26th, 2023; accepted: Apr. 21st, 2023; published: Apr. 29th, 2023

Abstract

Due to geographical and climatic factors, the Jiangnan region will enter the plum rain season every summer, with frequent precipitation. In order to more accurately analyze the precipitation in the plum rain season, it can better predict the possible flood disaster in the plum rain season, so as to prevent greater economic losses, in this paper, the precipitation in Nanjing during the plum rain season was taken as an example, the meteorological observation data from June to July of 2011 to 2022 were used to select the element data of precipitation weather, and the diagnosis and analysis

*通讯作者。

文章引用: 黄梦婷, 杨卫华. 基于贝叶斯网络的梅雨季节降水诊断分析[J]. 应用数学进展, 2023, 12(4): 1971-1980.
DOI: 10.12677/aam.2023.124202

of the precipitation occurrence in Nanjing was carried out based on Bayesian network model and decision tree model. The results show that the accuracy of Bayesian network model is 83.61%, and that of decision tree model is 81.97%. It can be seen that the accuracy of Bayesian network model is higher and the effect is better.

Keywords

Bayesian Networks, Decision Trees, Precipitation, Diagnostic Analysis

1. 引言

江南地区人口稠密，经济发达，但也因为地理原因，受到副热带季风东亚热带季风的影响，每年夏季都会有一段持续时间的降水，又被称为梅雨季节，每年的梅雨季节都有可能会发生洪涝灾害。江苏省南京市就位于江南地区，梅雨季节集中在每年的 6~7 月。针对梅雨季节的降水情况，尤其是暴雨天气进行预测，以预测可能到来的洪涝灾害有着极为重要的经济价值和应用价值。

随着社会信息化的发展，目前国内外越来越多的研究选择使用计算机手段来进行有关天气要素特征的预测和分析。赵琳娜等利用贝叶斯模型平均方法，得到具有预测性的概率密度函数，然后采用 EM 算法，对淮河流域 2008 年 7 月 1 日~8 月 5 日的降水数据预报进行了集成与订正[1]。史达伟等选择了机器学习中的分类与回归树算法、支持向量机算法、线性支持向量机算法、类支持向量机算法与类神经网络算法对连云港地区 2014~2016 年间大雾天气背景下发生特强浓雾天气的气象要素建立诊断模型，并发现这几种算法中，线性支持向量机算法的测试效果最好，但相较而言，决策树模型更加直观准确，且具有较高的泛化能力，综合效果更优[2]。张远汀等利用决策树算法，根据 2017 年 12 月~2018 年 2 月和 2007 年 12 月~2008 年 2 月国家测站日值数据中的各个气象要素，生成了一个二元判别决策树模型，用以预测当天是否会有积雪，但也有不足之处，此模型对于极端积雪天气的预测误差较大[3]。阮成卿等对影响华北地区汛期降水的年际分量的预报因子的强度进行逐年分类，针对每个分类都进行相应预报模型设计，利用条件降尺度法和偏相关预报因子挑选法，对原有的降水时间尺度分离模型进行改进，建立条件降水时间尺度分离统计降尺度模型，显著提高了预测准确率[4]。

朴素贝叶斯算法是一种应用广泛的分类预测算法。Mccallum A 等使用多变量伯努利模型和多项式模型对文本进行分类，并发现多项式模型效果更好[5]。任民宏等利用朴素贝叶斯算法对大盘指数的涨跌进行预测。Rish I 使用蒙特卡罗模拟法，系统研究了几个分类问题，发现朴素贝叶斯分类算法的准确性与特征向量之间的互信息的依赖程度并不直接相关[6]。Kohavi R 提出了朴素贝叶斯和决策树算法的混合算法：朴素贝叶斯树(NBTree)算法[7]。Peng F 等提出了一个广义的朴素贝叶斯模型——增强型朴素贝叶斯分类器(CAN)，相较于原本的朴素贝叶斯模型，CAN 放松了条件独立性假设，还能够直接运用来自统计语言建模的复杂平滑技术[8]。张文钧等提出一种双层的贝叶斯模型，结合随机森林和朴素贝叶斯算法，对文本进行分类[9]。黄宇达等对朴素贝叶斯和决策树算法进行了改进，引入了客观属性重要度这一参数，弱化了朴素贝叶斯必须的条件独立性假设，选用加权信息熵作为分类标准，得到的算法模型能够在一定程度上克服决策树算法原有的多指偏向问题[10]。邸鹏等改进了经典朴素贝叶斯分类算法，选择引入一个放大系数，从而降低先验概率的影响，扩大后验概率的权重[11]。李冬梅等选用 UCI 数据集，采用 10 倍交

叉验证法，把概率优化函数代入至朴素贝叶斯中，利用朴素贝叶斯与决策树混合分类法，既避免朴素贝叶斯易下溢与过度拟合的问题，也降低了决策数算法过度拟合的可能，对冠心病医辅助诊疗系统起到了有效作用[12]。蒋良孝提出了基于支持向量机的朴素贝叶斯分类算法。任民宏等利用朴素贝叶斯算法对大盘指数的涨跌进行预测[13]。

但是，朴素贝叶斯模型有着一个极大的限制条件，它要求所有的特征属性必须相互独立，这一要求在现实生活中不仅几乎难以实现，有些特征属性之间还会有着较强的相关性。由此，贝叶斯网络模型这一算法进入了人们的视线。宫秀军等利用主动学习这一方法，通过修正分类参数，提出了一种主动贝叶斯分类模型[14]。吴立增等基于贝叶斯网络模型建立了变压器的 TAN 故障诊断模型和朴素贝叶斯网络故障诊断模型[15]。陈雪等基于贝叶斯网络分类算法实现了对遥感影像的变化的检测[16]。

因此，本文选择了更高等级，应用范围也更广的贝叶斯网络模型来对南京市梅雨季节的降水情况进行相关的诊断分析。

2. 贝叶斯分类模型

2.1. 贝叶斯网络

贝叶斯网络(Bayesian network)，又称为信度网(Belief Network, BN)或有向无环图模型，是一种概率图模型，一般由有向无环图和一个条件概率表组成，其中，节点表示随机变量$\{X_1, X_2, \cdots, X_n\}$，有向线段表示变量之间有因果关系，并由此产生条件概率值[17]。

如图 1 是一个简单的贝叶斯网络。

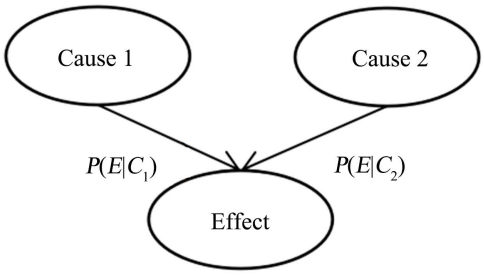

Figure 1. Simple Bayesian network diagram
图 1. 简单贝叶斯网络图

图中包含三个随机变量 Cause 1，Cause 2 和 Effect，很明显，节点 Effect 的取值或结果直接取决于 Cause 1 和 Cause 2，由此，在三个变量节点间会产生条件概率表，由 $P(E|C_1)$ 和 $P(E|C_2)$ 组成。

2.2. 贝叶斯网络分类原理

2.2.1. 贝叶斯相关理论

在了解贝叶斯网络分类原理之前，我们需要先了解一些贝叶斯相关理论。

条件概率公式：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \tag{1}$$

其中，$P(A|B)$是后验概率，即事件 B 发生的前提下，事件 A 发生的概率；$P(B|A)$是事件 A 发生的前提下，事件 B 发生的概率；$P(A)$是先验概率，即事件 A 发生的概率；$P(B)$是事件 B 发生的概率，$P(B)$不改变分类结果，是一个规范化因子，作用是获取后验概率的和等于 1。

贝叶斯公式：

$$P\left(A_i\middle|B\right)=\frac{P\left(B\middle|A_i\right)P\left(A_i\right)}{\sum_{j=1}^{n}P\left(B\middle|A_j\right)P\left(A_j\right)} \tag{2}$$

其中，$P\left(A_i\middle|B\right)$是后验概率，即事件 B 发生的前提下，事件 A_i 发生的概率；$P\left(A_i\right)$是先验概率，即事件 A_i 发生的概率，且对于所有的事件 A_i，有 $\sum_{i=1}^{n}P\left(A_i\right)=1$ 成立；$P\left(B\middle|A_i\right)$是条件概率。

对于一组离散随机变量$\left(X_1,X_2,\cdots,X_n\right)$，它们的取值分别为$\left\{x_1,x_2,\cdots,x_n\right\}$的联合概率为

$$\begin{aligned}P\left(x_1,x_2,\cdots,x_n\right)&=P\left(x_n\middle|x_{n-1},\cdots,x_1\right)P\left(x_{n-1}\middle|x_{n-2},\cdots,x_1\right)\cdots P\left(x_2\middle|x_1\right)P\left(x_1\right)\\&=\prod_{i=1}^{n}P\left(x_i\middle|x_{i-1},\cdots,x_1\right)\end{aligned} \tag{3}$$

而贝叶斯网络中，在给定了父节点的条件下，每个节点 X_i 都与其非后代节点条件独立，由于这一变量节点的局部独立性，上式可以化简为

$$P\left(x_1,x_2,\cdots,x_n\right)=\prod_{i=1}^{n}P\left(x_i\middle|Pa\left(x_i\right)\right) \tag{4}$$

式中 $Pa\left(x_i\right)\subseteq\left\{x_{i-1},\cdots,x_1\right\}$，取值通常是已知的，其含义是结点 y_i 的父结点的集合。

2.2.2. 贝叶斯网络分类步骤

1) 设随机变量$\left(X_1,X_2,\cdots,X_n\right)$，其取值分别为$\left(x_1,x_2,\cdots,x_n\right)$，这里的每个 x_i 都是随机变量 X 的一个特征属性，可以是离散数据，也可以是连续数据，又或者同时包含离散变量和连续变量。

2) 假定有 m 个类别 Y，记为$Y=\left\{y_1,y_2,\cdots,y_m\right\}$。

3) 所要得到的结果，就是给定一个未知分类的数据样本 X，在已知 X 的数据的情况下，预测该待分类项所须归属的类别，也就是，利用 $\max_{i=[1,\cdots,m]}\left\{p\left(y_i\middle|x\right)\right\}$ 来决定其所属类别，其中，

$$p\left(y_i\middle|x\right)=\frac{p\left(y_i\right)\times\prod_{j=1}^{n}p\left(x_j\middle|y_i;Pa\left(x_j\right)\right)}{p\left(x\right)} \tag{5}$$

3. 决策树模型

决策树又被称为判定树，是一种用于分类与回归的树形结构，一般可以认为是 if-then 规则的集合。决策树由结点和有向边组成，而结点则由内部结点和叶结点组成，其中，内部结点表示一种特征或者属性，而叶结点则表示一种类[18]。

3.1. 平均信息(熵)

信息：

$$I\left(x_i\right)=-\log_2 P\left(x_i\right) \tag{6}$$

平均信息(熵)：

$$H=-\sum_{i=1}^{N}P\left(x_i\right)\log_2 P\left(x_i\right) \tag{7}$$

一般来说，熵越大，事件的发生越无序；熵越小，事件的发生越确定。

3.2. 树的建立/划分规则

本质上来说，决策树的建立其实就是在寻找一种能够使得熵能够在最大程度熵变小的划分方案。一般来说，分为四个步骤：

1) 找到可以令平均熵最小的特征维度对数据集进行分割。
2) 对分割后的数据集再找寻可以使平均熵最小的特征维度，再对数据集进行分割。
3) 重复上面步骤直到用完所有特征、或者子集中目标标签全部相同。
4) 如果所有特征都用完，最终的子集中，目标标签仍不一致，则使用最多标签作为最终输出。

4. 仿真实验

4.1. 数据处理

本文利用贝叶斯网络模型和决策树模型对梅雨季节降水量进行研究，选取南京市 2011 年~2022 年每年 6 月 1 日~7 月 31 日的南京市机场 58,238 站逐日自动观测数据。

为了研究下雨天气背景下暴雨的气象观测要素的特征，文章选取地面以上 2 米处的当天最高温度(℃) (简称为最高温度)，地面以上 2 米处的当天最低温度(℃) (简称为最低温度)，气象站水平的大气压(mmHg) (简称为大气压)，地面高度 2 米处的相对湿度(%) (简称为相对湿度)，观测前 10 分钟内地面高度 10~12 米处的平均风速(m/s) (简称为平均风速)等观测要素。以降水量的多少作为分类类别，分为无降水天气与降水天气，其中，降水天气又分为普通降水天气(包括小雨，中雨，大雨，降水量为 0.1~49.9 mm)和暴雨天气(包括暴雨，大暴雨，特大暴雨，降水量在 50 mm 以上)。

实验中，选取南京市 2011~2021 年 6~7 月的气象数据作为训练数据，2022 年 6~7 月的气象数据作为测试数据。

4.2. 降水气象特征分析

本文以南京市 2011~2021 年每年 6~7 月的气象数据为研究背景，对南京市梅雨季节的降水以及降水量的多少的规律以及诊断模型进行分析。针对南京市梅雨季节降水天气的时间分布特征进行统计，对影响南京市 6~7 月降水天气的气象要素特征进行统计分析，也即研究南京市的降水天气的发生规律。

通过统计分析发现，如图 2 所示，降水天气累计发生频次 268 次，在每年的 6 月 1 日~7 月 31 日均有发生，其中 6 月 16 日~7 月 15 日发生频次较高，其余时间发生频次较低，发生频次最高时间段是 7 月 1 日~7 月 5 日，累计发生了 34 次降水天气，占总降水频次的 12.69%。未发生过特大暴雨。如图 3 所示，暴雨天气发生频次较为均匀，几乎每个时间段都有暴雨天气发生，且总频次也较低，只有 15 次，占全部降水天气的 5.60%。

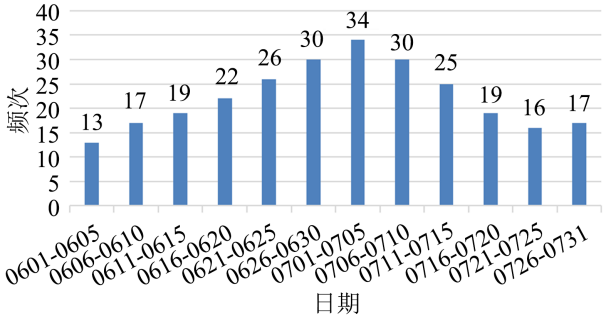

Figure 2. Time distribution diagram of rainy weather frequency
图 2. 降水天气频次的时间分布图

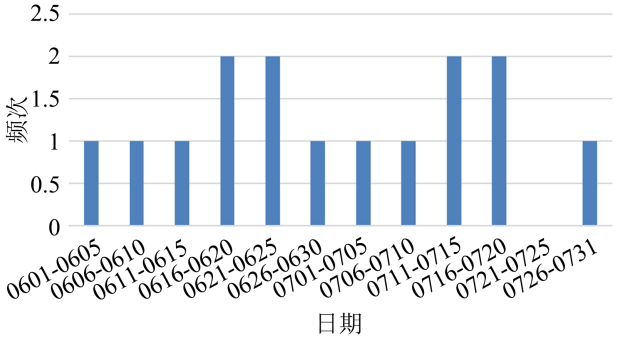

Figure 3. Time distribution diagram of rainstorm weather frequency
图 3. 暴雨天气频次的时间分布图

降水天气和暴雨天气发生时的大气压分布特征如图 4 与图 5 所示，降水天气的大气压在 738.7 mmHg~756.3 mmHg 区间内均有分布，主要分布在 748 mmHg~752 mmHg 这一区间，发生了 160 个时次，占比 59.7%。暴雨天气发生的大气压区间分布在 740.5 mmHg~755.2 mmHg，发生时次同样在 748 mmHg~752 mmHg 区间最为集中，占比 53.3%。

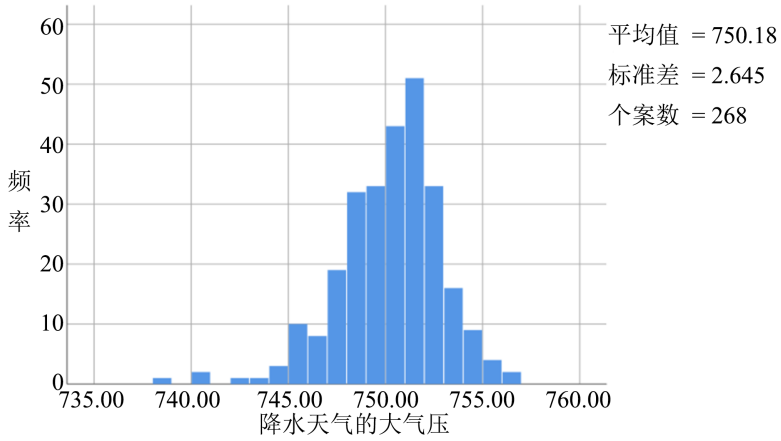

Figure 4. Graph of atmospheric pressure frequency in rainy weather
图 4. 降水天气的大气压频率图

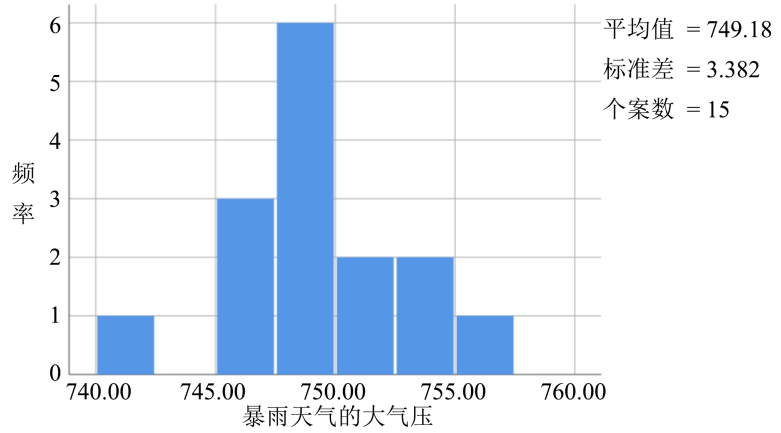

Figure 5. Graph of atmospheric pressure frequency in rainstorm weather
图 5. 暴雨天气的大气压频率图

降水天气和暴雨天气发生时的相对湿度分布特征如图 6 与图 7 所示，降水天气在相对湿度 62%以上均有分布，主要分布在 90%以上，发生了 198 频次，占比 73.9%。暴雨天气发生的相对湿度区间分布在 92%以上。

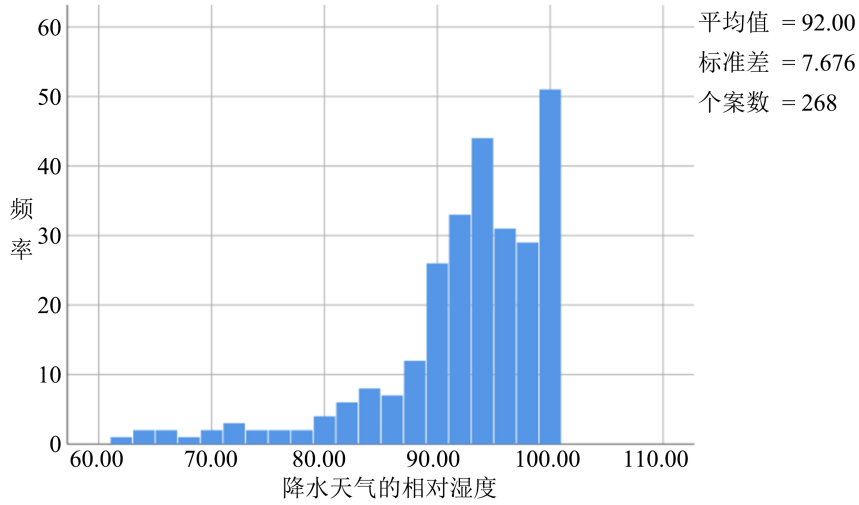

Figure 6. Graph of relative humidity frequency in rainy weather
图 6. 降水天气的相对湿度频率图

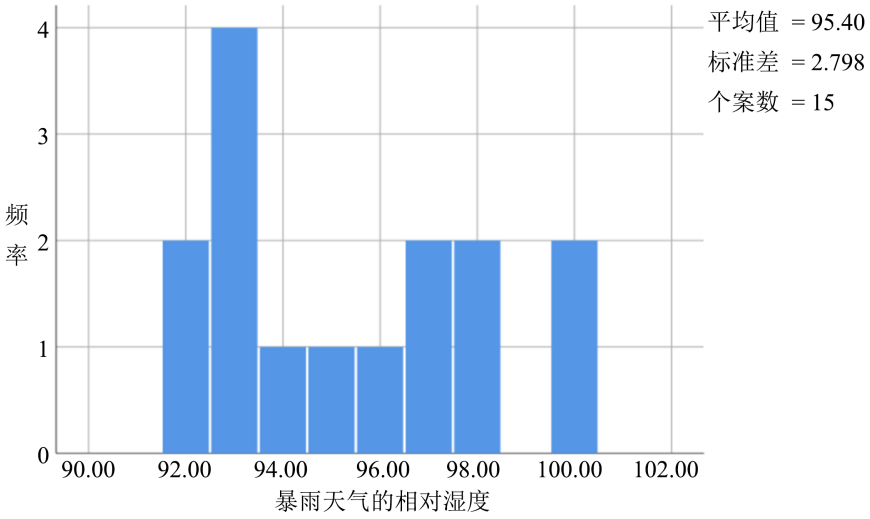

Figure 7. Graph of relative humidity frequency in rainstorm weather
图 7. 暴雨天气的相对湿度频率图

降水天气和暴雨天气发生时的平均风速分布特征如图 8 与图 9 所示，降水天气发生时，平均风速从 0 m/s~10 m/s 均有分布，平均风速在 1~4 时发生降水天气的时次占比 89.6%，其中平均风速为 2 m/s 时发生降水天气的时次最多，有 83 次，占比 31%。从图 9 可以看出，平均风速在 2 m/s~3 m/s 区间时，暴雨天气发生时次集中，且在平均风速为 2 m/s 时发生时次最多。

通过对降水天气和暴雨天气发生的气象特征要素进行统计分析，发现在发生降水天气时，是否会增强至暴雨天气需要具备以下几个条件才会更加有利：大气压在 748 mmHg~752 mmHg 区间内最为有利；相对湿度需在 92%以上；平均风速在 2 m/s~3 m/s 之间。

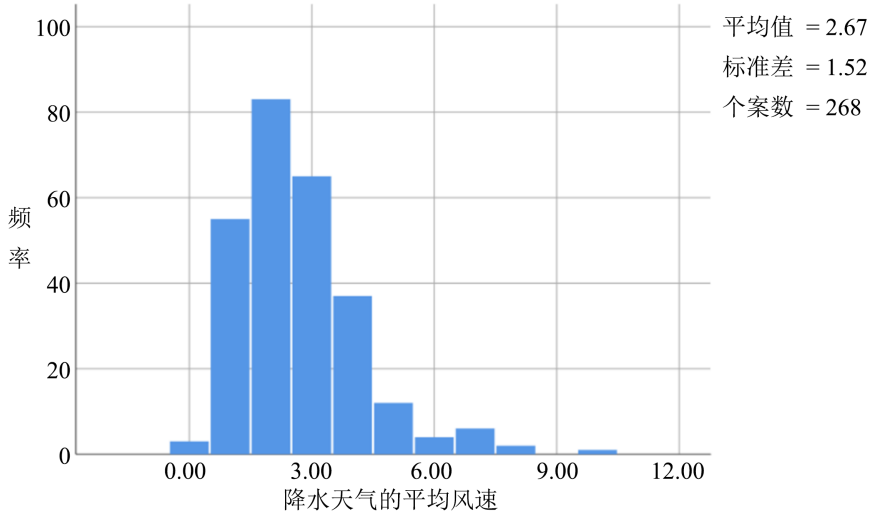

Figure 8. Graph of mean wind speed frequency in rainy weather
图 8. 降水天气的平均风速频率图

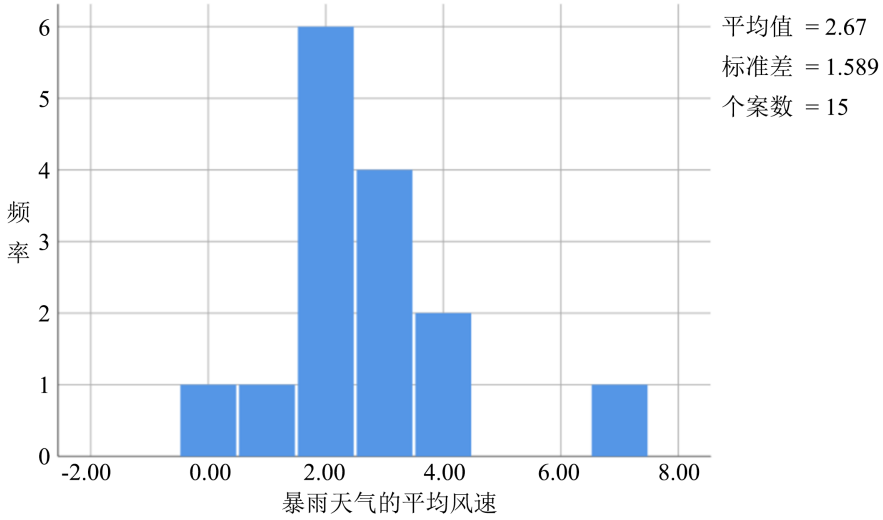

Figure 9. Graph of mean wind speed frequency in rainstorm weather
图 9. 暴雨天气的平均风速频率图

4.3. 实验结果与分析

通过上述分析，可以得到降水天气和暴雨天气发生的一些有利条件，但现实中，各种天气发生的条件与可能性是一个十分复杂的非线性过程，尤其虽然南京市 2011 年~2021 年 6~7 月间发生降水天气较多，但发生暴雨天气的频次并不算很高，具有较大的偶然性，想通过简单的定性条件就直接判定是否会发生暴雨天气是否发生并不容易。因此，本文会利用贝叶斯网络模型和决策树模型，对南京市的降水天气，尤其是暴雨天气是否发生建立诊断模型。

以是否为暴雨天气为模型的目标向量，模型的输入变量为最高温度，最低温度，大气压，相对湿度和平均风速。将预处理好的训练集和测试集数据输入模型，得到最终的结果。

建立一个贝叶斯网络，以是否降水为子节点，最高温度，最低温度，大气压，相对湿度和平均风速为父节点，得到的贝叶斯网络如图 10。

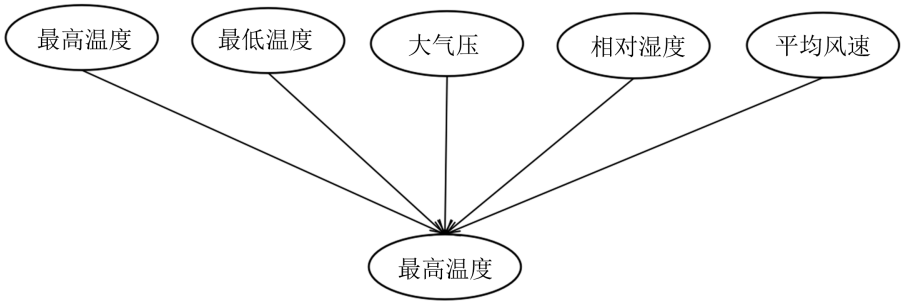

Figure 10. Bayesian network model diagram
图 10. 贝叶斯网络模型图

将训练数据代入程序中，训练贝叶斯网络模型和决策树模型，然后进行测试数据的实验，将南京市2022 年 6 月 1 日~7 月 31 日的降水数据与所得结果对比，最终发现，贝叶斯网络模型的准确率为 83.61%，而决策树模型的准确率为 81.97%。

贝叶斯网络模型与统计相结合，相较于其他的如朴素贝叶斯模型、决策树模型等模型，贝叶斯网络模型有着独特的优势。贝叶斯网络模型结合图论知识，利用图模型描述各个变量之间的关系，更清晰易懂，也便于理解；不像朴素贝叶斯模型那样需要所有的特征变量都保持完全独立，贝叶斯网络模型对特征变量间的独立关系要求没有那么严格，也更符合实际；贝叶斯网络模型包含了因果关系和概率性语义，可以用于学习数据间的因果关系；贝叶斯网络模型最大的优势是可以处理不完备的数据集，因为它反映的是数据间的概率关系模型，所以可以有一定的数据缺失。

5. 总结

本文针对南京市每年 6~7 月期间梅雨季节的降水状况以及暴雨天气进行了时间特征和天气要素特征的分析，并利用贝叶斯网络模型和决策树模型对南京市 58,238 站点降水天气背景下的暴雨天气特征建立了诊断模型，得到以下结论：

1) 通过对南京市降水天气与暴雨天气特征的统计分析，发现暴雨天气的发生相较于普通降水天气的发生有着更为苛刻的气象条件的要求。

2) 基于贝叶斯网络模型分类算法对预处理数建立的诊断模型准确率为 83.61%。

3) 基于决策树模型分类算法得到的决策树模型，可以看出，决策树的根节点是相对湿度，说明判断普通降水背景下的暴雨天气是否发生的最重要因素是相对湿度；决策树模型的准确率为 81.97%。

4) 贝叶斯网络模型比决策树模型的准确率要稍微好一些，相较而言，贝叶斯网络模型计算更为方便，准确率也要稍高。

随着信息时代的发展，利用计算机对天气数据进行预测已经并且正在走向更为宽广的未来。将图论知识与人工智能等技术进行结合与优化，也将成为未来的一个发展趋势。

参考文献

[1] 赵琳娜, 梁莉, 王成鑫. 基于贝叶斯模型平均的集合降水预报偏差订正[C]//中国气象学会. 第 28 届中国气象学会年会论文集: 2011 年卷. 厦门: 中国气象学会, 2011: 1-13.

[2] 史达伟, 李超, 史逸民, 等. 基于机器学习的大雾天气背景下特强浓雾本地化诊断研究[J]. 灾害学, 2018, 33(2): 193-199.

[3] 张远汀, 龚伟伟, 叶钰, 等. 应用机器学习技术预测强雨雪天气过程中的积雪[J]. 科学技术与工程, 2019, 19(15): 58-69.

[4] 阮成卿, 李建平. 华北汛期降水分离时间尺度降尺度预测模型的改进[J]. 大气科学, 2016, 40(1): 215-226.

[5] 任民宏, 肖海蓉. 基于朴素贝叶斯分类算法的股指预测研究[J]. 陕西理工学院学报: 自然科学版, 2014, 30(3): 68-73.

[6] Rish, I. (2001) An Empirical Study of the Naive Bayes Classifier. *Journal of Universal Computer Science*, **1**, 127.

[7] Kohavi, R. (1997) Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, 1997, 202-207.

[8] Peng, F., Schuurmans, D. and Wang, S. (2004) Augmenting Naive Bayes Classifiers with Statistical Language Models. *Information Retrieval*, **7**, 317-345. https://doi.org/10.1023/B:INRT.0000011209.19643.e2

[9] 张文钧, 蒋良孝, 张欢, 等. 一种双层贝叶斯模型:随机森林朴素贝叶斯[J]. 计算机研究与发展, 2021, 58(9): 2040-2051.

[10] 黄宇达, 王迤冉, 等. 基于朴素贝叶斯与 ID3 算法的决策树分类[J]. 计算机工程, 2012, 38(14): 41-43.

[11] 邸鹏, 段利国. 一种新型朴素贝叶斯文本分类算法[J]. 数据采集与处理, 2014, 29(1): 71-75.

[12] 李冬梅. 朴素贝叶斯与决策树混合分类方法的研究[D]: [硕士学位论文]. 大连: 大连海事大学, 2016.

[13] 蒋良孝. 朴素贝叶斯分类器及其改进算法研究[D]: [硕士学位论文]. 北京: 中国地质大学, 2009.

[14] 宫秀军, 孙建平, 史忠植. 主动贝叶斯网络分类器[J]. 计算机研究与发展, 2002, 39(5): 574-579.

[15] 吴立增, 朱永利, 苑津莎. 基于贝叶斯网络分类器的变压器综合故障诊断方法[J]. 电工技术学报, 2005, 20(4): 45-51.

[16] 陈雪, 戴芹, 马建文, 等. 贝叶斯网络分类算法在遥感数据变化检测上的应用[J]. 北京师范大学学报: 自然科学版, 2005, 41(1): 97-100.

[17] Daphne Koller, Nir Friedman. 概率图模型: 原理与技术[M]. 王飞跃, 韩素素, 译. 北京: 清华大学出版社, 2015: 51.

[18] 李航. 统计学习方法[M]. 第二版. 北京: 清华大学出版社, 2012: 67-68.