国际预测杂志30 (2014)395-401





目录可以在ScienceDirect上找到

国际预报杂志

杂志主页:www.elsevier.com/locate/ijforecast

基于特征工程的GEFCom 2012年风电功率预测方法



卢卡斯席尔瓦

*巴西贝洛奥里藏特的DTI Sistemas*

条信息

*关键词:*

GEFCom

特征工程梯度促进决策树线性回归

机器学习

时间序列

摘要

本文详细介绍了Leustagos团队对GEFCom 2012年风电预测轨迹的方法。**这项任务是提前48小时预测7个风电场每小时的发电量。**这个问题通过提取时间和天气相关的特征来解决，这些特征被用来**建立梯度增强的决策树和线性回归模型**。这种方法在公共和私人排行榜上都获得了第一名。

©2013国际预测机构协会。Elsevier B.V.版权所有

**1.介绍**

“GEFCom 2012-Wind Forecasting”比赛对7个风电场逐时风力发电预测提出了挑战。1 提供了一个数据集，其中包含这些风电场的历史功率测量数据，以及这些风电场级别的风力成分的气象预报。Hong, Pinson和Fan(2013)提供了数据集的详细描述。

*1.1。挑战*

这个任务中的一个大挑战是处理数据集的时间序列特性。由于**没有使用时间序列特定的模型**，所以必须时刻记住它。

另一个困难是创建有效的特性。在解决监督学习问题时，特征创建是最重要的步骤之一。为了做到这一点，从数据集派生了许多特性。特性创建步骤有两个主要的指导原则:

1.以常数、风力强度、风向和空气密度(代替)为基础，对风力发电方程进行建模。这些特征代表了风车的行为。

*电子邮件地址:lucas.eustaquio@gmail.com。*1)http://www.kaggle.com/c/GEF2012-wind-forecasting。

2.发现T时刻风力发电与T±n之间的关系。这里的目的是反映数据集的时间序列性质，因为建模技术不是特定于时间序列的。

*1.2。数据集*

可以看到,提供的数据(见Hong et al ., 2013)只包括一系列的每小时每个农场,风力发电值和预测风力强度和方向,每12 发布一次。这样做是为了模拟真实的操作条件下,不包括所有所需的信息,如预测温度和空气密度。因此，为了弥补这些缺失的信息，创建了许多代理功能，稍后将对此进行解释。此外，一个重要的步骤是仅使用the training period来构建一致的验证集。这个验证集允许构建一个数据不过拟合的模型。

*1.3。技术*

在这项工作中，使用了三种主要的机器学习技术，它们都在R统计环境中。表1简要描述了每种算法的使用方式。

*1.3.1。线性回归*

本文采用线性回归方法作为结合风力的预处理步骤

0169-2070/$ -见©2013国际预测机构爱思唯尔版权所有http://dx.doi.org/10.1016/j.ijforecast.2013.07.007

*L. Silva /国际预测期刊30 (2014)395-401*

396

**表1使用的技术。**

|  |  |
| --- | --- |
| 技术 | 用于 |
| 多元线性 | 风力和风向的影响 |
| 回归(盖尔 | 每个农场的组件 |
| 2003) | 系综 |
|  | 后期处理平滑预测(高 |
|  | 频率滤波器) |
| k - meansa | 相似模型来检测农场和整体 |
|  | 行为 |
|  | 模型由农场 |
| GBMb | 按时间间隔模型 |
|  | 整体模型 |

a R K-Means软件包，http://stat.ethz。ch / R-manual / R-devel /图书馆/统计/ html / kmeans.html。

b gbm:广义强化回归模型。http://cran.r-project。org/web/packages/gbm/index.html。

和一个单一特征的方向分量。它也被用来结合GBM训练的模型，最后作为一个后处理步骤来平滑预测值的移动平均值。

线性回归试图通过将线性方程拟合到观测数据上来建立两个或多个解释变量和一个响应变量之间的关系模型。2 (1)表示形式定义:



(1)

地点:

•i = 1，…， n为训练实例索引;

•βare回归系数;*p*

•xare功能;*ip* 和

•ε是残差。

*1.3.2。k - means*

在这项工作中，K-Means被用来创建相似特征。这些特性的目标是对应该具有类似行为的实例进行集群。

K-Means是一种简单的聚类分析学习算法。k - means算法的目标是将n个实体最优划分为k组，使组内成员与其对应的代表组的质心之间的总距离最小。K-Means算法的伪代码如图1所示。3

*1.3.GBM—generalized boosted models*梯度提升树算法

中间模型采用GBM进行训练。我们使用了高斯分布函数(Ridgeway, 2013)的R环境中的GBM包。

R GBM包实现了模型的增强，这是统计中常用的。GBM算法的伪代码如图2所示(参见Friedman, 2001)。在图2中,

•F (x)是要最小化的函数;

L(y, p)是损失函数;

•y是输出;

•p是预测;

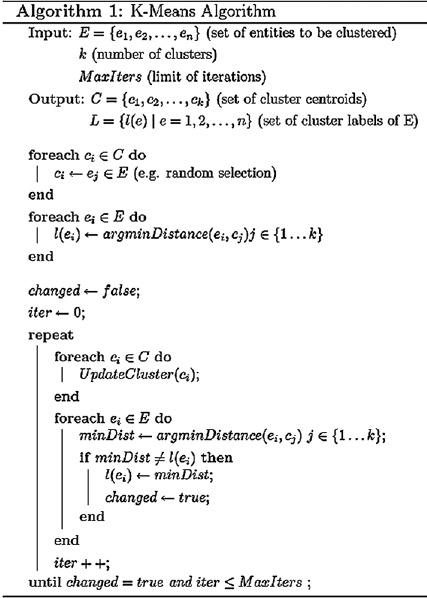
•M是树的数量;和

•h (x);*a)通常是“弱学习者”或“基础学习者”*

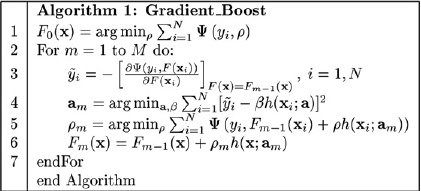
树。

2 http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm。

R/Clustering/K-Means中的3种数据挖掘算法。http://en。wikibooks.org/wiki/Data\_Mining\_Algorithms\_In\_R/Clustering/K-Means。



**图1所示。**k - means算法。



**图2所示。**增强算法。

*1.4。方法概述*

该方法的基本框架如图3所示。首先，在预处理步骤中，创建特性。接下来，利用处理后的数据对三种类型的模型进行训练:

•为每个农场培训的模型;

•为每个预测时间间隔(提前1-3小时，提前4-6小时，……，前进45至48小时);和

•整体模型的训练不需要分割样本(交叉验证除外)。

*1.5。论文组织机构*

本文详细介绍了所采用的特征和模型。第2节描述了所采取的方法，解释了特性创建和建模技术

*L. Silva /国际预测期刊30 (2014)395-401*

397

预处理：原始数据——特征提取——中间数据——相似特征，k-means分簇——预处理数据

模型训练：用GBM算法处理每个风场的数据——特定时间间隔的模型——全数据实例模型

后处理：模型集——模型平滑处理——预测

**图3所示。**基本方法框架。



**图4所示。**培训/验证分裂。

使用，并显示如何分割数据集以进行模型训练。第3节给出了结果并讨论了解决方案的各个方面，第4节给出了结论。

**2.方法**

*2.1。验证设置*

该方法的第一步是建立一个一致的验证集。它建立的目的是试图复制在评估期间发现的填充和缺失点之间的结构。为了复制这种结构，每个训练和评估周期被分割为312组，每组84个连续小时(36 + 48)，这些集被用来定义5倍验证分裂。每个训练/验证折叠将整个84小时的桶分配给训练或验证。

图4显示了如何进行集合分配。可以看到奇数集从00:00开始，偶数集从12:00开始。这也发生在评估阶段的缺失点上。至于5倍分裂，它是这样做的:

•验证折叠:折叠1包含集1、6、11等，作为验证集，其余集作为训练集。折叠2包含集2、7、12等，用于验证，其他集用于训练。3、4和5的折叠也进行了类似的分割。

•测试折叠:所有可用的填充数据都用于

建立模型以预测缺失的数据。

在这个场景中，需要对每个模型进行6次培训:一次针对每个验证折叠，以便调整模型参数和培训集成，总共5次，另一次使用所有可用的数据来预测排行榜的值。这个过程允许我们在大多数情况下产生一致的结果，这意味着在可用数据中发现的改进几乎总是泛化到缺失的未知数据(排行榜数据)。

在这个问题中，数据中有一个时间组件。因此，基于实例的随机交叉验证往往高估了模型的性能，因为它隐含地向模型提供了关于每个48小时点周围环境的更多信息。这在早期的基准测试中得到了验证，这也是使用此验证方案的原因。？？？

*2.2。功能创建*

创建特性的第一步是创建训练实例。所提供的数据集没有显式地包含这些。要创建这些实例，文件需要“训练”。csv”和“基准。处理“csv”，为每个日期和每个farm创建一个条目。此时还创建了小时、月和年功能。图5显示了生成的数据是什么样的，表2解释了每个特征。

下一步是处理预测。在这个步骤中使用了每个农场的每个预报。每一次,

*L. Silva /国际预测期刊30 (2014)395-401*

398

**表2**

在第一步中创建的功能。

名称类型

描述

*日期日期和时间农场分类(1-7)wp数字*

*时间分类(01-24)月份分类(01-12)年分类(2009-2012)*

这是每一次风力测量的日期和时间;它主要用于连接表示每个farm的不同功能表

风电功率值有待预测

每小时风力测量

每月风力测量一次

每一年的风力测量

**表3预测特征。**

名称类型

描述

*日期日期及时间*

*农场分类(值1-7)开始日期和时间*

*dist Categorical(取值范围01-48)改为Categorical(取值范围00和12)设置Categorical(取值范围1-312)*

*ws数值*

*wd数值*

*wd\_cut分类的([0,30]…*(330,360)

目标预测日期

代表每一个农场

天气预报发布的日期

开始和日期的时间差异;预报距离

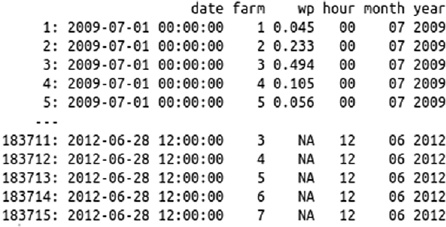
表示预报开始时间

每组代表36小时+ 48小时;利用该变量对预测的风力进行交叉验证训练

预测风向

风向的分类版本

|  |  |  |
| --- | --- | --- |
| **表4** |  |  |
| 历史特征。 |  |  |
| 的名字 | 类型 | 描述 |
| *开始* | 日期和时间 | 天气预报发布的日期 |
| *农场* | 分类(1 - 7) | 风力发电场 |
| *经销*  *wp\_hnXX (XX = [01, 06])* | 分类(01-48)  数值 | 开工时间和日期的差异;预报距离  X以前的已知值:01是开始日期和时间的场值，02是开始 |
|  |  | 时间-1小时，等等 |

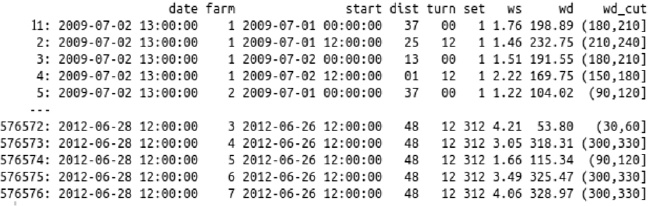


**图5所示。**实例创建，步骤1。

邮票有四个不同的预测，除了在排行榜集，其中只有预测时可用的预测。在此步骤中还导出了其他一些特性。图6和表3展示了数据和特征的含义。

接下来创建的是历史特性。对于每个预报，使用预报开始前的6个值。在测试集中，许多值是未知的，这些实例在部分训练期间被丢弃。出于验证的目的，为每一个单独的实例计算那些以前的值是有用的，即使是在训练集中，因为它们在实际情况中是已知的。历史特征见表4。

下一个功能旨在统一的风力强度和风向在一个单一的数值变量。这种统一后来在相似模型的建立和移动平均线预测的修正中得到了利用。对于风力发电来说，不仅要考虑风力的强度，还要考虑风力的方向。风车在不同的方向上有不同的效率水平。风能转换与常数和ws(三次风强度)成正比。3 它也取决于空气密度。



**图6所示。**使用预测创建的功能。

*L. Silva /国际预测期刊30 (2014)395-401*

399

**表5**

特点统一的ws和角度。

|  |  |  |
| --- | --- | --- |
| 的名字 | 类型 | 描述 |
| *日期* | 日期和时间 | 目标预测日期 |
| *农场* | 分类(1 - 7) | 风力发电场 |
| *经销* | 分类(01-48) | 开始和日期的时间差异 |
| *wp* | 数字 | 风力发电的预测 |
| *ws.angle* | 数字 | ws和wd对日期的影响 |
| *ws.angle.p3* | 数字 | *ws。日期的角度为- 3小时。如果没有找到，默认为最近的ws.角度* |
| *ws.angle.p2* | 数字 | *ws。日期的角度为- 2小时。如果没有找到，默认为最近的ws.角度* |
| *ws.angle.p1* | 数字 | *ws。角为日期- 1小时。如果没有找到。*默认为最近的ws.s角 |
| *ws.angle.n3* | 数字 | *ws。角度为日期+3小时。如果没有找到，默认为最近的ws.角度* |
| *ws.angle.n2* | 数字 | *ws。角度为日期+2小时。如果没有找到，默认为最近的ws.角度* |
| *ws.angle.n1* | 数字 | *ws。日期的角度+1小时。如果没有找到，默认为最近的ws.角度* |

**表6**

相似的特性。

名称类型

描述

*日期日期和时间农场分类(1-7)区分类(01-48)pos分类(1-12)开始分类*

*clust。farm Categorical(42 = 7×6)clust Categorical (1-24)*

目标预测日期

风力发电场

开始和日期的时间差异;预报距离

预测在12个连续数值内的位置。与(dist - 1)除以12的余数相同

小时的第一个值的12小时序列

12个连续的小时分配到的集群

一般的集群

该特征生成包括为每个农场构建一个简单的回归模型。为了避免过度拟合，使用了5倍交叉验证。交叉验证中一个非常重要的问题是它必须使用前面计算的集合特性。每个集合必须完全包含或完全排除。基于实例的抽样将导致过度拟合，因为回归将推断出每个“36小时+ 48小时”周期的相关信息。

这次训练使用的R公式是Eq.(2)，但是wd\_cut与之前的描述并不完全相同。对于这个训练，它是一个分类值，以8度的间隔分割wd(这种分割是迭代发现的)。这个训练的结果叫做ws.angle。然后用这些值构建另一个功能图。在这个映射中，ws的前三个值和后三个值。*角度预测也包括在内。*结果如表5所示。



(2)

最后一个特征是相似特征。由于预测值是一个时间序列，连续预测值之间存在一定的相关性。让模型了解这一点的一种方法是在最后一步中完成的。包含上一个和下一个期望值将使模型了解T±n和T之间的某种程度的依赖关系。另一种可能的方法是通过相似性:为一组预测提取农场行为。

利用K-Means计算相似特征。对于每个12小时的预测期([1 h - 12 h]， [13 h - 24 h]， [25 h - 36 h]， [37 h - 48 h])，构建一个实例。为每个农场计算了6个集群，为包括所有农场的数据集计算了24个集群。结果见表6。

最后的特征是通过加入所有已经产生的特征得到的。要连接这些特性，只需匹配公共列并对其余列进行笛卡尔积。以表2和表6为例

例如，我们可以看到它们的共同特征是日期和农场。要联接这些表，请根据字段的值匹配它们。在这种情况下，对于表2中的每个匹配值，将在表6中找到4条记录，因为每个wp预测都发布了4次。连接结果将在表6(笛卡尔积)的每个匹配中附加表2的特性。

有些特征指的是将来的值(日期+ n小时)。需要指出的是，这些值都是预测值，因此这里的整个特征工程只使用在预测时实际可用的数据。

*2.3。建模技术*

所有的模型都使用相同的5倍验证方案进行训练。这种训练的一个非常重要的特点是，验证集的构建是为了将36小时+48小时的块作为一个整体来包含或删除，就像反复说的那样(这真的很重要)。逐个实例抽样既不能反映实际情况，也不能反映测试集的实际情况。通过这样做，可以获得一致的验证集。显著的改进几乎总是导致排行榜的改进。

如前所述，这些模型被训练有两个目的。

•基于常数、风力强度、风向和空气密度，对风力发电方程进行建模。最后一个是由季节特征代替的，比如时间、时间和年份。这些因素表达了风车的行为。

•为了发现T点风力发电与T±n之间的关系，这里的目的是介绍模型的时间序列性质。

*L. Silva /国际预测期刊30 (2014)395-401*

400

为了实现这一点，对以下类型的模型进行了培训。

1.高斯分布的GBM和Eq.(3).这个模型的目的是研究惯性行为，发现空气密度和其他偶然环境影响的影响。惯性行为是通过ws习得的。*角度(T±3)。利用时间、月份和年份的分类变量来推断偶尔的影响。*使用温度和/或湿度测量可能会在真实环境中获得更好的性能。时间和月份与昼夜和季节的关系有关。farm值将处理任何farm细节，而dist将处理与预测距离相关的错误(距离越长，错误越大)。

2.GBM具有高斯分布和Eq.(4).如前所述，但包括最新已知的风电值。这个值可以提供一些有价值的上下文信息和时间序列的起点。这两个模型存在的原因是为了训练这个模型而丢弃了测试期间的许多实例(它们缺少wp\_hn01)。在现实生活中，这是不必要的，因为这些信息总是可用的。

3.为了研究环境的影响，建立了高斯分布的GBM和Eq.(5)。将这些特性与之前的特性混合使用并没有显示出任何改进;然而，单独的训练是有效果的。

*wp∼ws + farm + dist + ws。角+ ws.angle。p1 + ws.angle。p2 + ws.angle。p3 + ws.angle。n1 + ws.angle。n2 + ws.angle。n3 +小时+月(3)*

*wp∼ws + farm + ws。角+ ws.angle。p1 + ws.angle。p2 + ws.angle。p3 + ws.angle。n1 + ws.angle。n2 + ws.angle。n3 +小时*

+月+年+ dist + wp\_hn01

(4)

*wp∼farm + dist + wp\_hn02 + wp\_hn03*

+ wp\_hn04 +小时+月+ clust。*farm + clust + begin + clust.pos。*

(5)

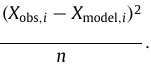
所有提到的模型都被训练成三种风格:每个农场，每个集群距离((0,3)，(3,6)…(45,48])，和一个没有区别

(图3)。

最后一个模型是前面提到的模型(9个模型)的线性集合。为了定义集成权重，使用了一个简单的线性回归优化RMSE (Eq.(6))。它被训练使用交叉验证的预测模型。它还使用farm和dist作为特征，因为它们被用来进行分割，并且与每个模型的优缺点有关系

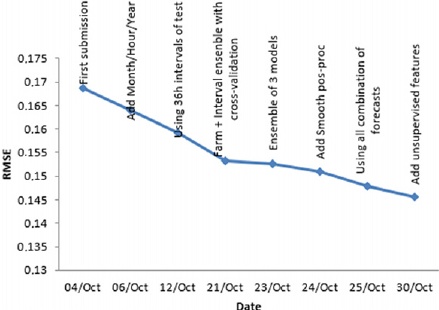
 n 

 i=1



RMSE =

(6)



**图7所示。**分数进化。

**表7最终排名。**

|  |  |  |  |
| --- | --- | --- | --- |
| 排名 | 团队 | 公共的分数 | 私人的分数 |
| 1 | Leustagos | 0.14574 | 0.14567 |
| 2 | DuckTile | 0.14872 | 0.14720 |
| 3. | 吉尔伯托Titericz小 | 0.14792 | 0.14822 |
| 4 | Stefan Henß | 0.14684 | 0.14854 |
| 5 | Mz | 0.14804 | 0.14916 |

之后，又训练了另一个模型，这次是平滑模型。它采用集成的结果，并对预测结果进行T±n (n =−3到3)内插。当使用ws时，该模型几乎无用。*角度(T±n)在之前的预测中有所提及，但预测结果确实有所提高，这对比赛有一定的帮助。*

**3.结果与讨论**

*3.1。结果*

评价指标为RMSE (Eq.(6))，即均方根误差。这个度规的值随着模型精度的上升而下降。图7显示了该解决方案中里程碑的分数改进情况。可以看出，第一个模型仅包含预测和之前已知的值，得分为0.1685 RMSE。然后，通过添加一些季节特征(小时、月、年)，将其降低到0.16393。然后将测试期可用的36小时间隔纳入set中，再次得到改善。改变交叉验证方案，使用36 + 48进行验证分割后，分数跃升至0.15315。接下来，添加总体模型和最终预测的平滑模型，得到0.15103。得分上一个大的飞跃是包含了所有日期和预测的组合，每个已知实例有四个预测，得分为0.14779。最后，加入无监督的集群特性，最终的最佳得分为0.14567。

查看表7中的最终结果，我们可以看到整个过程概括得很好，在公共和私人排行榜中都产生了一致的结果。

*L. Silva /国际预测期刊30 (2014)395-401*

401

*3.2。讨论*

在这里，我们提供了一个有效的方法来预测风力发电在七个不同的风电场。该解决方案由两部分组成:第一，提取特性;第二，为了创建模型，对数据应用不同的聚合模式。最后，这种方法使用了许多模型，比实际情况所需的模型要多，但这只是因为采用这种方法是为了赢得竞争，在必须考虑每一点性能时，忽略了构建它所需的计算开销。此外，一些在实际操作中已知的数据被隐藏起来，增加了训练过程的复杂性。比赛结束后，用模型进行了一些测试，得到了一个更简单的版本，这个版本的性能几乎一样，但复杂程度要低得多。在该方法中，只使用了两种总体GBM回归，抛弃了用来合并风力和风向的特征。通过加入更多的特征，如温度和湿度预报，可能会达到更高的精度。

**4.结论**

我们已经证明，利用一些著名的算法(GBM、线性回归和K- Means)的巧妙结合，有可能实现高水平的风电功率预测精度。最重要的步骤是为培训和创建良好的特性选择可靠的验证方案。结果的关键是永远不要忘记数据集的时间序列特性。许多特性、一些后处理步骤和验证方案都在设计中使用了这个特性，这是这种方法成功的原因。提出的解决方案有些复杂，但很精确。一些测试也表明，它可以大大简化，只有边际性能下降。

**参考文献**

弗里德曼，J. H.(2001)。*贪心函数近似:一个梯度增强机。*那个宿舍叫赖茨讲座。

盖耶，C. J.(2003)。R中的广义线性模型。

http://www.stat.umn.edu/geyer/5931/mle/glm.pdf。

在香港,T。mike vanderboegh, P。范(2013)。2012年全球能源预测大赛。2013年GEFCom。

山脊路,g (2013)。gbm:广义强化回归模型。arxiv:cran.r-project.org/package=gbm。