# 基于数据挖掘的日负荷曲线预测与修正①

杨 莉[1], 李鹏举[2]

[1](上海海事大学 信息工程学院, 上海 201306)
[2](中国矿业大学(北京) 机电与信息工程学院, 北京 100083)

**摘 要**: 分析了传统负荷预测方法的缺点, 提出了一种基于数据挖掘技术的负荷预测方法. 利用决策树算法进行负荷预测, 根据预测结果找出负荷不正常点. 依靠关联规则算法, 对不正常负荷进行修正, 从而使预测结果更加精确.

**关键词**: 数据挖掘; 关联规则; 决策树算法; 日负荷曲线

## Forecasting and Modification of Daily Load Curve Based on Data Mining

YANG Li[1], LI Peng-Ju[2]

[1](College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)
[2](School of Mechanical Electronic & Information Engineering, China University of Mining & Technology(Beijing), Beijing 100083, China)

**Abstract**: Analyzing the shortcoming of traditional load forecast method, a method of load forecast based on data mining is put forward. Using the decision tree algorithm to make load forecasting, abnormal load data is found among the load forecasting results. This paper relies on association rules algorithm to modify the abnormal load, so that the load forecasting results will be more accurate.

**Key words**: data mining; association rule; decision tree algorithm; daily load curve

电力系统短期负荷预测问题是电力部门日常重要工作, 根据预测结果, 可以方便地进行电能生产. 利用数据挖掘技术, 建立数学模型, 进行负荷预测成为近年来研究人员工作的重点. 负荷曲线反映了某一时间段内负荷随时间而变化的规律, 按时间段长短分, 可分为日负荷曲线和年负荷曲线[5].

传统预测方法包括数理统计、混沌理论、小波回归分析等方法, 文献[2、4]利用小波变换对负荷序列进行分解, 得到不同频率的各个负荷分量, 然后利用数据分类和多元回归分析对各个分量进行负荷预测. 文献[6]利用多分辨分析的小波变换对短期电力负荷序列进行分解处理, 将负荷序列投影到不同尺度上, 根据各尺度上子序列的特性分别进行回归分析, 最后将预测结果叠加. 传统预测方法要求样本有较好的分布规律, 当预测长度大于原始数据长度时, 其预测结果的精度不能保证[2].

电力负荷的数据具有数据冗杂、数据量大、种类繁多等特点, 而数据挖掘在处理海量数据方面具有一定的优势, 因此, 越来越多的研究人员利用数据挖掘来进行负荷预测. 文献[1]针对传统 K 均值聚类算法中的不足, 提出了逐级均值聚类算法, 解决了传统聚类算法解的局部最优性问题. 文献[3]针对影响电力负荷预测的众多因素问题, 引入数据挖掘中的粗糙集约简算法, 有效地判断和选择出能够改善电力负荷预测的相关因素. 本文将利用数据挖掘的关联规则算法就日负荷曲线的预测进行分析研究.

## 1 关联规则简介

关联规则是一种常见的数据挖掘方法, 表明两个或多个变量的取值之间存在某种规律性. 设 $I=\{I_1, I_2, \ldots, I_m\}$是 m 个不同项目的集合, D 是数据库, 数据库中的每条记录 T 是一组属性, $T \in I$. 关联规则就是一种形如“$A \Rightarrow B$”的蕴含式, 其中 $A \in I$, $B \in I$, 并且 $A \cap B=\Phi$. 一般使用支持度、可信度、期望可信度和作用度四个

① 收稿时间:2014-04-10;收到修改稿时间:2014-05-15

参数来描述关联规则属性, 表 1 列举了它们的定义及含义.

表 1 关联规则属性的四个参数

| 参数名称 | 计算公式 | 概率表示 | 意义 |
|---|---|---|---|
| 置信度(Confidence) | $\frac{A和B的记录总数}{包含A的记录数}$ | P(B\|A) | 属性 A 出现的前提下, B 出现的概率 |
| 支持度(Support) | $\frac{A和B的记录总数}{记录总数}$ | P(A∩B) | 属性集 A 和 B 同时出现的概率 |
| 期望置信度 (Expected Confidence) | $\frac{包含B的记录总数}{记录总数}$ | P(B) | 属性集 B 出现的概率 |
| 作用度(Lift) | $\frac{置信度}{期望置信度}$ | P(B\|A)/ P(B) | 置信度对期望置信度的比值 |

为了发现有意义的关联规则, 给出两个阈值: 最小支持度和最小置信度. 如果规则的支持度大于最小支持度, 则认为此规则是频繁项集, 反之为非频繁项集. 如果同时满足最小支持度和最小置信度, 并且作用度大于 1 的规则, 则称为强关联规则. 关联规则挖掘的目的就是从数据库中挖掘满足用户要求的最小支持度与最小置信度的强关联规则.

## 2 预测过程及结果

首先我们需要定义负荷变化率, 定义如下:

$$\triangle P_t=\frac{P_{t+1}-P_t}{P_t}\times 100\%$$

其中, $\triangle P_t$为负荷变化率, $P_t$, $P_{t+1}$分别是一天内第 t, t+1 时刻的负荷值.

本文利用数据挖掘的决策树算法进行日负荷曲线的预测, 具体的实现过程在文献[7]有详细描述, 这里仅作简要介绍.

(1)把一天 24 小时等分为 96 份(每间隔 15 分钟分一点), 分别对每一点进行预测.

(2)采集每日气象数据、负荷数据.

(3)对以上数据进行预处理.

(4)根据历史数据, 建立决策树模型.

(5)利用聚类算法选择基准日.

(6)把基准日的气象数据、负荷数据以及与预测日的气象数据的差值代入决策树模型, 得出负荷变化率.

(7)根据基准日的负荷数据以及得出的负荷变化率得到预测日的负荷量.

本文选择某地区的负荷数据作为研究对象, 根据文献[7]的决策树算法, 预测出某一天的日负荷曲线, 并且与实际的日负荷曲线进行比较. 本文选择的负荷点间隔是 15 分钟, 一天共有 96 个点, 每一个单位代表 15 分钟, 结果如图 1 所示.

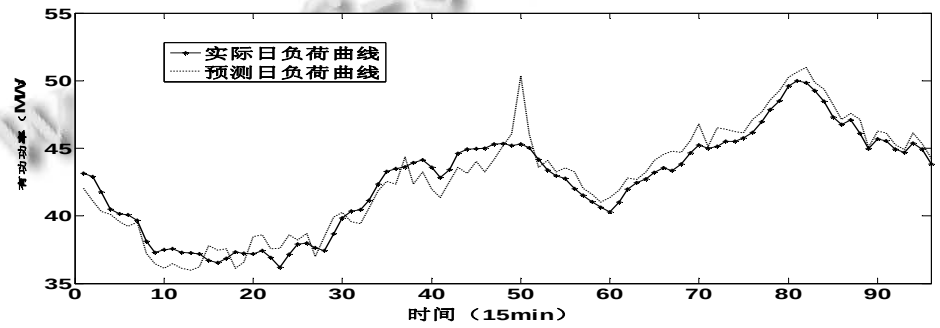


图 1 预测结果与实际日负荷曲线的比较

## 3 预测结果的修正

由以上算法预测的负荷曲线, 往往会存在某时刻的负荷值与实际不符. 电力负荷中, 当某点的负荷变化率超过 7%时, 则认为该点为冒大数点[8]. 如图 1 中 12 点的负荷变化率为 0.09, 大于 7%, 需要修正. 本文采用关联规则对冒大数点进行修正. 设某天 11: 45 到

12 点为事件 A, 12 点到 12 点 15 分为事件 B, 要修正的时间点为 12 点 15 分, 即事件 B. 下面将利用关联规则求 A⇒B 的频繁规则集.

(1)负荷变化率离散化

欲对 12 点 15 分的数据进行修正, 根据关联规则定义, 需要知道 12 点时刻的负荷. 由上文知, 12 点 15 分的数据应介于 12 点数据的±(1+7%)之间, 即[42.84,49.28], 负荷差为 6.44, 假设按照 0.1 的精度进行分度, 至少需要 64 个点. 根据负荷实际数据可知, 0.1 的精度依然很大, 如果我们按照 0.01 的精度进行分度, 则需要 644 个点, 那样将会是非常巨大的计算量, 因此我们需要按照负荷变化率进行分度. 根据本文采用的某地区历史负荷数据, 经过上文的负荷变化率公式计算可知, 本地区负荷变化率在-5%到 4%之间, 我们按照 1%的精度进行分度, 仅仅需要 8 个点, 因此我们按照负荷变化率进行计算. 为了简化计算, 把负荷变化率离散为以下状态, 负荷变化率状态表如表 2 所示.

表 2　负荷变化率状态表

| 状态编号 | 负荷变化率(%) | 状态编号 | 负荷变化率(%) |
|---|---|---|---|
| -4 | -5— -4 | 0 | -1—1 |
| -3 | -4— -3 | 1 | 1—2 |
| -2 | -3— -2 | 2 | 2—3 |
| -1 | -2— -1 | 3 | 3—4 |

(2)支持度表的生成

根据历史负荷数据, 统计出来事件 A, B 对应的所有负荷变化率的情况, 以本文选取时刻为例, 事件 A 的负荷变化率为 $\frac{45.21-45.36}{45.36}\times100\%=-0.33\%$, 事件 B 的负荷变化率为 $\frac{45.30-45.21}{45.21}\times100\%=0.19\%$, 查表 2 可知 A 对应的状态 0, B 对应的状态是 0, 因此 A 对应的 0 行与 B 对应的 0 列的数据加 1, 同理把其他时刻的负荷变化率按照上述方法进行统计, 统计结果如表 3 所示.

表 3　支持度表

| B / A | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|
| -4 | 0 | 0 | 1 | 5 | 6 | 3 | 1 | 0 |
| -3 | 0 | 0 | 2 | 5 | 7 | 6 | 4 | 0 |
| -2 | 1 | 3 | 6 | 12 | 23 | 13 | 6 | 2 |
| -1 | 0 | 3 | 7 | 21 | 22 | 14 | 9 | 6 |
| 0 | 2 | 5 | 8 | 35 | 28 | 23 | 12 | 3 |
| 1 | 1 | 7 | 13 | 33 | 31 | 24 | 16 | 8 |
| 2 | 2 | 8 | 15 | 45 | 36 | 21 | 16 | 6 |
| 3 | 0 | 1 | 3 | 6 | 4 | 2 | 0 | 0 |

根据专业人员的经验和实际运行总结, 设支持度阈值为 0.13. 由上表可得, 事件 A 取-4 时的支持度为 $\frac{16}{616}$=0.026, 同理可得其他状态的支持度, 如表 4 所示.

表 4　事件 A 取不同值对应的支持度

| A | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|
| 支持度 | 0.026 | 0.039 | 0.109 | 0.136 | 0.192 | 0.221 | 0.247 | 0.026 |

由表 4 可知, 仅当 A 取-1, 0, 1, 2 时, 其支持度大于 0.13, 接下来由表 3 数据计算 A 取-1, 0, 1, 2 的情况

下的置信度.

(3)置信度表的生成

由置信度的定义可知, 当 A=-1, B=-3 时, 即(-1) ⇒ (-3) 的置信度为 $\frac{3}{(3+7+21+22+14+9+6)}$=0.03, 同理可得其他情况下的置信度, 如表 5 所示.

表 5 置信度表

| B / A | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|
| -1 | 0 | 0.03 | 0.08 | 0.25 | 0.26 | 0.17 | 0.11 | 0.07 |
| 0 | 0.01 | 0.04 | 0.07 | 0.31 | 0.24 | 0.19 | 0.11 | 0.02 |
| 1 | 0.01 | 0.05 | 0.09 | 0.24 | 0.23 | 0.18 | 0.12 | 0.06 |
| 2 | 0.01 | 0.05 | 0.11 | 0.31 | 0.24 | 0.14 | 0.11 | 0.04 |

由此表可以找出关联规则 $A \Rightarrow B$ 的频繁规则集, 例如规则(-1) ⇒ (-2)的的支持度是 0.136, 置信度是 0.08. 表示“当 11: 45 到 12 点的负荷变化率为-0.02 到 -0.01 之间时, 则 12 点到 12 点 15 分的负荷变化率为 -0.03 到-0.02 之间”这条规则的支持度是 0.136、置信度是 0.08”. 同理, 可得其他类似的规则, 本文不再一一说明.

(4)结果修正

传统方法对于冒大数点的修正是利用“线性插值”法, 这种方法认为负荷曲线的变化是线性的. 因此, 12 点 15 分的负荷数据为其前一时刻和后一时刻的负荷和的一半, 即(46.06+46.03)/2=46.045. 传统方法修正负荷数据虽然简单易行, 但是准确性不高, 下面根据关联规则原理进行修正.

由负荷变化率的定义可知, 欲修正 12 点 15 分的数据, 即计算出 12 点的负荷变化率. $P_1$ 定义 11 点 45 分的负荷变化率, $P_2$ 为 12 点的负荷变化率. 按照关联规则, 具体计算如下:

11 点 45 分到 12 点的负荷变化率为:

$$P_1=\frac{46.06-45.22}{45.22}=0.018>0.01$$

则 A 取值为 1, 由表 5 得最大置信度为 0.24, 对应 B 值为-1, 即 12 点到 12 点 15 分的负荷变化率为 -0.02—-0.01 之间, 本文取上限值, 则:

$$\frac{P_2-46.06}{46.06}=-0.01$$

$$P_2=45.59$$

下面用实际数据分别与两种方法修正出来的数据进行比较, 如表 6 所示.

表 6 两种修正方法结果比较

| 时刻 | 实际值(MW) | 线性插值法(MW) | 关联规则法(MW) | 误差百分比% | 误差百分比% |
|---|---|---|---|---|---|
| 12:15 | 45.30 | 46.045 | 45.59 | 1.64 | 0.64 |

可见, 使用关联规则法修正过的数据更接近于实际值, 提高了预测的准确性. 把修正过的数据重新带入预测数据, 生成负荷曲线, 如图 2 所示, 与实际日负荷曲线进行比较, 发现结果基本吻合, 没有突变点.

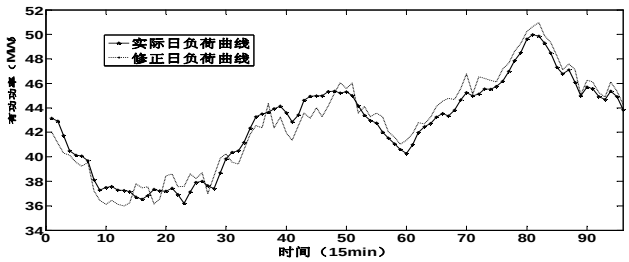


图 2 修正后的日负荷曲线

## 4 结论

本文提出了一种基于数据挖掘技术的负荷预测和修正方法, 通过决策树算法建立数学模型, 预测出未来的日负荷曲线. 在实际日负荷曲线的预测中, 可能会出现某点数据是冒大数点, 与实际数据偏差较大, 因此需要对这点预测数据进行修正. 传统修正方法大多采用线性插值法, 线性插值法虽然简单, 但是预测精度不高. 相较传统修正方法, 关联规则修正的预测结果更接近于实际负荷数据, 一定程度上提高了预测精度. 根据某地实际负荷数据, 对预测出的负荷曲线分别利用两种修正方法对同一点异常数据进行修正, 发现关联规则修正的数据的准确性更高, 同时修正后的负荷曲线更接近于实际负荷曲线, 取得了预期的结果.

## 参考文献

1 刘小华.数据挖掘在电力系统短期负荷预测智能化建模中的应用研究[学位论文].武汉:华中科技大学,2003.

2 李冬伟.基于数据挖掘的电力系统短期负荷预测研究[学位论文].大连: 大连理工大学,2007.

3 薛美娟.基于数据挖掘技术的电力系统短期负荷预测[学位论文].西安: 西安理工大学,2007.

4 黄达文.基于数据挖掘的电力系统短期负荷预测应用研究[学位论文].广州:华南理工大学,2009.

5 陈珩.电力系统稳态分析(第 3 版).北京:中国电力出版社,2007.

6 张涛,朱建良.小波回归分析法在短期电力系统负荷预测中的应用.哈尔滨理工大学学报,2008.2:74−76.

7 洪流,张海勤,肖明军,等.一种基于数据挖掘技术的电力负荷预测系统.小型微型计算机系统,2004,3:434−437.

8 耿波,仲红,徐杰,等.用关联分析法对预测结果进行二次处理.计算机技术与发展,2008,4:171−173.