**基于数据挖掘的森林火灾预测**

目录

[一、数据背景 3](#_Toc327791115)

[摘要 3](#_Toc327791116)

[数据描述 4](#_Toc327791117)

[资料来源 4](#_Toc327791118)

[相关论文 4](#_Toc327791119)

[数据集信息 5](#_Toc327791120)

[属性信息 6](#_Toc327791121)

[二、数据建模 7](#_Toc327791122)

[（1）数据的读入 7](#_Toc327791123)

[（2）建立个别属性的支持向量机（SVM）模型 10](#_Toc327791124)

[（3）建立所有属性的支持向量机模型 17](#_Toc327791125)

[三、数据分析与挖掘 19](#_Toc327791126)

[（1）对使用“日期条件”生成的数据分析 19](#_Toc327791127)

[（2）对使用“坐标条件”生成的数据分析 22](#_Toc327791128)

[（3）对使用“自然条件”生成的数据分析 26](#_Toc327791129)

[（4）对使用“FWI指数条件”生成的数据分析 29](#_Toc327791130)

[（5）对使用所有属性生成的数据分析 33](#_Toc327791131)

[四、结论 41](#_Toc327791132)

# 一、数据背景

数据来源：<http://archive.ics.uci.edu/ml/datasets/Forest+Fires>

**Abstract**:

Forest fires are a major environmental issue, creating economical and ecological damage while endangering human lives. Fast detection is a key element for controlling such phenomenon. To achieve this, one alternative is to use automatic tools based on local sensors, such as provided by meteorological stations.

In effect, meteorological conditions (e.g. temperature, wind) are known to influence forest fires and several fire indexes, such as the forest Fire Weather Index (FWI), use such data. In this work, we explore a DataMining (DM) approach to predict the burned area of forest fires. Five different DM techniques, e.g. Support Vector Machines (SVM) and Random Forests, and four distinct feature selection setups (using spatial, temporal, FWI components and weather attributes), were tested on recent real-world data collected from the northeast region of Portugal.

The best configuration uses a SVM and four meteorological inputs (i.e. temperature, relative humidity, rain and wind) and it is capable of predicting the burned area of small fires, which are more frequent. Such knowledge is particularly useful for improving firefighting resource management (e.g. prioritizing targets for air tankers and ground crews)。

摘要：

这是一份困难的的回归分析任务，任务目的是通过利用气象学和其他学科的综合数据来预测葡萄牙东北地区将要发生森林火灾的地区。

森林火灾是造成经济和生态危害的主要自然现象之一，而控制这种现象的关键方法就是快速发现并行动。为了达到这个目的，其中一个方法是使用自动化工具进行预测，例如使用基于气象站提供的局部传感器数据。

事实上，气象条件（如温度，风速等）能够显著影响森林火灾的发生以及几种火灾指标，例如“森林火险天气指数”（FWI）。在这份研究中，我将会通过数据挖掘的方法预测森林火灾发生的区域。可以使用以下几种数据挖掘工具：决策树（DT）、人工神经网络（NN）、支持向量机（SVM），并且通过四组不同的数据属性进行设置（空间坐标、时间因数、FWI指数以及气象数据）。

本实验中主要采用SVM进行数据挖掘。

通过实验，可以发现使用SVM对四个气象输入数据（温度、相关湿度、降雨量和风速）进行处理后，能较准确预测到小型火灾的发生区域。这个结论将有效提高消防资源的管理（譬如优先对将要发生火灾的区域配置救火飞机和相应的消防人员）。

以上数据来自葡萄牙东北地区的真实观测记录。

**Description**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 517 | **Area:** | Physical |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 13 | **Date Donated** | 2008-02-29 |
| **Associated Tasks:** | Regression | **Missing Values?** | N/A | **Number of Web Hits:** | 101935 |

## 数据描述

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集特征 | 多元 | 实例数 | 517（条） | 领域 | 物理 |
| 属性特征 | 真实 | 属性数 | 13（列） | 发表日期 | 2008-02-29 |
| 关联任务 | 回归 | 丢失值 | 无 | 网页访问数 | 101935（次） |

Source

资料来源:

Paulo Cortez, pcortez@dsi.uminho.pt, Department of Information Systems, University of Minho, Portugal.

Anbal Morais, araimorais@gmail.com, Department of Information Systems, University of Minho, Portugal.

**Relevant Papers:**

相关论文**：**

[Cortez and Morais, 2007] P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimares, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9

**Data Set Information:**

In [Cortez and Morais, 2007], the output 'area' was first transformed with a ln(x+1) function.

Then, several Data Mining methods were applied. After fitting the models, the outputs were post-processed with the inverse of the ln(x+1) transform. Four different input setups were used. The experiments were conducted using a 10-fold (cross-validation) x 30 runs. Two regression metrics were measured: MAD and RMSE. A Gaussian support vector machine (SVM) fed with only 4 direct weather conditions (temp, RH, wind and rain) obtained the best MAD value: 12.71 +- 0.01 (mean and confidence interval within 95% using a t-student distribution). The best RMSE was attained by the naive mean predictor. An analysis to the regression error curve (REC) shows that the SVM model predicts more examples within a lower admitted error. In effect, the SVM model predicts better small fires, which are the majority.

数据集信息：

在这份名为[Cortez and Morais, 2007]的数据集中，输出属性‘area’首先使用一个ln(x+1)函数进行转化。

然后，应用几种数据挖掘提供的方法对其进行分析。在成功对模型进行拟合后，通过使用四种不同的输入，得到逆ln(x+1) 转换后的输出结果。使用折叠次数为10的交互验证进行30次实验，再此期间测定了两个回归指标：MAD和RMSE。再使用一个高斯支持向量机（SVM）对四个直接气象条件（temp，RH，wind和rain）进行操作，获得最佳MAD取值：12.71 +- 0.01（通过t-student distribution获得分布在95%的信心区间的值），且RMSE的目标值需要接近原始的平均预测值。回归误差曲线（REC）表明，SVM模型预测到更多的例子，同时错误更少。事实上，SVM模型能更好地预测到较小规模的火灾，这是主要的优势所在。

注：原论文中使用了两个自行计算的MAD和RMSE回归指标，因过于复杂以下不进行计算，本实验专注于使用决策树、人工神经网络和支持向量机进行数据分析与挖掘。

**Attribute Information:**

For more information, read [Cortez and Morais, 2007].   
1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9   
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9   
3. month - month of the year: 'jan' to 'dec'   
4. day - day of the week: 'mon' to 'sun'   
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20   
6. DMC - DMC index from the FWI system: 1.1 to 291.3   
7. DC - DC index from the FWI system: 7.9 to 860.6   
8. ISI - ISI index from the FWI system: 0.0 to 56.10   
9. temp - temperature in Celsius degrees: 2.2 to 33.30   
10. RH - relative humidity in %: 15.0 to 100   
11. wind - wind speed in km/h: 0.40 to 9.40   
12. rain - outside rain in mm/m2 : 0.0 to 6.4   
13. area - the burned area of the forest (in ha): 0.00 to 1090.84   
(this output variable is very skewed towards 0.0, thus it may make   
sense to model with the logarithm transform).

属性信息：

获取更多信息请参阅《Cortez and Morais, 2007》

|  |  |  |  |
| --- | --- | --- | --- |
| 属性序号 | 属性名称 | 属性说明 | 取值范围 |
| 1 | X | Montesinho公园地图的X轴坐标 | 1~9 |
| 2 | Y | Montesinho公园地图的Y轴坐标 | 2~9 |
| 3 | month | 一年中的月份 | jan~dec |
| 4 | day | 一周中的日期 | mon~sun |
| 5 | FFMC | FWI系统中的FFMC指标 | 18.7~96.20 |
| 6 | DMC | FWI系统中的DMC指标 | 1.1~291.3 |
| 7 | DC | FWI系统中的DC指标 | 7.9~860.6 |
| 8 | ISI | FWI系统中的ISI指标 | 0.0~56.10 |
| 9 | temp | 温度（摄氏单位） | 2.2~33.3. |
| 10 | RH | 相对湿度（百分率） | 15.0~100 |
| 11 | wind | 风速（km/h） | 0.40~9.40 |
| 12 | rain | 室外雨量（mm/m2） | 0.0~6.4 |
| 13 | area | 森林的燃烧面积 | 0.00~1090.84 |

（由于输出值相当偏向于0.0，因此建立对数转换模型是有意义的。）

注：

FWI system—加拿大森林火险气候指数系统，是加拿大森林火险等级系统（CFFDRS）的重要组成部分。加拿大森林火险等级系统是当前世界上发展最完善、应用最广泛的系统之一，是世界上唯一能适应从区域到全球任何尺度的系统。加拿大火险气候指数系统（FWI）以实际中平衡含水率理论为基础，通过天气条件的变化计算可燃物含水率的变化，然后根据不同位置或大小的可燃物含水率划分森林潜在火险等级。

# 二、数据建模

## （1）数据的读入

首先使用“源”中的可变文件节点读入数据，在“文件”选项卡（图1.1）下可以看到Clementine正确读入了数据项。勾选“读取文件中的字段名”，使软件自动读取文件中属性的名称；由于数据集内没有标题，“跳过标题字符”保持为0；同样，开头和结尾没有空格，无须去除；由于数据集没有无效字符，故无需替换，保持“丢弃”即可；可以从预览框内看到，数据集的定界符主要是“，”和“换行符”，因此在“定界符”处作相应设置；而数据集中没有单、双引号，因此保持“丢弃”即可；其余设置保持默认值；

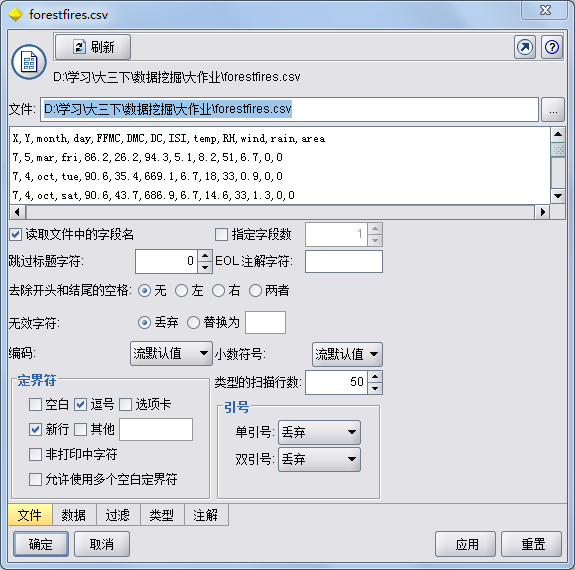


图1.1

在“类型”选项卡下点选“读取值”按钮，可以看到正确读取出来的13个属性（图1.2），及其相应的类型和值，设置保持默认；



图1.2

在“注解”选项卡输入节点名称“Forest Fires”（图1.3），点击确定按钮，即可生成节点。

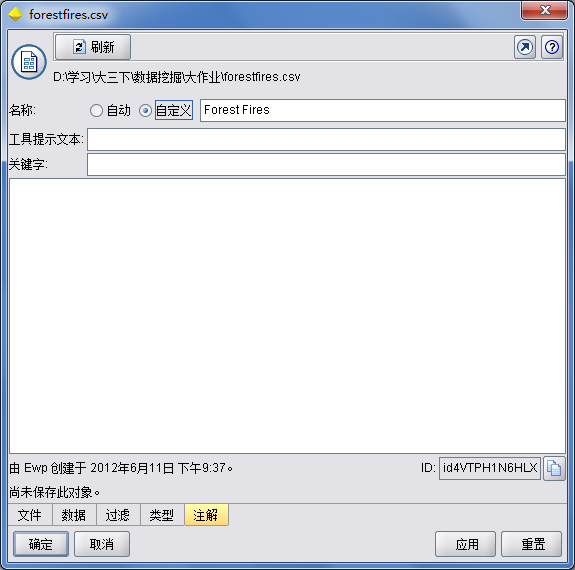


图1.3

创建一个与该节点相连接的“表”节点，执行后能看到数据以表单的形式被读入（图1.4）。

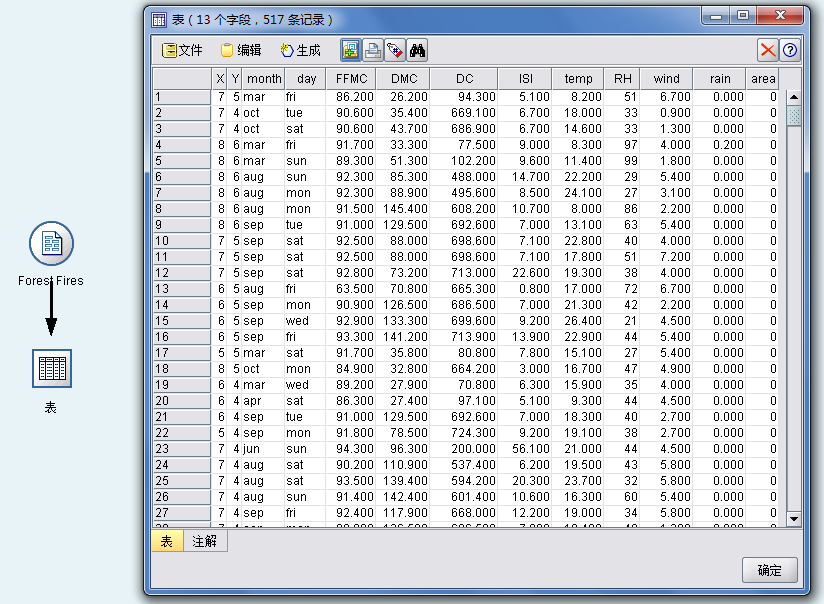


图1.4

注：由于本数据集没有丢失数据，故无需对其进行丢失数据过滤。

## （2）建立个别属性的支持向量机（SVM）模型

首先创建“类型”节点。在“字段选项”中双击“类型”节点，建立“Forest Fires”到“类型”节点的连接；双击“类型”节点，在“类型”选项卡中把 “area”一属性的类型设为“集”，“值”设为“读取”，“方向”设为“输出”；然后分别把“方向”一列中的“X、Y”，“month、day”，“FFMC、DMC、DC、ISI”，“temp、RH、wind、rain”中的一组设为输入，其余设为“无”，最后点击“读取值”按钮，以建立四个不同的类型节点。

如图1.5~1.8所示。



图1.5



图1.6



图1.7



图1.8

在“建模”选项卡中选择“SVM”模型，建立四个SVM节点，并与各个条件类型节点相连接，如图1.9所示。

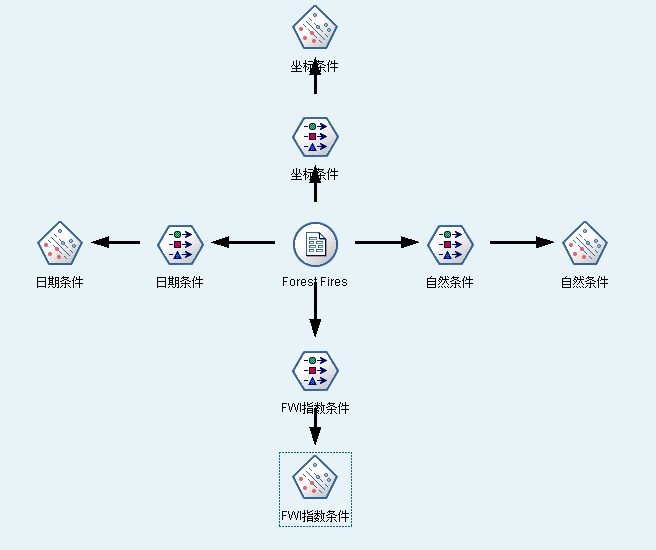


图1.9

以其中的“坐标条件”模型节点为例，双击该节点，在弹出的窗口中点选“专家”选项卡，内核类型选择“RBF”，其余选项保持不变，如图1.10所示



图1.10

在“分析”选项卡中，勾选“计算变量重要性”一项，之后点击“执行”按钮，如图1.11所示



图1.11

把在“模型”选项卡中生成的模型拖入主窗口，并建立与“坐标条件”节点的连接，如图1.12所示

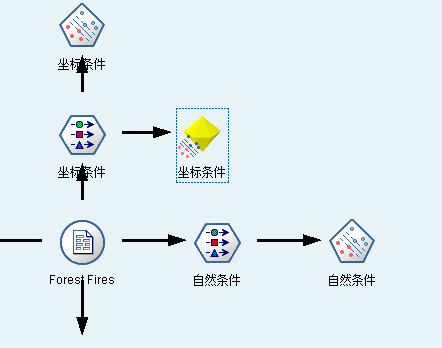


图1.12

双击按钮，改变“内核类型”为“多项式” ，再执行一次，以此类推同样生成另外两个内核“sigmoid”和“线性”的模型，并按顺序相连接，如图1.13所示

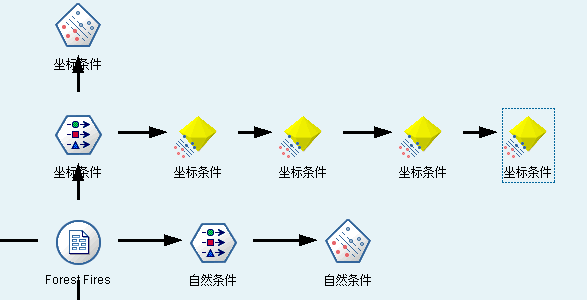
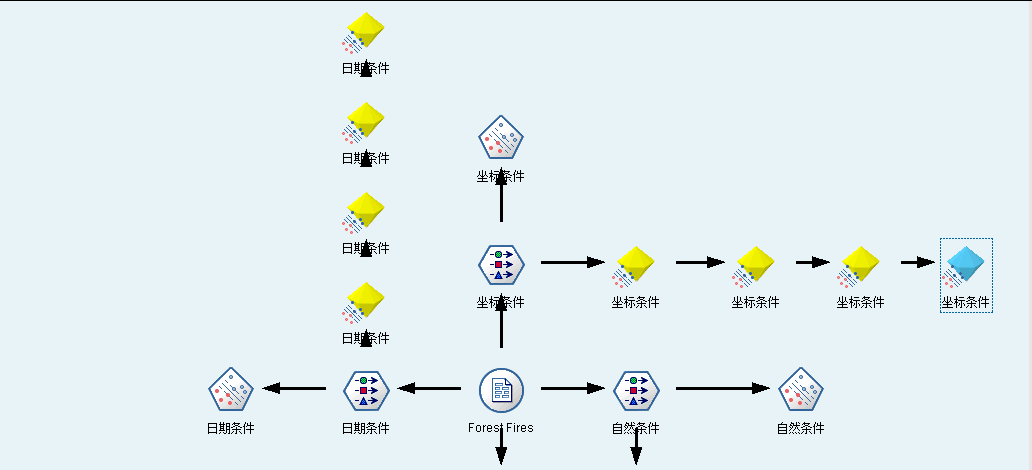


图1.13

用相同方法对令外几个类型节点建立模型，最终结果如图1.14所示



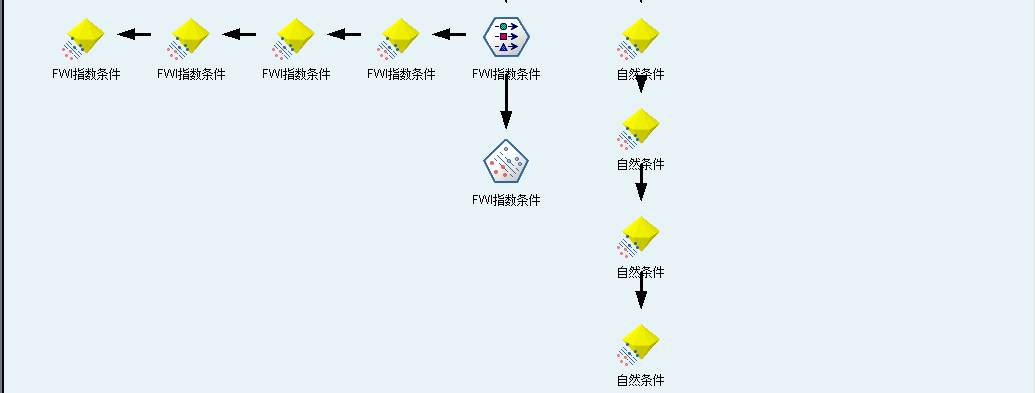


图1.14

在每个条件模型的最后一个节点连接一个“分析”节点，双击后点击“执行”按钮，即能查看各个模型的分析结果。

## （3）建立所有属性的支持向量机模型

以上步骤基本为本数据集相应的原论文[Cortez and Morais, 2007]中提到的数据挖掘方法。

但在数据分析的过程中我发现，四个类型（空间坐标、日期、FWI指数、自然条件）单独使用分析出来的结果绝对正确率都不高（在53%~54%之间）；而经过观察数据集发现，这四个类型的属性并不冲突，因此能否通过建立所有属性的支持向量机模型来提高预测的正确率呢？

接下来让我们尝试一下。

首先回到最初的起点，如图1.15所示

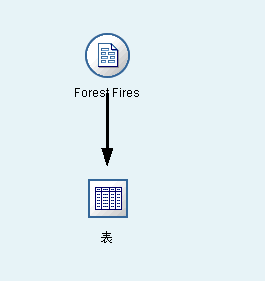


图1.15

接下来在建立“类型”节点时，双击后在“类型”选项卡下把除了“area”属性作为输出外，其余属性全部用于输入。

如图1.16所示。

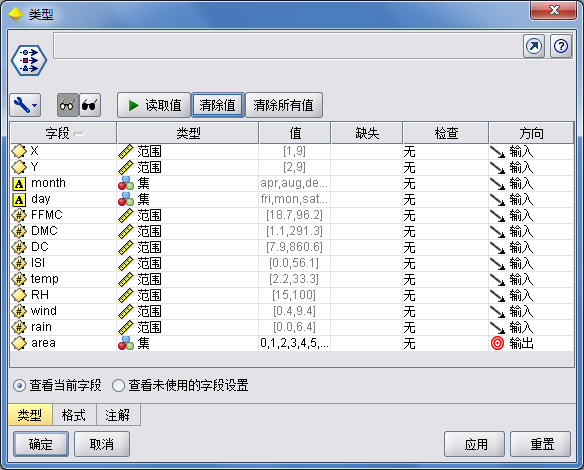


图1.16

点击“确定”后，其余步骤同对各个类型属性单独操作，建立四个内核的SVM模型，并在最后部分连接一个“分析”节点和一个“表”节点。

最终结果如图1.17所示

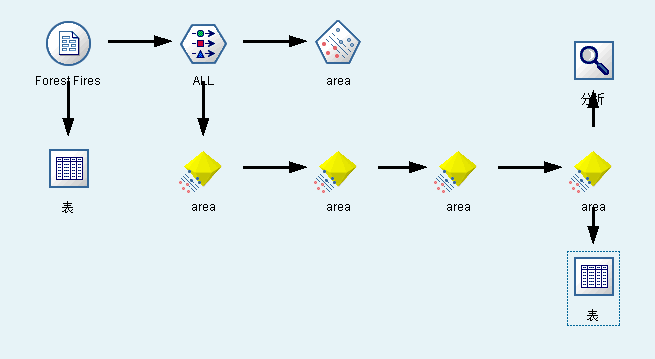


图1.17

# 三、数据分析与挖掘

## （1）对使用“日期条件”生成的数据分析

执行“分析”节点后，结果如图2.1所示



图2.1

从每个模型计算结果与现实数据的对比可以看出，使用“RBF”内核的预测结果与使用“多项式”的正确率、错误率都相同，分别为54.93%和45.07%；虽然正确率已经是相对最高，但绝对值仍然很低。

使用“sigmoid”内核的正确率只有53%，为四种内核中最低的。

使用“线性”内核的正确率也只有53.97%，同样不高。

双击各个模型，可以看到在不同内核下各个变量的重要性。如图2.2~图2.5所示



图2.2



图2.3



图2.4

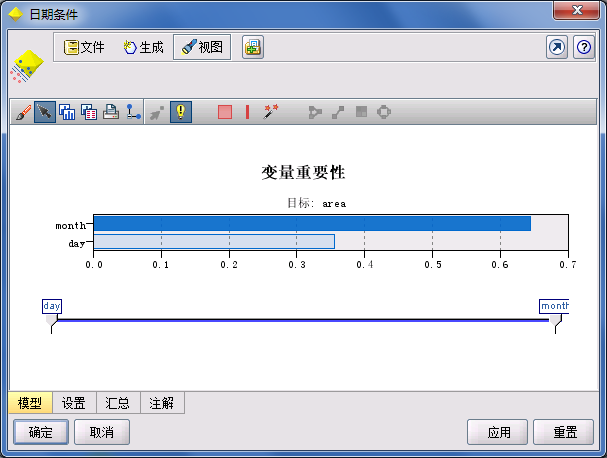


图2.5

上图从上至下分别为“RBF”、“多项式”、“sigmoid”和“线性”内核下的两变量“month”以及“day”的重要性。

可以看出，总体“month”的重要性较高。“month”的重要性在“多项式”下并未达到0.6，是四个内核中最小的；而“RBF”中的则最高，接近0.7。

相应的，“day”属性的重要性比重较低，且“多项式”中比重最大，超过了0.4；“RBF”中最小，恰恰比0.3多一点。

有意思的是，比重相对来说完全相反地“RBF”和“多项式”内核恰巧是正确率相对较高的内核，且正确率相等。

## （2）对使用“坐标条件”生成的数据分析

执行“分析”节点后，结果如图2.6所示

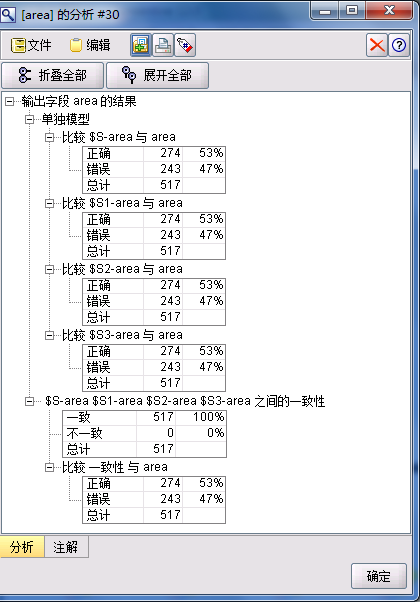


图2.6

从图中可以看出，四个内核预测出的结果正确率完全一致，都为53%，处于偏低位置。

且四个内核的结果同样完全一致，一致性为100%。

双击各个模型，可以看到在不同内核下各个变量的重要性。如图2.7~图2.10所示

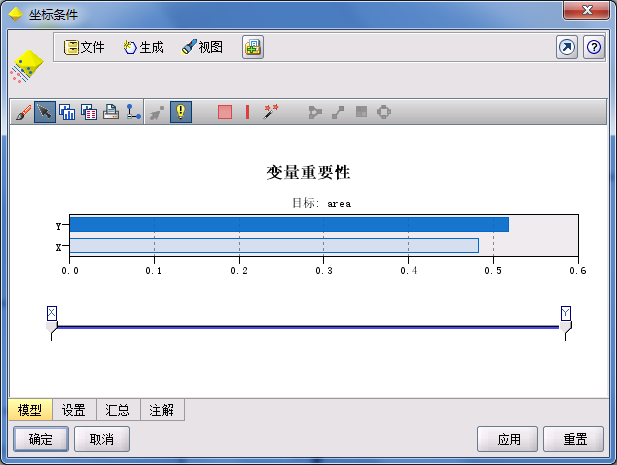


图2.7

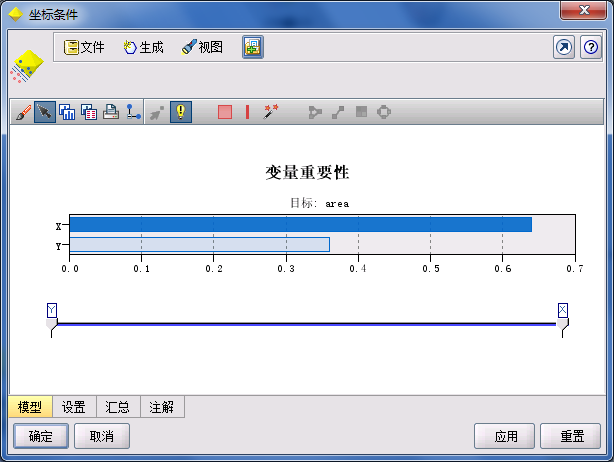


图2.8

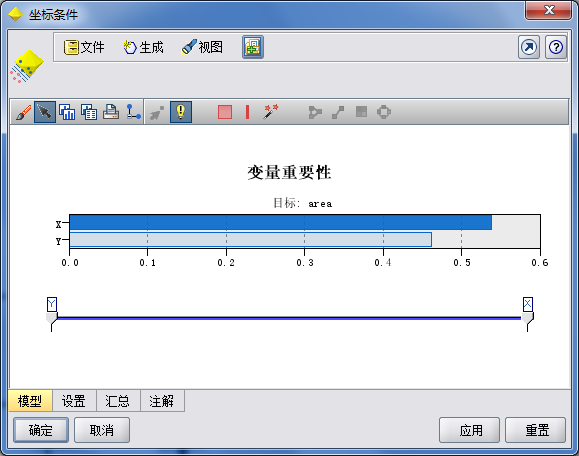


图2.9

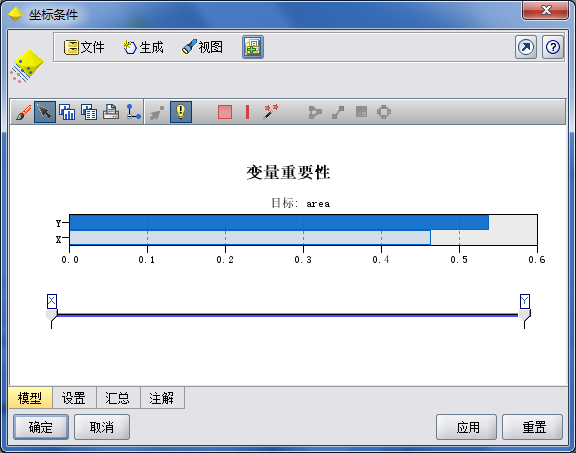


图2.10

上图从上至下分别为“RBF”、“多项式”、“sigmoid”和“线性”内核下的两变量“X”以及“Y”的重要性。

可以看到，尽管分析的结果相同，但各个内核对变量重要性仍然有着泾渭分明的差别。“X”轴坐标是“多项式”和“sigmoid”内核共同的相对重要变量，而“RBF”和“线性”则正好相反，“Y”轴坐标才是相对重要变量。

另外值得注意的是，只有“多项式”内核的比重差距较大，另外三个内核的两属性比率都相近。

## （3）对使用“自然条件”生成的数据分析

执行“分析”节点后，结果如图2.11所示



图2.11

分析结果表明，使用“多项式”内核能更准确地预测到火灾发生区域的面积，尽管其正确率仅有53.97%。其余的内核的预测正确性在53%左右，与“多项式”内核的偏差较小。

双击各个模型，可以看到在不同内核下各个变量的重要性。如图2.12~图2.15所示

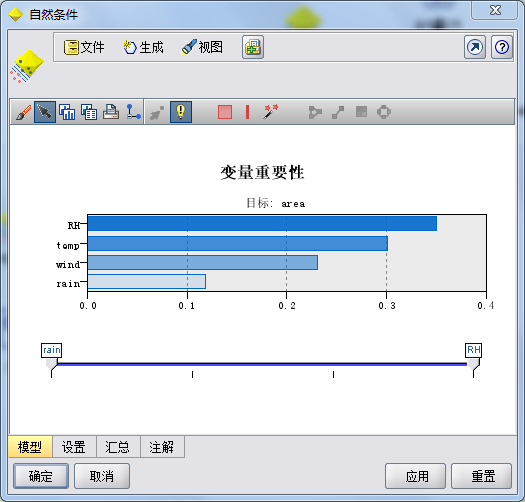


图2.12

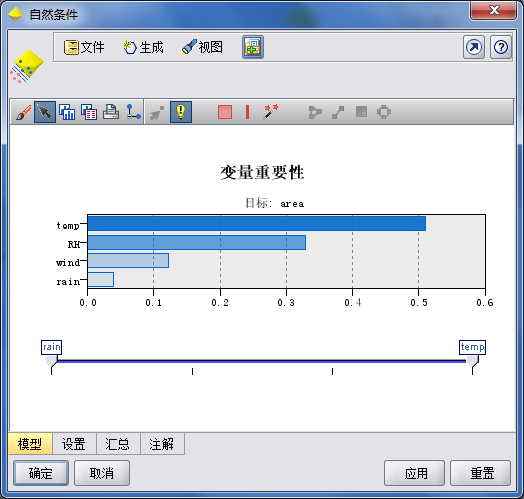


图2.13

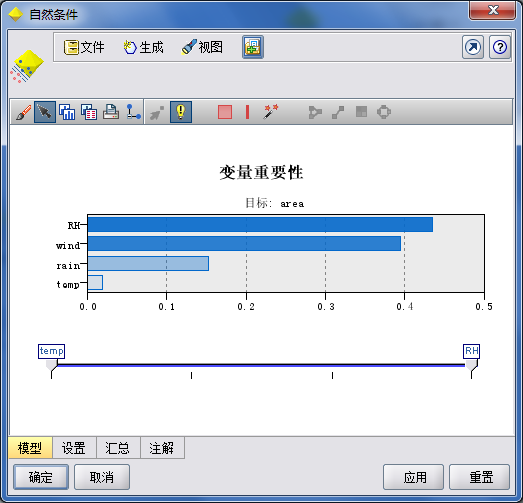


图2.14

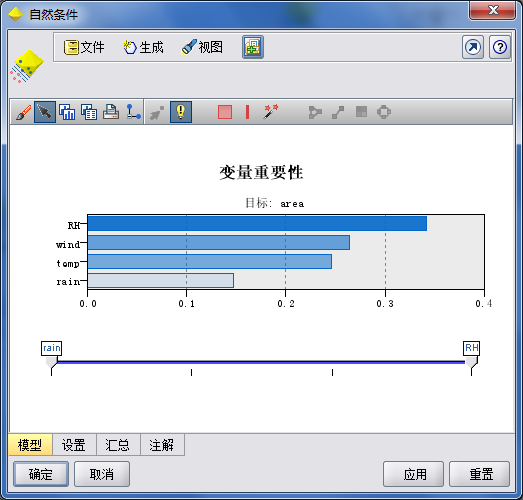


图2.15

上图从上至下分别为“RBF”、“多项式”、“sigmoid”和“线性”内核下的四变量温度“temp”、风速“wind”、相对湿度“RH”以及降雨量“rain”的重要性。

每一个内核对的变量重要性排序都各不相同。值得注意的是，真确率相对最高的“多项式”内核计算出的最重要变量是温度“temp”，且其所占的比率超过了50%。

## （4）对使用“FWI指数条件”生成的数据分析

执行“分析”节点后，结果如图2.16所示



图2.16

巧合的是，使用“FWI指数条件”所得到的结果与使用坐标条件所得到的结果相似：四个内核预测出的结果正确率完全一致，都为53%，处于偏低位置；且四个内核的结果同样完全一致，一致性为100%。

双击各个模型，可以看到在不同内核下各个变量的重要性。如图2.17~图2.20所示

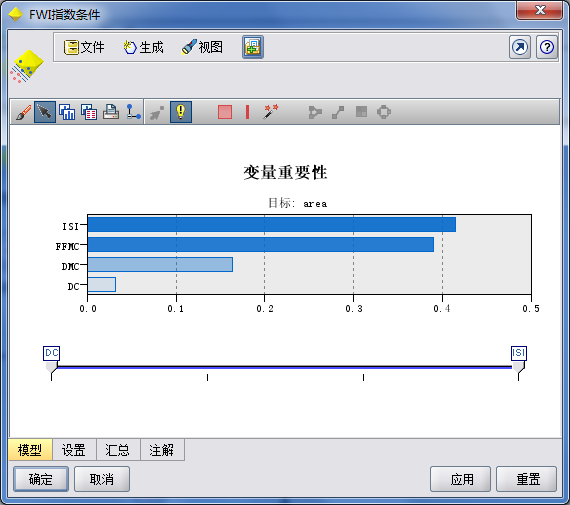


图2.17

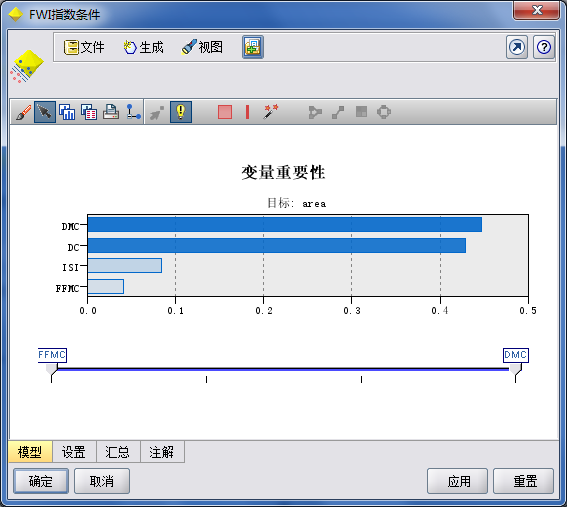


图2.18

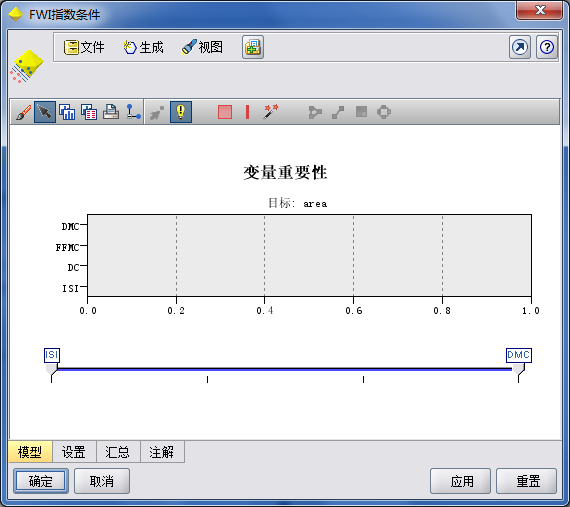


图2.19

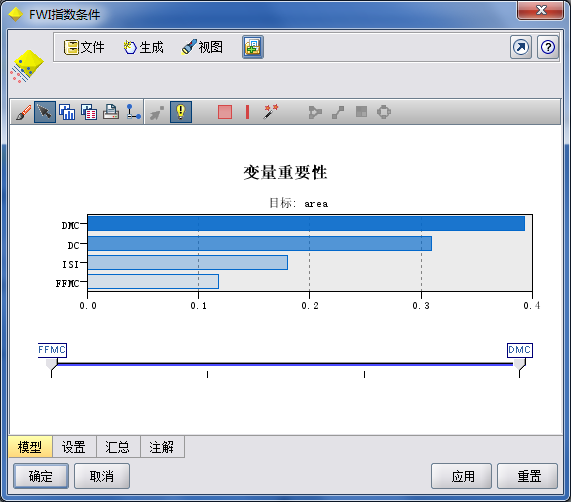


图2.20

上图从上至下分别为“RBF”、“多项式”、“sigmoid”和“线性”内核下的四变量 “DMC”、“DC”、“ISI”以及“FFMC”的重要性。

同样的各有不同，但使用“sigmoid”内核却没办法显示各个变量的重要性。在重复操作了两次后得到的都是相同的结果，这也让我百思不得其解。

## （5）对使用所有属性生成的数据分析

执行“分析”节点后，结果如图2.21所示



图2.21

结果令人震惊：使用“多项式”内核预测出来的结果和现实观测结果的拟合度达到了83.37%。这意味着能通过使用“多项式”内核的SVM对现实数据（方位、日期、FWI指数、自然条件）进行数据挖掘，能预测到超过五分之四的森林火灾区域，能极大地帮助相应防火防灾工作的展开。

而且这还只是保持“多项式”内核默认参数的情况下得到的数据，如果能获得更多的数据，不断对SVM进行调整训练，获得最佳的参数方案设置后，正确率还有可能达到更高的水平。

双击各个模型，可以看到在不同内核下各个变量的重要性。如图2.22~图2.25所示

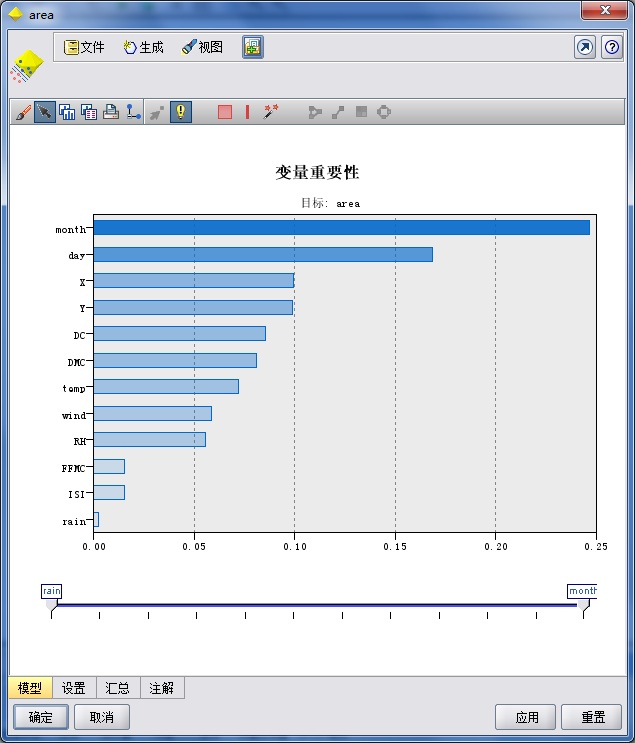


图2.22

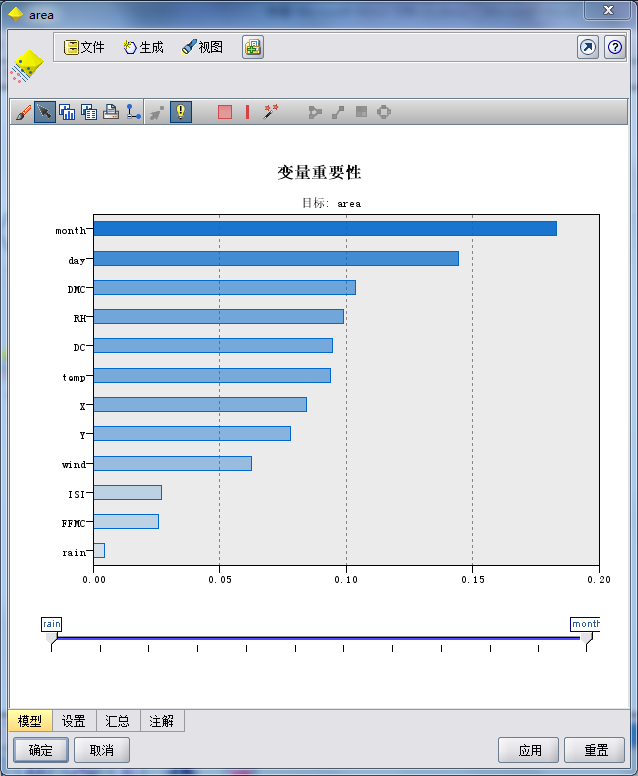


图2.23

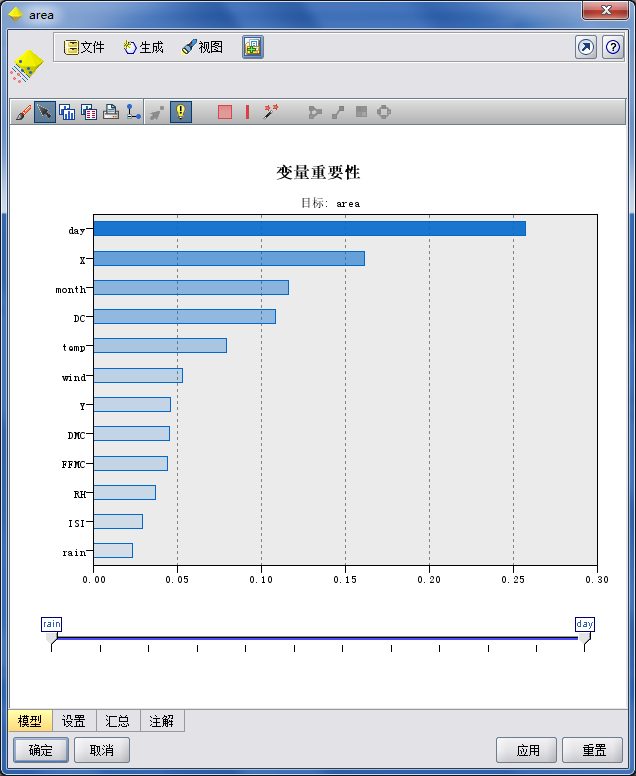


图2.24

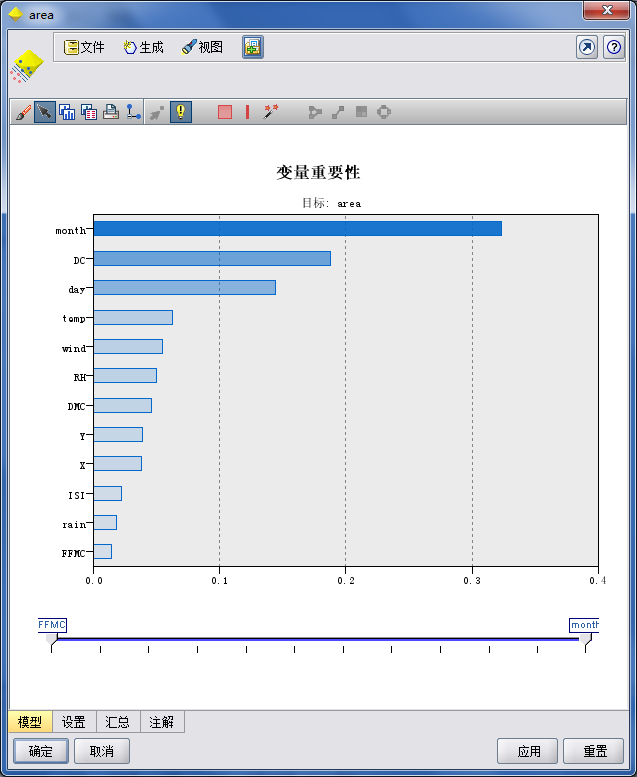


图2.25

上图从上至下分别为“RBF”、“多项式”、“sigmoid”和“线性”内核下除输出属性“area”外其余属性变量的重要性。

相对正确率较高的两个内核“RBF”和“多项式”都把日期属性（“month”、“day”）放在最重要的位置，且降雨量“rain”都处于末位，是最为不重要的因素。这与我们的认知有着微妙的出入。

让我们继续关注“多项式”内核的结论。从FWI指数组成因素来考虑（图2.26），处于重要性第3~6位的因素“DMC”、相对湿度“RH”、“DC”、温度“temp”，其实同样主要与相对湿度、温度相关，因此可以认为这两个属性是次重要的；而因为同样被包括降雨量“rain”本身的重要性较低，所以可以排除。

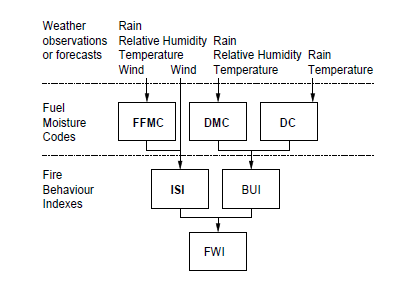


图2.26

接下来再让我们看看具体数据。双击“表”节点，点击“执行”按钮，可以看到具体的表单结果。如图2.27所示

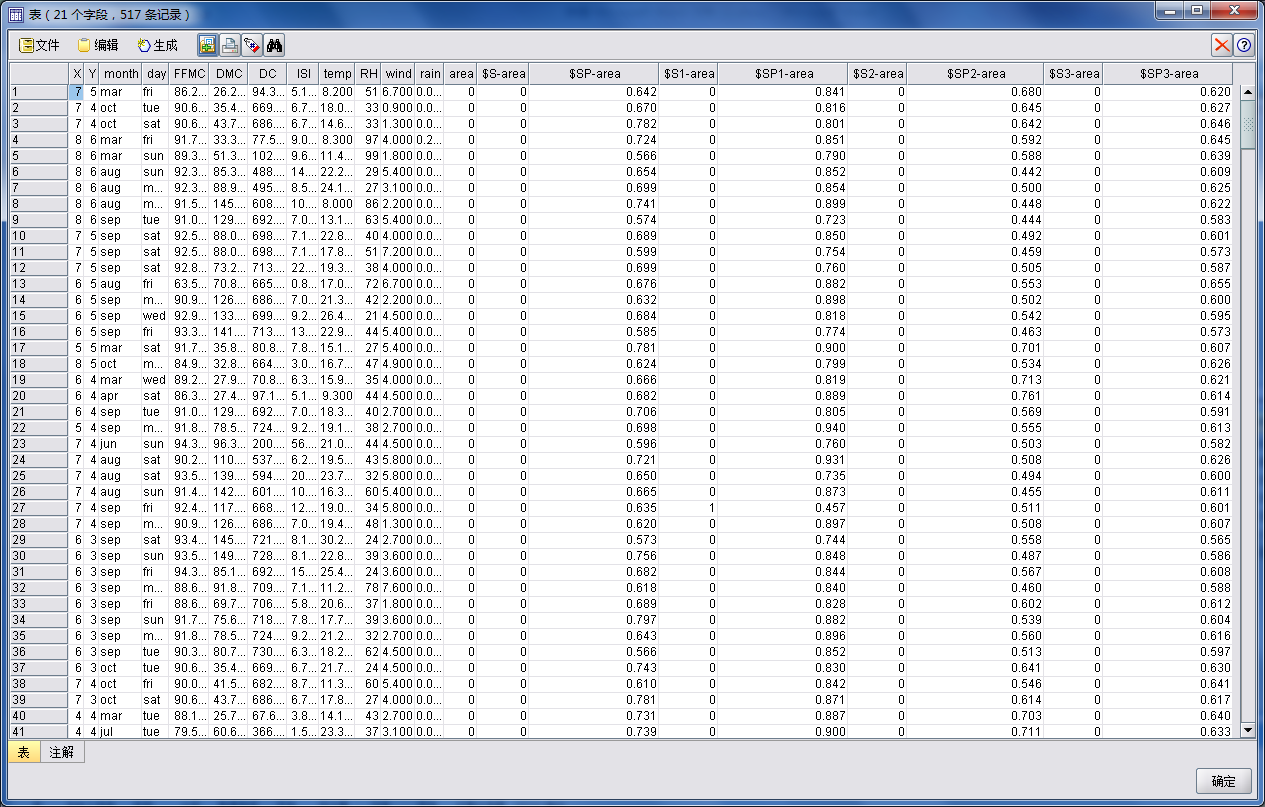


图2.27

因为数据过于复杂，我们仅关注正确率较高的“多项式”内核的结果。在第二个模型处连接一个“表”节点，执行后如图2.28所示。

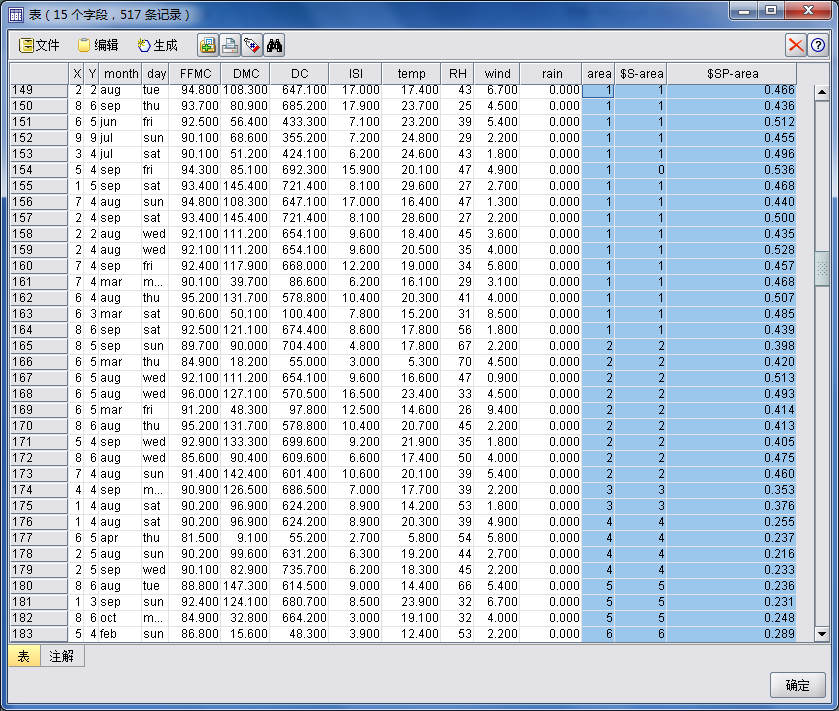


图2.28

图中“$S-area”列为预测出的火灾面积，“area”列为实际火灾面积。观察图2.28中的蓝色部分，“多项式”内核SVM能十分准确地预测到小面积火灾的情况。

这点和原论文中提到的情况相一致。

然而，当火灾规模增大时，SVM预测的结果则出现明显错误。如图2.29所示。

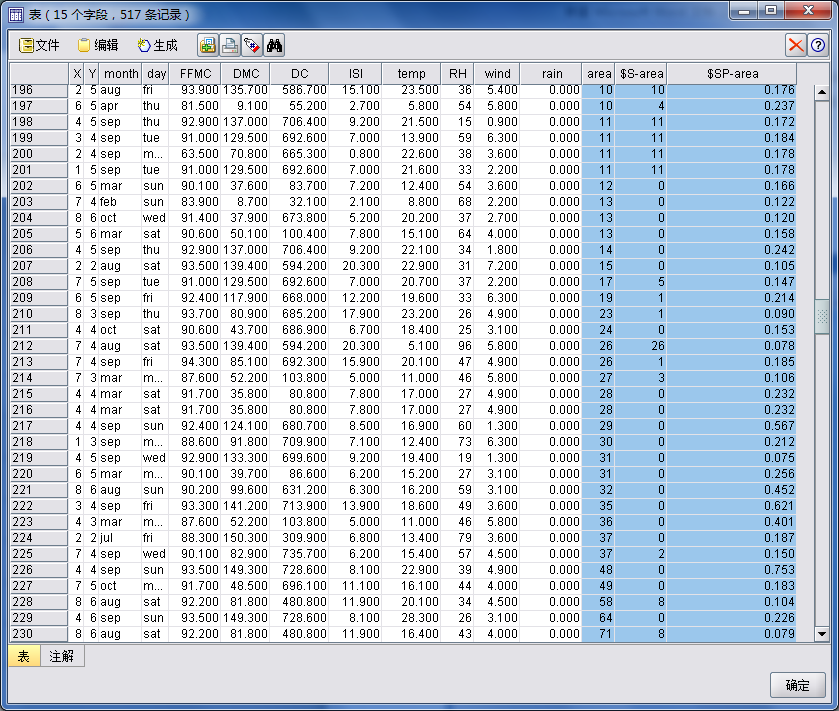


图2.29

蓝色部分的实际火灾面积逐渐增大，但SVM并未能计算出相应的数据。

同样地，当超大规模火灾发生时，SVM也没办法预测。如图2.30所示。

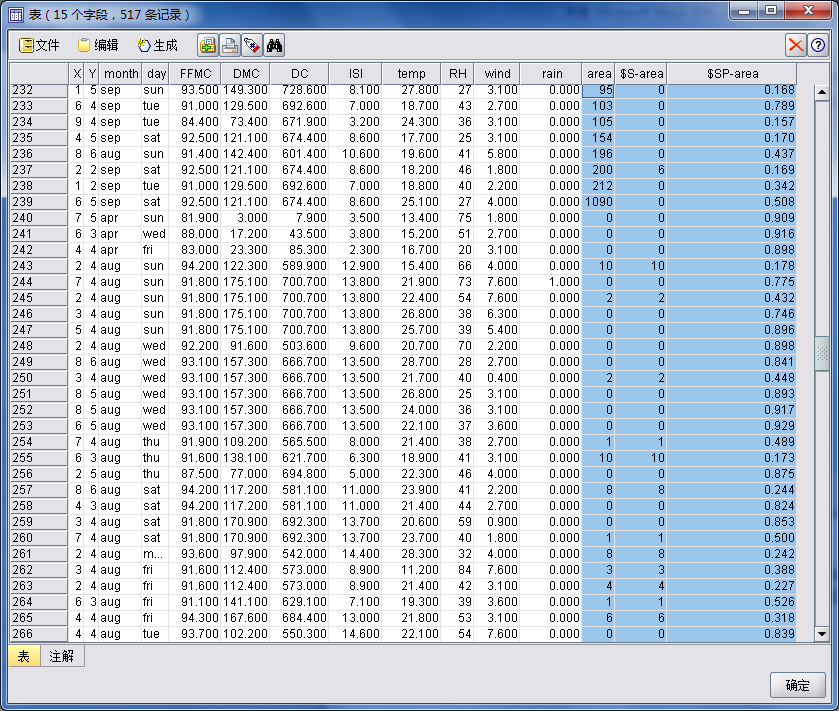


图2.30

上图中有几场“area”过百甚至上千的火灾发生，但SVM都没有相应预测。

在这里大胆推测，时间恰逢8、9月，这几场火灾很可能并非天然引起，而是有人为失误造成的；亦或是这些火灾在发生后因为其他因素迅速蔓延而造成更大面积受灾。因此SVM并未预测出来。

# 四、结论

经过以上建模、分析、数据挖掘的步骤之后，可以得到以下结论：

使用“多项式”内核的支持向量机SVM能较为准确地预测出相应地区是否发生火灾及其受灾面积，且准确率在参数足够的前提下能达到惊人的83%，而即使数据不足或较为片面，预测准确率也有53%~54%之间；该SVM对小面积火灾有较好的预测能力，而对大规模或超大规模火灾的预测能力较弱。