无锡横新电站预测简报

发电预测（没有准确的预测不能并入公网），发电板检测（正常预测）

1、数据

无锡横新电站的数据分为四列：
第一列是电站发电量的采样时间。间隔并不等距，但大多数是 4 分钟或 5 分钟；
第二列是到该时间为止的累计发电量；
第三列是日辐射测量时间点，时间间隔也不想等；
第四列是该时间点的辐射值。

2、数据预处理

数据初步清洗：日发电量有些值远超正常值，我将这些值用前后比较正常的观测值做线性插值这些异常值。

由于发电量在每日内是累计值，为避免伪回归，首先将发电量转为单位时间发电量(姑且认为是发电功率)。做法是将日发电量做差分，然后除以时间差(单位是分钟).。

数据清洗：数据中所有的辐射量为 0，单位发电量为 0 的观测值属于平凡样本。首先去掉。这些值虽然都是正常的，但对我们分析没有任何帮助。进一步观察，在辐射量小于 10 的情况下，只有不到 1/3 的对应样本的发电功率大于 0，而且值都比较小。为此，我们将所有辐射量小于 10 的样本清除。这些样本也是正常的，但是会干扰我们的分析。另外，有些样本发电功率明显大于正常值，这些样本虽然只有十几个，它们也会影响分析结果。所以我们也把它们剔除。最后剩余供分析的样本约为总样本量的 47%。

3、样本标记

b spline 拟合曲线
建立（光伏量与发电量之间的函数）
根据行业经验，在正常发电量均值以下 30%可以认为异常。经过我们的分析。在上述可用于分析的样本中若以均值 30%以下为异常样本，比例过大。传统上，一般把 5%, 1%或 0.1%作为一个样本中异常的比例。我们采用常用的 5%作为标准。在这一标准下，以低于（）发电功率 55%作为异常，数据中大致有 5%的异常值。因此，我们用这一标准进行数据标记。正常数据标记为 1，异常数据标记为 0（习惯上，SVM 中负样本应该标记为-1。但为了后续分析比较便利起见，我们将它们标记为 0）。

由于数据限制，我们考虑以辐射值为解释变量，发电功率为被解释变量进行建模。通过数据散点图，我们不难发现

1、 数据非线性情况明显
2、 数据方差呈扩散趋势
为便于建模，我们按日发电量的时间戳匹配最近时间的日辐射值。

考虑到发电量受天气因素影响，我们收集了无锡对应日期的气象数据，包括：最高气温，最低气温，风速和空气质量 AQI。由于只有每日数据，并没有小时气象数据，因此我们匹配日期，将这些气象数据一并纳入建模范畴。

4、One-class 建模

由于我们将采用 one-class 方法进行数据分类，故仅使用正样本建模。我们将所有有效正样本按 70%:15%:15%用于建模，校正（用来检测损失函数值是否减小）和检验。将所有负样本随机等分为两部分。第一部分与用于校正的正样本合并为校正样本。另外一部分与用于检验的正样本合并为检验样本。

训练样本使用 70%的正样本，优化目标是训练模型使得其应用于校正样本时，其错分样本数最少。我们使用常用的 RBF 为核（高斯核）。分析结果用检验样本最后评判模型的泛化能力。具体结果如下：

真实标签 / 计算标签	校正样本		检验样本	
	0	**1**	**0**	**1**
0	1	16	12	18
1	263	1501	253	1499

注意到 one-class 分类算法中的参数均由优化算法自动获得（预设）。在错分样本数最少的目标下，由上表的结果，one-class 分类在本应用中，控制虚警率较好，但同时几乎没有识别负样本的能力。

另外，超球算法一般是基于线性假设的 one-class 算法。若以线性 one-class 进行分类，效果如下：

真实标签 / 计算标签	校正样本		检验样本	
	0	**1**	**0**	**1**
0	113	457	32	491
1	151	1060	233	1027

其控制虚警率和负样本识别，从泛化检验的结果可知，这两项都较差。比核技巧得到的结果更差。

5、预测建模

根据第四部分所得结果，我们得到了分类正样本。分类正样本是我们在实际工作中由算法得到的正常情况（在实践中，用 one-class 的场合中，可观测得到的负样本一般难以获得或样本极少，我们只能认为，使用算法标记的正样本为正常情况的值）。基于这些算法标记的正样本，我们利用辐射值，天气状况来预测发电功率。我们以辐射值构造 b-样条，结合匹配时间的气温，风速和 AQI 作为解释变量，以发射功率作为被解释变量构造 b-样条回归模型。结果如图 1 所示。结果对应的 adjusted-R^2为 0.78（用于评估回归模型的效果）。效果尚可。

以上均为估计发电功率的模型。为估计日发电量，只需将上述功率在给定的日辐射强度条件下按时间累计即可得到。

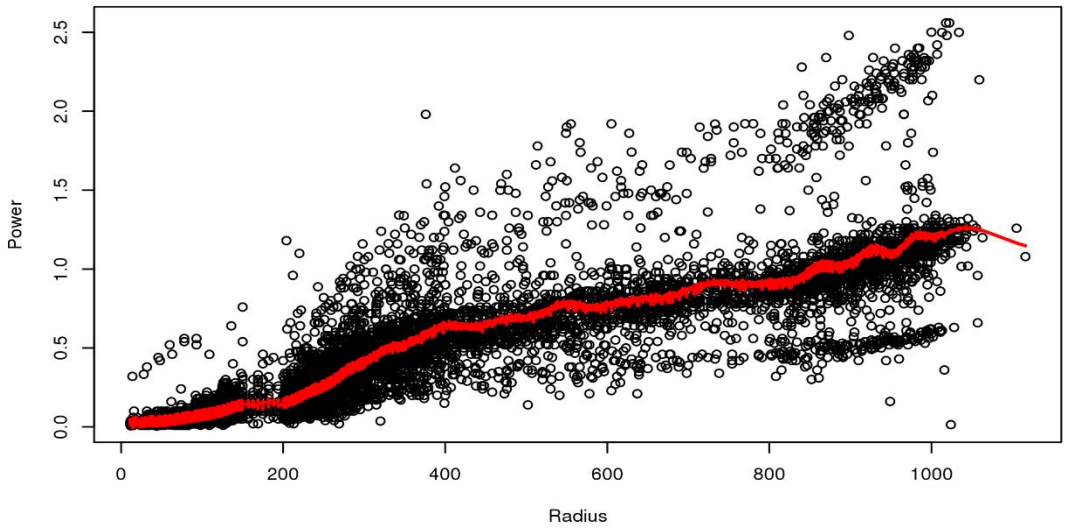

图 1 无锡横新光伏发电功率（红色为拟合曲线）