

用关联分析法对负荷预测结果进行二次处理

耿波, 仲红, 徐杰, 闫娜娜

(安徽大学 计算机信息学院, 安徽 合肥 230039)

摘要:短期电力负荷预测是电力系统管理现代化的重要内容之一, 是对发电、输电和电能分配等合理安排的必要前提, 对提高电力系统的各方面的效益有非常重要的影响。随着电力系统管理现代化的发展, 电力部门对短期负荷预测精度的要求也越来越高。介绍了现有的电力负荷预测技术的研究状况, 并在此基础上提出了利用关联规则分析法对负荷预测结果进行二次处理。实验结果表明负荷预测结果通过关联规则分析法的处理, 其准确率得到一定程度的提高。

关键词:负荷预测; 数据挖掘; 关联规则; 聚类; 决策树

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2008)04-0171-03

Using Correlation Analysis to Treat Load Forecasting Results

GENG Bo, ZHONG Hong, XU Jie, YAN Na-na

(School of Computer and Information, Anhui University, Hefei 230039, China)

Abstract: The short-term load forecast is an important aspect of power management system modernization. It's very important to improve benefits of power system. Along with the development of electrical power system management modernization, the request of the short-term load forecast from electric power department become more and more. Describes the existing power load forecasting technology development, and on this basis, using the association rule analysis to treat the load forecast. Experimental results show that the results of the load forecasting accuracy is enhanced to a certain degree.

Key words: load forecast; data mining; association rule; clustering; decision tree

0 引言

短期电力负荷预测是电力系统管理现代化的重要内容之一, 对提高电力系统的经济效益和社会效益, 保障电力系统的安全经济运行与国民经济的发展具有非常重要的影响^[1-3]。

文献[4]提出了一种利用决策树算法预测短期负荷的系统的具体实现, 利用天气和负荷量的关系建立模型进行预测, 在得到结果后, 对预测结果中明显不正常的数据利用线性插值算法进行处理^[1]。这样虽然可以在一定程度上减小误差, 但这种简单地认为负荷量是线性变化的设计仍然会影响预测的准确率, 尤其是在无法线性内插, 而使用线性外插算法对不正常数据点进行处理的情况下误差可能会很大。而文献[5]提出了利用小平方误差对预测结果曲线进行平滑处理, 从而在一定程度上降低了预测结果的误差。其他一些

学者也提出了一些方法对预测结果进行处理, 但始终还是利用预测结果曲线本身的特点进行处理, 没有充分利用海量历史数据中潜在的规律。在笔者设计的算法中, 在得到预测结果后, 利用关联规则算法从历史大量负荷数据中挖掘出连续时间点负荷量的关联规则, 再利用这种关联性规则对预测结果中不正常数据进行二次处理。实验结果表明负荷预测的准确率得到一定的提高。

1 预备知识

负荷预测:从已知的用电状况及用电趋势出发, 研究政治、经济、气候和人民生活等相关因素影响规律, 对未来的用电需求所做出的预测。

短期负荷预测:预测未来1~7日的负荷。用于火电分配、水火电协调、机组经济组合交换功率计划等。

关联规则挖掘:设 $I = \{i_1, i_2, \dots, i_n\}$ 是项的集合, 设任务相关的数据 D 是数据库事务的集合, 其中每个事务 T 是项的集合, 使得 $T \subseteq I$, 每个事务有一个标识符, 称作 TID。设 A 是一个项集, 事务 T 包含 A 当且仅当 $A \subseteq T$, 关联规则是形如 $A \Rightarrow B$ 的蕴涵式, 其中

收稿日期: 2007-07-10

基金项目: 安徽省自然基金项目(070412051); 安徽高校省级重点自然科学研究项目(KJ2007A43)

作者简介: 耿波(1982-), 男, 硕士研究生, 研究方向为数据挖掘; 仲红, 硕士生导师, 研究方向为数据挖掘。

$A \subset I, B \subset I$, 并且 $A \cap B = \emptyset$, 规则 $A \Rightarrow B$ 在事务集 D 中成立, 具有支持度 s , 其中 s 是 D 中包含 $A \cup B$ 的百分比, 即它是概率 $P(A \cup B)$, 规则 $A \Rightarrow B$ 在事务集 D 中具有置信度 c , 它是 D 中包含 A 的事务同时也包含 B 的百分比, 这是条件概率 $P(B | A)$ 。即 $\text{Support}(A \Rightarrow B) = P(A \cup B)$, $\text{Confidence}(A \Rightarrow B) = P(B | A)$, 挖掘关联规则就是产生那些支持度和置信度分别大于用户给定的最小支持度和最小置信度阈值的规则, 阈值可以由用户或领域专家根据经验设定^[6,7]。

2 用关联分析法对负荷预测结果进行二次处理

首先简单介绍现有的基于数据挖掘算法的负荷预测的实现过程, 具体的实现过程在文献[4]中也有详细描述, 这里只做简单介绍。

具体的做法是将 1 天等分成 96 个时间点(每隔 15 分钟一点), 对每一点进行预测, 大致过程如下:

1) 采集每天的气象预报数据、气象实况数据和负荷数据。

2) 对采集的数据进行预处理。

3) 利用历史的海量数据建立决策树模型。

4) 通过聚类算法选择基准日(在近期的气象数据中选出与预测日的欧式距离最近的一天)。

5) 将基准日的气象数据和负荷数据以及与预测日的气象数据差值代入决策数模型得出负荷变化率。

6) 通过基准日负荷数据和得到的负荷变化率计算出预测日的负荷量。

最后, 将预测出的 96 点的负荷量用曲线图显示, 再由人工调整。

下面重点介绍用关联规则挖掘算法对预测结果曲线图中需要修改的数据点进行调整。

先通过举例说明使用关联规则的原因:

假设通过决策树模型得到如下结果(见表 1)。

表 1 负荷预测 (kW)

时间点	负荷量
15 点	99
15 点 15 分	96
15 点 30 分	96
15 点 45 分	96
16 点	106
17 点 15 分	94
17 点 30 分	98

可以观察到 16 点的负荷数据量很大, 而电力负荷数据有一个规律, 在连续时间点的负荷量变化率不会

大于 7%, 否则认为是冒大数, 在以往的各种负荷预测系统中都采用了人工调整或程序调整的方法处理这些冒大数, 但均是采取线形插值的算法, 即简单地认为负荷量是线形变化的。在上面假设的例子中 16 点的负荷量将会被调整为 95(即 $(96 + 94)/2$)。这种处理方式虽然简单但却会影响预测的准确率, 尤其在某些情况下无法使用线形内插值而只能采用线性外插的情况下, 误差率将会更大。在笔者设计的算法中不是简单地认为负荷量是线性变化的, 而是设计算法从历史海量的数据中挖掘出连续时间点负荷变化的关联规则, 用这些关联规则对冒大数点进行处理。笔者采取类 Apriori 算法寻找频繁规则集合^[6], 再利用这些频繁规则处理预测错误的数据。

下面举例说明具体的算法:

设 A 为 2 点 45 分到 3 点的负荷变化率, B 为 3 点到 3 点 15 分的负荷变化率。用下面的做法来找到 $A \Rightarrow B$ 的频繁规则集:

首先, 将数据离散化, 为了算法简化, 将负荷变化率离散化成 11 类, 0 表示变化率在 -0.01 到 0.01 之间, 1 表示变化率在 0.01 到 0.02 之间, 2 表示变化率在 0.02 到 0.03 之间, 3 表示变化率在 0.03 到 0.04 之间, 4 表示变化率在 0.04 到 0.05 之间, 5 表示变化率在 0.05 到 0.06 之间, -1 表示变化率在 -0.01 到 -0.02 之间, -2 表示变化率在 -0.02 到 -0.03 之间, -3 表示变化率在 -0.03 到 -0.04 之间, -4 表示变化率在 -0.04 到 -0.05 之间, -5 表示变化率在 -0.05 到 0.06 之间, 再构造 A 和 B 的关联表。

假设有如下的关联表(见表 2)。

表 2 支持度表

B \ A	-5	-4	-3	-2	-1	0	1	2	3	4	5
-3	0	0	2	5	5	8	3	4	0	0	0
-2	0	1	9	7	17	33	29	15	5	0	0
-1	0	0	9	14	21	39	20	21	12	1	0
0	0	2	9	22	26	50	22	19	11	3	0
1	0	1	5	13	20	53	31	18	7	1	0
2	0	0	6	10	11	37	29	19	8	2	0
3	0	0	1	0	2	7	1	1	0	0	0

根据相关工作经验, 支持度阈值设为 0.13。通过表中的概率数据计算得出, 仅有 A 为 $-2, -1, 0, 1, 2$ 这 5 类的支持度大于 0.13, 分别为 0.145, 0.201, 0.225, 0.204, 0.168。再通过表 2 中的数据计算 A 为 $-2, -1, 0, 1, 2$ 的情况下, $A \Rightarrow B$ 的置信度。

例如: $(-2) \Rightarrow (-3)$ 的置信度为 $9/(0 + 1 + 9 + 7 + 17 + 33 + 29 + 15 + 5 + 0 + 0) = 0.09$, 其他不再一一说明。得到如下置信度表(见表 3)。

表3 置信度表

$\begin{matrix} B \\ A \end{matrix}$	-5	-4	-3	-2	-1	0	1	2	3	4	5
-2	0	0.01	0.09	0.07	0.16	0.31	0.27	0.14	0.05	0	0
-1	0	0	0.06	0.10	0.14	0.27	0.14	0.14	0.09	0.01	0
0	0	0.01	0.05	0.13	0.16	0.30	0.13	0.12	0.07	0.02	0
1	0	0.01	0.03	0.09	0.13	0.36	0.21	0.12	0.05	0.01	0
2	0	0	0.05	0.08	0.09	0.30	0.22	0.16	0.07	0.02	0

接下来,找出置信度大于置信度阈值的关联规则即为 $A \Rightarrow B$ 的频繁规则集,例如从表中可以看到有如下规则 $(-2) \Rightarrow (-1)$ 的支持度为 0.145, 置信度为 0.16 (意思是:“如果 2 点 45 分到 3 点的负荷变化率在 -0.02 到 -0.03 之间,则 3 点到 3 点 15 分的负荷变化率在 -0.01 到 -0.02 之间”这条规则的支持度为 0.145, 置信度为 0.16)。其他规则不再一一说明。在得到频繁规则集后,如果 3 点 15 分的预测数据冒大数,就用得到的频繁规则集对该点进行处理。实验证明通过这种关联规则分析的方法对预测结果进行处理后得到的最终预测结果的准确率比原系统有所提高。

3 实验结果

笔者采用 2003 年 6 月到 2005 年 6 月的黄山市负荷数据作为训练数据,对 2005 年 7 月到 12 月黄山市的负荷量进行预测。

通过对几个月的负荷量的预测,和与实际负荷量的比较得到各月的平均准确率见表 4。可以看到采用关联规则处理后的准确率比原系统有了明显提高。

4 结束语

在原有负荷预测系统基础上,对其算法进行改进,

并采取关联规则算法对预测后的结果进行处理。并且在实验中证明,预测准确率得到一定的提高。

然而,在电力负荷数据库中隐藏着很多有价值的规则,远远没有被充分挖掘。

表4 月负荷预测准确率对比表

日期	采用了线性插值方法的准确率	采用关联规则处理的准确率
2005 年 7 月	96.90%	96.93%
2005 年 8 月	95.61%	96.11%
2005 年 9 月	94.31%	94.12%
2005 年 10 月	96.72%	96.88%
2005 年 11 月	95.13%	96.11%
2005 年 12 月	95.44%	96.12%

参考文献:

- [1] 蔡佳宏. 超短期负荷预测中相似日的选择方法[J]. 华北电力大学学报, 2006(1): 38-41.
- [2] 路广. 数据仓库与数据挖掘技术在电力系统中的应用[J]. 电网技术, 2001(8): 54-57.
- [3] Niu xiao-dong. Techniques and application of electric load prediction[M]. Beijing: China Electric Power Press, 1998.
- [4] 洪流. 一种基于数据挖掘算法的电力负荷预测系统[J]. 小型微型计算机系统, 2004(3): 434-437.
- [5] 龙立波. 短期电力负荷预测中的数据处理技术[J]. 电力需求侧管理, 2007(1): 11-14.
- [6] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 等译. 北京: 机械工业出版社, 2001.
- [7] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C] // In Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data. Washington, D.C.: [s. n.], 1993: 207-216.

(上接第 167 页)

- [7] 孟令奎, 史文中, 张鹏林. 网络地理系统原理与技术[M]. 北京: 科学出版社, 2005.
- [8] 孙美娟. 基于 XML 技术的空间信息发布方法研究[D]. 北京: 北京林业大学, 2006.
- [9] Zimmermann R, Ku Wei-Shinn, Wang Haojun, et al. A Distributed Geotechnical Information Management and Ex-

change Architecture[J]. IEEE Computer Society, 2006(10): 26-33.

- [10] 芦东昕, 李典蔚, 柳长安. 基于 AJAX 和 Servlet 的 Web GIS 的研究与实现[J]. 计算机技术与发展, 2007, 17(3): 193-195.

(上接第 170 页)

- [4] Rosenberg J, Schulzrinne H. An RTP Payload Format for Generic Forward Error Correction[S]. IETF Draft. RFC 2733. 1999.
- [5] Wang Yao, Wenger S, Wen Jingtao, et al. Error resilience Video Coding Techniques, Real-Time video Communication

over Unreliable Networks[J]. IEEE Signal Processing Magazine, 2000(7): 61-82.

- [6] Roth G, Sjöberg R, Liebl G, et al. Common Test Conditions for RTP/IP over 3GPP/3GPP2 - Amendments and Software [R]. ITU-T SG16 Doc. VCEG-M37. Santa Barbara, CA, USA: [s. n.], 2001.