Statistics and Application 统计学与应用, 2016, 5(2), 163-171
Published Online June 2016 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2016.52016

Hans汉斯

Prediction and Analysis of Forest Fire Based on Machine Learning

Dan Liu

College of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: dan-jx@163.com

Received: Jun. 8th, 2016; accepted: Jun. 27th, 2016; published: Jun. 30th, 2016

Abstract

Forest fire is a kind of destructive and huge disaster, which causes irreparable damage in the ecological environment and brings great harm to human survival and life. Especially since the 1980s, the global warming has continued, and forest fires occur more frequently, leading to huge economic losses to the world each year. So how to predict, prevent or reduce the hazards of forest fires become the common concern of many science disciplines. Rapid detection is an effective way to predict forest fire. To achieve this goal, one approach is to use automated tools based on sensor data, such as the data that meteorological stations offer. The study found that the meteorological conditions (such as temperature, wind speed) are important factors influencing forest fires and some fire indicators (such as forest fire weather index). Therefore, we will explore several machine learning methods to predict forest fire area. Using the data collected from Montesinho National Park in Northeastern Portugal, and a variety of different machine learning techniques, such as support vector machines (SVM) and random forests, four different characteristics (distribution of space, time, climate indicators and FWI system indicator) were analyzed. The best results were obtained using support vector machines and four basic meteorological inputs (such as temperature, relative humidity, wind speed and precipitation), which could accurately predict the damage area of small-scale and frequent fires. The above prediction methods are of great significance for improving the management and allocation of fire-fighting resources.

Keywords

Forest Fire, Machine Learning, Support Vector Machine, Random Forest

基于机器学习对森林火灾的预测分析

刘　丹

文章引用: 刘丹. 基于机器学习对森林火灾的预测分析[J]. 统计学与应用, 2016, 5(2): 163-171.
http://dx.doi.org/10.12677/sa.2016.52016

云南财经大学统计与数学学院，云南 昆明
Email: dan-jx@163.com

收稿日期：2016年6月8日；录用日期：2016年6月27日；发布日期：2016年6月30日

摘 要

森林火灾是一种破坏性及其巨大的灾难，在对生态环境造成难以挽回的破坏的同时还对人类生存与生活带来极大的危害，特别是20世纪80年代以来，全球气候持续变暖，林火有上升的趋势，每年发生的森林火灾都给世界各国造成了巨大的经济损失，使得对于如何预测、防治或减少森林火灾的危害成为许多学科领域共同关注的科学任务。而快速检测正是预测森林火灾的一个有效途径。为了实现这一目标，一种方法是使用基于传感器的自动工具，如气象观测站所提供的数据。研究发现，气象条件(如气温，风速)是影响森林火灾发生和一些火灾指标(如森林火险天气指数)的重要因素。因此，我们将探讨几种机器学习预测森林火灾面积的方法。利用来自葡萄牙东北部的Montesinho国家公园采集测试的真实数据，使用多种不同的机器学习技术，如支持向量机(SVM)和随机森林，对四组不同的特征(分布空间，时间，气候指标和FWI系统指标)进行分析。最好的结果是使用支持向量机和四个基本气象输入(如气温，相对湿度，风速和降水量)，它能够准确预测规模较小且发生频繁的火灾的受灾面积。上述预测方法对于提高消防资源的管理和调配有重大意义。

关键词

森林火灾，机器学习，支持向量机，随机森林

1. 引言

森林火灾已经成为备受关注的环境问题，不仅影响森林保护，还会造成巨大经济损失和严重的生态破坏，给人类的生活带来灾难性影响。森林火灾的发生源于多种原因(如人为疏忽和闪电)，尽管越来越多的国家斥巨资来控制这场灾难，全世界每年仍有数百万公顷的森林葬身火海。

近几年，快速检测已慢慢成为预测火灾的关键要素，但由于传统的监视费用昂贵且受主观因素的影响较大，人们逐渐重视并发展自动化的解决方案。这些方案大致可分为三类：卫星，红外扫描仪和局部传感器[1]。由于卫星定位的延迟和扫描仪高昂的设备成本和维护成本，这些方案不能用来解决所有的情况。研究表明，天气条件，如气候和相对湿度，是影响火灾发生的关键因素。而自动气象站[2]通常可以提供有效数据，这些数据可以实时采集且成本低廉。

在过去，气象数据已纳入量化指标体系，用以预防火灾危险、警告公众和支持消防管理决策。特别是，加拿大森林火险天气指数(FWI)系统[3] [4]的设计，在上世纪 70 年代计算机还十分稀缺的情况下它只需要利用手动收集的四个气象观测读数(气候，相对湿度，风速和降水量)进行简单的计算。目前该指数系统在加拿大和其他一些国家广泛使用。

现今，由于计算机技术的快速发展，使得对数据的采集越发的实效和便捷。机器学习就是信息技术进步的一个体现，使用自动化的数据挖掘工具分析原始数据可以为高层决策者提取有效信息。事实上，机器学习技术已经应用到火灾探测领域[5] [6]。例如采用神经网络(NN)预测人类引起的森林火灾；红外扫描仪和神经网络结合在减少森林火灾误报率方面达到 90%的成功率；北美森林大火的卫星图像应用支持向量机获得了 75%的准确率在森林火灾可能性上；使用卫星和气象数据应用逻辑回归、随机森林和决策

树来探测斯洛文尼亚森林火灾。

学习上述方法，我们利用机器学习对森林火灾的发生做出预测，并分析的模型的错判率。我们使用从葡萄牙东北部的 Montesinho 国家公园采集的最新数据预测森林火灾的受灾面积。应用多种方法(即多元回归，支持向量机和随机森林)对四类指标进行分析(即分布空间，时间，气候指标和 FWI 系统指标)。将对四类不同性质的指标分别进行基于机器学习的数据分析，如气候指标(即气候，相对湿度，风速和降水量)与支持向量机相结合，能够预测森林火灾的燃烧面积，构建火灾燃烧等级对未来的火灾防治和消防管理决策是非常有用的。

2. 数据分析

2.1. 数据介绍

论文涉及的森林火灾数据来自葡萄牙东北部的 Montesinho 国家公园的数据库，信息包含 13 个变量：Montesinho 国家公园的空间坐标；信息采集的月份和每周的其中一天；FWI 系统的指数变量 FFMC (细小可燃物湿度码)、DMC (粗腐殖质湿度码)、DC (干旱码)和 ISI (初始蔓延指数)；四种可直接测量的气温、相对湿度、风速和降水量的气象数据；森林火灾燃烧的面积。

2.2. 变量解释

FWI 系统是由 6 个部分组成：3 个代表可燃物湿度的基本子指数，分别为细小可燃物湿度码(FFMC, fine fuel moisture code)，粗腐殖质湿度码(DMC, duff moisture code)和干旱码(DC, drought code)；2 个代表可燃物扩散速率和消耗率的中间子指数，分别为初始蔓延速度(ISI, initial spread)和累积指数(BUI, build up)；1 个代表火强烈程度的最终指数，FWI。火险气候指数系统中所涉及的元素由每天测量的气温、相对湿度、风速和降水量的气象数据中计算得到。

2.2.1. 细小可燃物湿度码 FFMC

FFMC 代表的是森林中地被物干质量为 0.25 kg·m^{-2}，厚度为 1.2 cm 的枯枝落叶和其他的已经固化的细小燃料的含水率。FFMC 是代表细小可燃物的可燃性和易燃性的指标，它受温度、降水、相对湿度和风速的影响，值随着燃料含水率的变化而改变，其核心是一个简单的水分交换的指数模型：$m_0 = 147.2 \times (101.0 - c_{FFMC}) / (59.5 + c_{FFMC})$。其中 m_0 为前一天的燃料含水率。

2.2.2. 粗腐殖质湿度码 DMC

DMC 代表的是森林地被物最上层厚度约为 7 cm，干质量为 5.00 kg·m^{-2} 的有机物质的含水率。DMC 用来表明中等下层落叶层和中型木质物质的燃料消耗，DMC 模型是一个简单的水分交换的指数模型：$M_0 = 20.00 + \ln\left[\frac{c_{DMC} - 244.73}{-43.43}\right]$。其中 M_0 表示前一天的地表可燃物含水率。

2.2.3. 干旱码 DC

DC 代表的是森林地被物中干质量为 25.00 kg·m^{-2}，厚度为 18 cm 的深层可燃物和粗死木残体的含水率。干旱码用于衡量季节性干旱对森林燃料以及深层下层落叶层和大型段木的影响指标。DC 模型的核心是一个简单的指数模型：$Q_0 = 400 \times e^{-CD/400}$。其中 Q_0 表示前一天干旱码的湿度指标。

2.2.4. 初始蔓延指数 ISI

ISI 代表的是火灾蔓延的潜在等级，由 FFMC 和风速两个指标决定。ISI 一直是表示火灾蔓延等级的很好指标。

3. 模型描述

3.1. 多元线性回归模型

将给定 $x_{1i}, x_{2i}, \cdots, x_{3i}$ 条件下y_i的均值

$$\mathrm{E}\left(y_i \mid x_{1i}, x_{2i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \tag{1}$$

定义为总体回归函数(Population Regression Function, PRF)。定义 $y_i - \mathrm{E}\left(y_i \mid x_{1i}, x_{2i}, \cdots, x_{ki}\right)$ 为误差项(error term)，记为μ_i，即 $\mu_i = y_i - \mathrm{E}\left(y_i \mid x_{1i}, x_{2i}, \cdots, x_{ki}\right)$，这样 $y_i = \mathrm{E}\left(y_i \mid x_{1i}, x_{2i}, \cdots, x_{ki}\right) + \mu_i$，或

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i \tag{2}$$

由于多元线性回归模型只能学习线性映射，它拟合本数据效果很差，本文为了解决这个问题，一种替代的方法是使用基于树结构的方法，如决策树(DT)，或非线性函数，如支持向量机(SVM)。

3.2. 决策树(DT)模型

决策树大多是用来分类的。选择分类属性的标准是信息增益最大，涉及到熵的概念。而在做回归树的时候，我们希望和回归有多一点联系，因此选择变量的标准我们用残差平方和。我们知道回归分析的最小二乘的解就是最小化残差平方和。在决策树的根部，所有的样本都在这里，此时树还没有生长，这棵树的残差平方和就是回归的残差平方和。然后选择一个变量也就是一个属性，这个变量使得通过这个进行分类后的两部分的残差平方和的和最小。然后在分叉的两个节点处，再利用这样的准则，选择之后的分类属性。一直这样下去，直到生成一颗完整的树。

3.3. 支持向量机(SVM)模型

支持向量机(support vector machine)是一种分类方法，而由 SVM 发展出来的回归方法称为支持向量回归(support vector regression)。假定其目的是把空间中的两类点($y=-1$ 或 $y=1$)用超平面 $\omega^{\mathrm{T}} x + b = 0$ 分开(在严格线性可分的情况下，存在这样的超平面)。而且希望这个超平面距离两类点的距离最大，也就是说，使得隔离带宽 $\rho = \dfrac{2}{\|\omega\|}$ 最大。这等价于用 Lagrange 乘子法求下式的极小值

$L\left(\omega, b, a\right) = \dfrac{1}{2}\omega^2 - \sum_{i=2}^{n} \alpha_i \left[y_i \left(\omega^{\mathrm{T}} x_i + b\right) - 1 \right]$，根据得到的解 ω^*，b^*，α^* 得到最优分割超平面方程 $\omega^{*\mathrm{T}} x + b^* = 0$。任意点 x 的函数值 $\omega^{*\mathrm{T}} x + b^*$ 的符号确定了该点的分类，或者说判别函数为 $\mathrm{sgn}\left(\omega^{*\mathrm{T}} x + b^*\right)$。上面介绍的是严格线性可分的情况，如果允许一些错误，则称为近似线性可分问题，结果与此有同样的形式。

3.4. 随机森林模型

随机森林作为一种组合分类器，采用 bootstrap 抽样技术从原始数据集中抽取 n_{tree} 个训练集，每个训练集的大小约为原始数据集的三分之二。为每一个 bootstrap 训练集分别建立分类回归树(Classification and Regression Tree, CART)，共产生 n_{tree} 棵决策树构成一片“森林”，这些决策树均不进行剪枝(unpruned)。在每棵树生长过程中，并不是选择全部 M 个属性中的最优属性作为内部节点进行分支(split)，而是从随机选择的 $m_{try} \le M$ 个属性中选择最优属性进行分支。集合 n_{tree} 棵决策树的预测结果，采用投票(voting)的方式决定新样本的类别。

随机森林在训练过程中的每次 bootstrap 抽样，将有约三分之一的数据未被抽中，这部分数据被称为袋外(out-of-bag)数据。随机森林利用这部分数据进行内部的误差估计，产生 OOB 误差(out-of-bag error)。Breiman 通过实验证明，OOB 误差是无偏估计，近似于交叉验证得到的误差。

4. 实验验证及结果

4.1. 线性相关性分析

做出细小可燃物含水率(FFMC)与气象因子之间的相关性矩阵，查看他们之间的相关性。

从表 1 可以看出 temp 与 FFMC 成正相关，他们的相关性为 0.432；RH 与 FFMC 成负相关，他们的相关性为−0.301；wind 与 FFMC 成负相关，他们的相关性为−0.028；rain 与 FFMC 成正相关，他们的相关性为 0.057。同时我们可以看出 temp 和 RH 与 FFMC 的相关性不是很强，wind 和 rain 与 FFMC 几乎不相关，在做线性回归时，可以不考虑变量 wind 和 rain。

用 R 里面的程序包"car"中的 scatterplotMatrix 函数，画出各个量之间函数图(图 1)以及散点图(图 2)。

从图 1 和图 2 我们可以看出气候变量与 FFMC 的线性相关性并不强，他们之间可能存在某种非线性关系，本文主要考虑存在交互项和高次幂项的多元回归，同时也考虑机器学习方法中的决策树、随机森林以及支持向量机等方法，我们用五折交叉验证方法验证模型的优劣性，选出最好的模型来预测 FFMC。也许模型效果不好，我们认为主要缺少前一、两天 FFMC 的数据，前一、两天 FFMC 的数据对预测下一天 FFMC 的数据起着比较重要的作用。

4.2. 多元线性模型

用 temp 和 RH 作为变量的多元线性模型的输出结果可以看出，回归方程为：$y_{FFMC} = 0.378 \cdot X_{temp} - 0.102 \cdot x_{RH}$，变量 X_{temp} 和变量 x_{RH} 的 t 统计量的估计值为 8.106 和−2.181，由对应的 P 值都比显著水平 0.05 小，可得两个偏回归系数在显著水平 0.05 下均显著不为零，进一步估计地剩余方差 σ^2 的估计值为 0.899，统计量的估计值为 61.73，由对应的 P 值为 2.2e−16，说明回归方程显著，可决系数 $R^2 = 0.194$，修正的可决系数 $R^2 = 0.191$，说明方程的拟合效果很差，主要原因是 temp 和 RH 的相关性不是很高，导致他们线性关系较差。

我们考虑非线性回归，在原模型上添加 temp 和 RH 的交互项和二次项，模型输出结果可以看出，回归方程为：$y_{FFMC} = 0.09 + 0.42 \cdot X_{temp} + 0.01 \cdot x_{RH} + 0.18 \cdot x_{temp} \cdot x_{RH}$，变量 X_{temp}、变量 x_{RH} 和变量 $x_{temp} \cdot x_{RH}$ 的 t 统计量的估计值为 9.120、0.287 和 5.247，除 x_{RH} 由对应的 P 值比显著水平 0.05 大，其余都比显著水平 0.05 小，可得除 x_{RH} 偏回归系数在显著水平 0.05 下均显著为零，其余两个偏回归系数在显著水平 0.05 下均显著不为零，进一步估计地剩余方差 σ^2 的估计值为 0.8773，F 统计量的估计值为 52.46，由对应的 P 值为 2.2e−16，说明回归方程显著，可决系数 $R^2 = 0.235$，修正的可决系数 $R^2 = 0.230$，说明方程的拟合效果很差，主要原因是 temp 和 RH 的相关性不是很高，导致他们线性关系较差。而逐步回归的结果：可决系数 $R^2 = 0.194$，修正的可决系数 $R^2 = 0.191$，同样说明方程的拟合效果很差，故用传统的多元统计回归模型不能很好的拟合该数据，可能数据受到一些极端值的影响，导致拟合效果很差。

Table 1. Table of linear correlation coefficient of meteorological factors
表 1. 气象因子数据线性相关系数表

	FFMC	temp	RH	wind	rain
FFMC	1.000	0.432	−0.301	−0.028	0.057
temp	0.432	1.000	−0.527	−0.227	0.069
RH	−0.301	−0.527	1.000	0.069	0.099
wind	−0.028	−0.227	0.069	1.000	0.061
rain	0.057	0.069	0.099	0.061	1.000

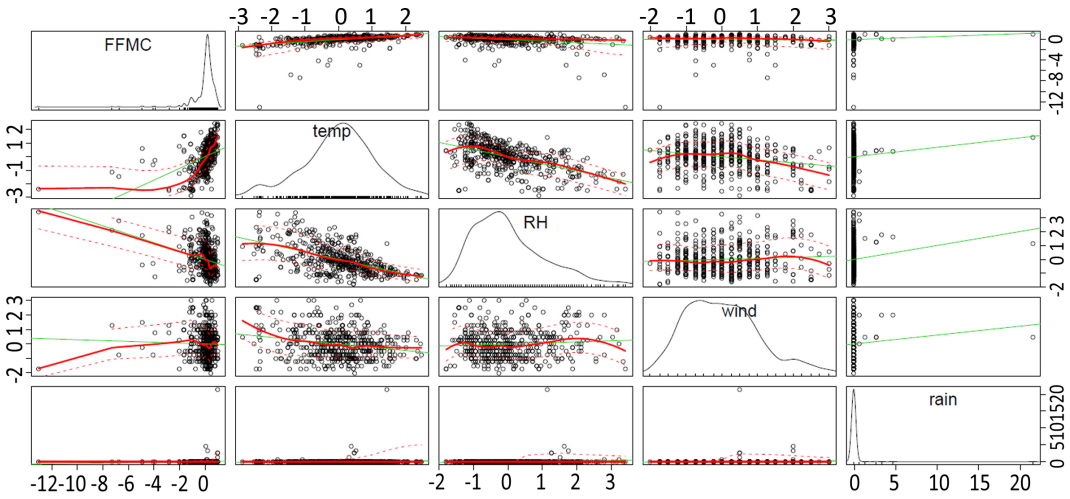

Figure 1. Function diagram of data among variables
图 1. 数据各变量之间的两两函数图

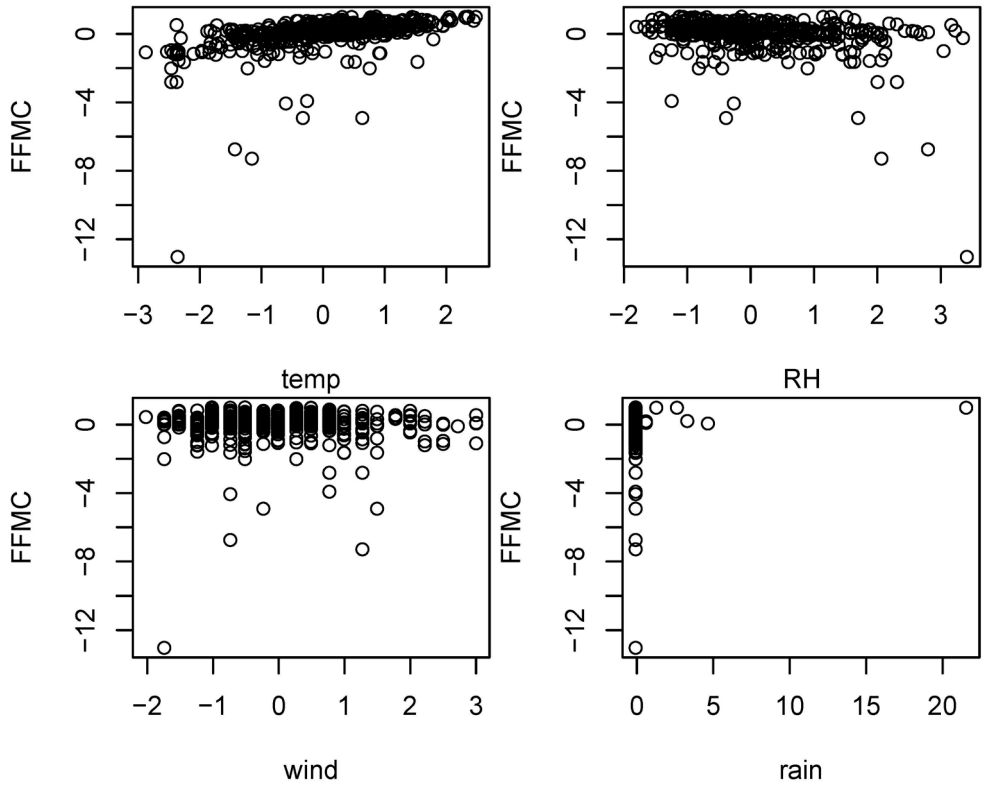

Figure 2. Scatter plots of climate variables on FFMC
图 2. 气候变量对 FFMC 的散点图

4.3. 决策树和随机森林

下面运用机器学习方法来拟合数据，首先使用机器学习中的回归树模型拟合数据，输出结果如图 3，可以看出第一个节点是在 temp = 0.17 进行分支的，然后在 RH = −1.2 和 temp = −2.4 进行分支从决策树的生成过程可以知道，主要根据 temp 和 RH 进行分支，说明 temp 和 RH 在决策树生成过程中起主要作用，其他两个变量几乎不起作用。

使用机器学习中的随机森林模型拟合数据，输出结果如图 4，我们可以看出随机森林回归过程中 temp

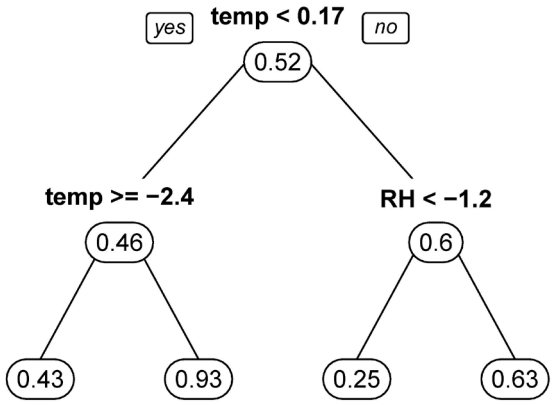

Figure 3. Decision tree
图 3. 决策树

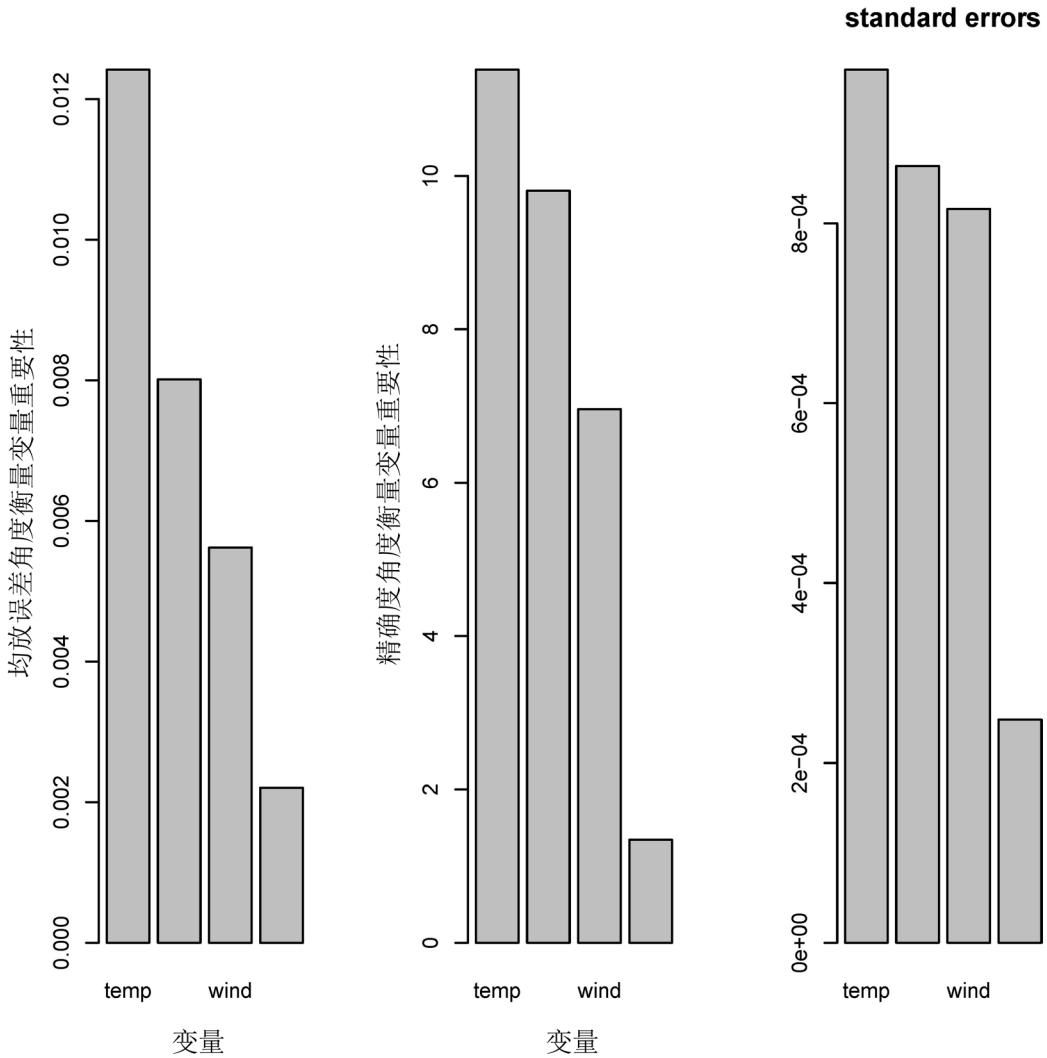

Figure 4. Importance of random forest variables
图 4. 随机森林变量重要性

起主要作用，RH 其次，wind 再次之，而 rain 起作用最小。我们通过五折交叉验证来判断上述模型的优劣性。

五折交叉验证结果如表 2，我们看出多元统计模型在这些模型中是最优的，但是多元统计模型的 MSE = 1.259，说明模型拟合数据很很不好，可能是数据本身波动性较大。也许该数据适合用来回归。

4.4. 机器学习方法

由于数据回归效果不好，但该数据可以用来分类回归，在本文主要用 Logistic 回归和机器学习方法，如决策树，人工神经网络回归，随机森林，支持向量机和 K 近邻等方法处理数据。

Logistic 回归，决策树，人工神经网络回归，随机森林，支持向量机和 K 近邻等的实验结果见表 3。

从表 3 可以看出这些方法的误判率都比较高，都在 40%左右，效果不是很好，这可能跟数据受到干扰有关，随机森林和支持向量机的结果相对还是好一点，也许这些方法不能很好的拟合该数据，我们以后会寻找更好的方法来拟合数据，使得我们的模型预测的精确度提高。

5. 结论

根据所做模型相关性分析气候向量必然与火灾发生有一定的相关性，直接利用 4 种基本气候指标拟合的线性回归并不是显著的，意味着 4 种基本指标与森林燃烧面积可能是非线性相关的(本文由于数据的非完整性并没能证明它们是非线性相关的)。

改进模型检测可以看出气候变量与 FFMC 的线性相关性也并不强，他们之间可能存在某种非线性关系，考虑到存在交互项和高次幂项的多元回归，我们用五折交叉验证方法验证模型的优劣性，选出最好的模型来预测 FFMC。得出模型效果仍然不好，我们认为 FFMC 数据的记录缺少时间上的连续性，并不能在某个时间段中连续的观测出 FFMC 的数值。

模型中我们先将燃烧面积大于 0 的数据看作森林火灾发生一次，生成一个新变量，变量为 1 时，火灾发生，变量为 0 时，利用机器学习方法来回归预测是否发生森林火灾。可以看出随机森林回归过程中

Table 2. Half off cross validation results
表 2. 五折交叉验证结果

回归方法	MSE
多元统计回归	1.259
决策树	1.285
随机森林	1.278

Table 3. Machine learning half off cross validation results
表 3. 机器学习五折交叉验证结果

分类方法	误判率
Logistic 回归	0.487
决策树	0.412
随机森林	0.325
人工神经网络	0.433
支持向量机	0.362
K 近邻	0.507

temp 起主要作用，RH 其次，wind 再次之，而 rain 起作用最小，虽然检测模型的误判率一度达到 40%，但我们仍觉得在森林火灾的发生很大程度上取决于 temp，控制 temp 的临界值(或者临界区间)可以很好的预防森林火灾的发生。

参考文献 (References)

[1] Cortez. P. and Morais, A. (2007) A Data Mining Approach to Predict Forest Fires Using Meteorological Data. http://www3.dsi.uminho.pt/pcortez/fires.pdf

[2] 曲智林, 胡海清. 基于气象因子的森林火灾面积预测模型[J]. 应用生态学报, 2007, 18(12): 2705-2709.

[3] 袁建, 江洪, 信晓颖. 基于 FWI 的浙江省森林火险等级划分[J]. 福建农林大学学报: 自然科学版, 2013, 42(3): 283-288.

[4] 田晓瑞, Douglas J. McRae, 舒立福, 赵凤君, 王明玉. 大兴安岭地区森林火险变化及 FWI 适用性评估[J]. 林业科学, 2010, 46(5): 127-132.

[5] 王正旺, 庞转棠, 魏建军, 杨艳萍, 杨梅红. 森林火险天气等级预测及火情检测应用[J]. 自然灾害学报, 2006, 15(5): 154-161.

[6] 牛若芸, 翟盘茂, 孙明华. 森林火险气象指数及其构建方法回顾[J]. 气象, 2006, 32(12): 3-9.