决策树技术分析气象因子对电力负荷预测的影响

高 霞[1,2] 曾 新[3] 马 骋[3]

(1. 河北省保定市气象局,071000; 2. 河北省气象与生态环境实验室; 3. 华北电力大学)

提 要:基于决策树技术,对气象因子和日电力负荷的最高、最低值、平均值进行联合建模,量化气象因子对电力负荷的影响,从而确立一种有效的基于气象因子的短期电力负荷预测方法,用以生成日特征负荷决策树预测模型。通过该模型,结合预测日的气象、属性(日期、节日等)等信息,可进行日特征负荷的预测。预测结果表明,该模型具有自动化程度高、预测结果准确率高的特性。以河北省保定市气象数据和电力负荷数据为例进行了训练和预测,研究结果证明这种方法能较大地提高日电力负荷预测的精度。

关键词:决策树 气象因子 负荷预测 ID3算法

Influence of Meteorological Factors on Load Forecasting Based on the Decision Tree

Gao Xia[1] Zeng Xin[2] Ma Cheng[3]

(1. Baoding Meteorological Office, Hebei Province 071000; 2. Hebei Province Key Lab for Meteorology and Eco-environment; 3. North China Electric Power University)

Abstract: Based on the decision tree and combined with the meteorological factors and the highest, lowest and average of electrical load, a model was established to measure the influence of meteorological factors on the load. An effective short-term forecasting method was put forward to establish the forecasting model of daily-characteristic-load decision tree. Daily characteristic load is forecasted according to the model and the date-forecasted information such as weather, attributes (date, workday or weekend). The forecasted results show that the proposed method has high-automation and high-accuracy. The method applies the data of meteorological factors and load of Baoding in Hebei Province to train and forecast, the results show that the new method can largely improve the precision of load forecasts.

Key Words: decision tree meteorological factor load forecasting ID3 algorithm

收稿日期:2007年9月29日; 修定稿日期:2008年1月29日

引　言

电力负荷预测是电力调度、用电、计划、规划等管理部门的重要工作之一。影响负荷的因素很多,气象因子的影响是其中之一,对此做一研究,对提高负荷预测技术水平、更合理地进行用电管理、安排电网运行方式和机组检修计划,提高电力系统的经济效益和社会效益都是很有益的[1-3]。

众多研究者提出了许多短期负荷预测方法,如时间序列法、专家系统、人工神经网络法等,这些方法各有自己的优点,但由于负荷变化的影响因素多且过程复杂,它们都存在一定的缺陷性和局限性。时间序列法计算量小、速度较快,但预测精度不能保证满足工程需要,且不具备自适应学习能力,预测系统的鲁棒性没有保障[4]。专家系统可以避开复杂的数值计算,但通用性较差,缺乏学习能力[5-6]。人工神经网络具有很强的鲁棒性、记忆能力、非线性映射能力以及强大的自学习能力,但收敛速度慢和可能收敛到局部最小点,且知识表达困难,难以充分利用调度人员经验中的知识[7-8]。

决策树(Decision Tree)[9-12]学习是以实例为基础的归纳学习算法,它着眼于从一组无次序、无规则的事例中推理出决策树表示形式的分类规则。利用决策树技术对气象因子和日电力负荷进行联合建模,通过训练历史数据,量化气象因子对电力负荷预测的影响,然后通过生成的预测决策树模型对历史数据进行检验、对未来数据进行预测,通过对预测值与真实值的对比,验证得到负荷预测的优化值。在负荷预测模型的生成中,综合考虑了气温、湿度等气象信息及节假日因素对日特征负荷的影响,不仅具有较好的预测结果,而且可以在一定程度上揭示出影响日特征负荷因素的相对重要性。

1　决策树技术及剪枝

决策树方法是数据挖掘中非常有效的分类方法,用样本的属性作为结点,用属性的取值作为分支的树结构,利用信息论原理对大量样本的属性进行分析和归纳而产生的。决策树的根节点是所有样本中信息量最大的属性;树的中间结点是该结点为根的子树所包含的样本子集中信息量最大的属性;决策树的叶子结点是样本的类别值。

由于 ID3 算法的基础理论清晰,算法较简单,学习能力较强,能够处理大规模的学习问题,通过 ID3 这种决策树分类算法,采用基于信息熵定义的信息增益度量来选择内节点的测试属性。

1.1　ID3 算法

设 S 是 n 个数据样本的集合,将样本集划分为 c 个不同的类 $C_i(i=1,2,\cdots,c)$,每个类 C_i 含有的样本数目为 n_i,则 S 划分为 c 个类的信息熵或期望信息为:

$$\mathrm{E}(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{1}$$

其中,p_i 为 S 中的样本属于第 i 类 C_i 的概率,即 $p_i = n_i/n$。

S_v 是 S 中属性 A 的值为 v 的样本子集,即 $S_v = \{s \in S \mid A(s) = \mathrm{v}\}$,选择 A 导致的信息熵定义为:

$$\mathrm{E}(S,A) = \sum_{v \in \mathrm{Value}(A)} \frac{|S_v|}{|S|} \mathrm{E}(S_v) \tag{2}$$

其中,$\mathrm{E}(S_v)$是将 S_v 中的样本划分到各个类的信息熵。属性 A 相对样本集合 S 的信息增益 $\mathrm{Gain}(S,A)$定义为:

$$\mathrm{Gain}(S,A) = \mathrm{E}(S) - \mathrm{E}(S,A) \tag{3}$$

$\mathrm{Gain}(S,A)$是指因知道属性 A 的值后导致的熵的期望压缩。$\mathrm{Gain}(S,A)$越大,说明选择测试属性 A 对分类提供的信息越多。ID3

算法就是在每个节点选择信息增益(S,A)最大的属性作为测试属性。

1.2 决策树的剪枝

创建决策树时,由于训练样本太少或数据中存在噪音和孤立点,许多分枝反映的是训练样本集中的异常现象,建立的决策树会过度拟合训练样本集。剪枝方法可以减少训练样本集中噪音的影响,剪枝的时机选择是很关键的。

(a) 本方法中选择最小分度值为 5,即扩展的叶子结点为每 5 个数据的汇总。

(b) 基于误差的剪枝。

本方法中利用生成的原始决策树,对每个叶子结点进行代价计算,建立差分方程,使得当斜率较为平缓时即可进行剪枝。

2 数据源

原始数据中经常存在噪声数据及缺测数据,对于缺测数据当时间的跨度范围并不是很大时($n\leqslant 5$ 天),可考虑用 3 次样条插值方法进行数据的插值填充;当时间跨度范围较大时只能将数据舍去。对于噪声数据可通过降噪手段(主要方法为插值、拟合)进行数据的修复,若噪声很大时则将其舍去。

系统使用 Visual C# 2005 作为开发平台,使用 Matlab7.0 做后台运算和图形显示工具,SQL Server 2005 作为预测数据库。数据库中包括 2004 年河北省保定市每日的电力负荷,以及相应的与负荷有关的气温、湿度、降水量等气象数据,这些数据可以通过 SQL Server 2000 的 DTS 包定期加载。其中将训练集合内的负荷数据作为历史数据,训练集合外的负荷数据作为新数据。电力负荷数据中对每日 96 时刻[1)] 的负荷值建立差分模型,计算出日最高电力负荷及日最低电力负荷。

3 决策树负荷预测模型的实现

以河北省保定市 2004 年 1 月到 2004 年 12 月的气象资料通过决策树技术建模。将日气象数据中的各种气象因子分别作为属性值,构造分类决策树,包括月份、日期、气压、温度、水汽压、降水、相对湿度、云量、蒸发量、地温、日照时数、平均风速、湿球温度、能见度、星期、节日类型等。

根据改进的 ID3 算法输出节点表(如表 1 所示),将表中“节点”字段通过“连接”字段与“父节点”字段连接起来,形成负荷预测决策树,图 1 为以文本形式表示的最高值负荷预测决策树。

表 1 改进 ID3 算法输出表结构

序号	字段名	数据类型	键	空值	索引
1	节点号	整型	主	否	是
2	节点	字符型		否	
3	父节点	字符型		是	
4	连接	短整型		是	

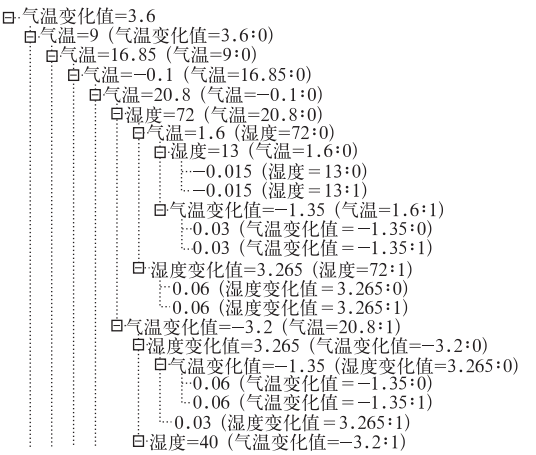

图 1 部分最高值负荷预测决策树图

1)本文以保定地区所属的保定市区、阜平、涞源、涿州四个区域的电力负荷值作为研究样本,数据来源于 SCADA 系统,每天 96 个采本点(每 15 分钟一个点)。整点时刻一共 24 时刻,进行日 24 点或 96 点负荷预测,基本能够满足负荷预测的需要。

决策树中的叶子结点即为在其父节点属性条件下的负荷值。通过将训练的气象数据带入生成的决策树得到预测的负荷值,并将随着决策树叶子结点规模扩展得到的真实值与预测值的误差平方和的序列作图得到“决策树规模——代价(误差平方和)”的决策树评价图(图2、图3)。可以看出,当决策树仅有一个结点时,其代价最大。随着叶子结点规模的扩大,决策树代价也在不断减小。当叶子结点规模即决策树的规模达到一定程度时,决策树代价的减少率趋于缓和,利用差分方程将拐点前的结点信息分离提取即可得到最终相关的多个因子的优化决策树(图4):

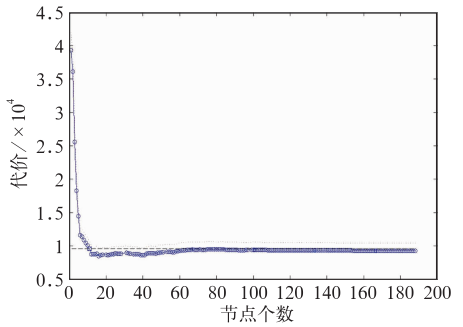

图 2　全因子日负荷最高值决策树评价图

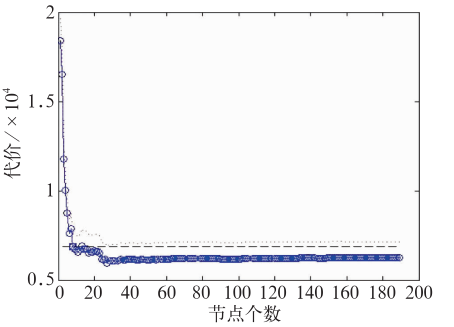

图 3　全因子日负荷最低值决策树评价图

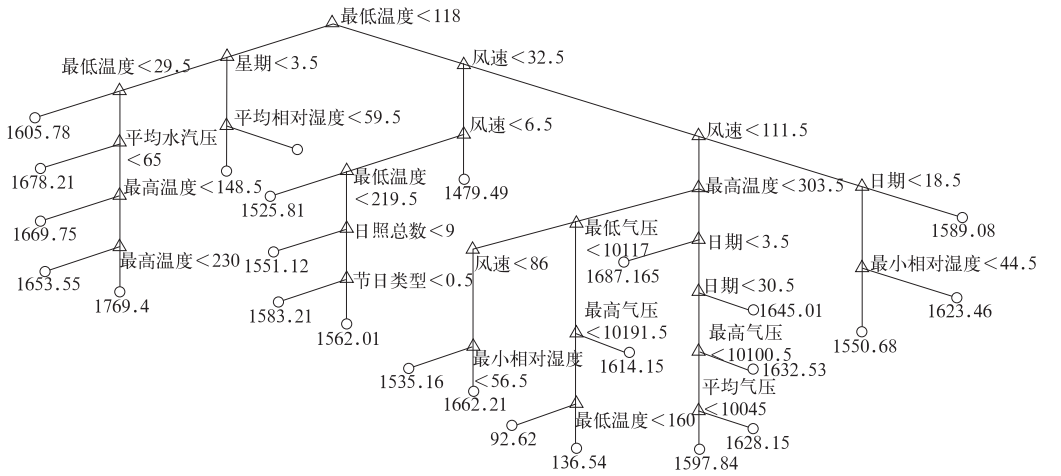

图 4　多个气象因子的日负荷最高值决策树

4　预测结果分析

利用决策树技术对多组不同训练及生成决策树并进行历史数据预测得到结果如表2。

利用决策树技术对多组不同训练及生成决策树并进行新数据预测得到结果如表3。

(1) 历史负荷预测检验

通过模型建立,负荷预测值的平均相对误差≤3.502%,最高负荷平均相对误差≤5.005%,最低负荷平均相对误差≤1.249%。

表2 优化决策树历史数据预测结果

训练集时间段	预测时间段	日负荷最高值		日负荷最低值	
		e_M	$\bar{e}$	e_M	$\bar{e}$
2004.7.9－2004.10.6	2004.8.7－2004.9.6	3.4%	0.9%	4.5%	1.2%
2004.7.29－2004.10.26	2004.7.29－2004.10.26	4.9%	1.2%	4.1%	0.7%
2004.11.6－2005.1.4	2004.11.20－2004.12.20	1.9%	0.5%	2.8%	0.8%
2004.11.16－2005.1.14	2004.11.16－2005.1.14	4.0%	0.8%	3.0%	1.0%
2004.12.6－2005.1.4	2004.12.6－2005.1.4	1.8%	0.7%	2.6%	0.8%

注：e_M、$\bar{e}$ 分别为最高相对误差和平均相对误差。

表3 优化决策树新数据预测结果

训练集时间段	预测时间段	日负荷最高值		日负荷最低值	
		e_M	$\bar{e}$	e_M	$\bar{e}$
2004.9.17－2004.12.15	2004.12.16－2004.12.30	5.7%	2.9%	5.3%	2.1%
2004.9.17－2004.12.15	2004.12.16－2004.12.22	2.9%	1.7%	3.7%	1.8%
2004.10.27－2004.12.25	2004.12.26－2005.1.9	5.7%	3.1%	5.0%	2.1%
2004.10.17－2004.12.15	2004.12.16－2004.12.22	2.9%	1.7%	3.7%	1.8%
2004.2.10－2004.3.10	2004.3.11－2004.3.17	2.9%	1.1%	2.9%	1.6%

(2) 未来负荷预测检验

训练集外的新负荷预测值得平均相对误差≤5.108%，最高负荷平均相对误差≤5.712%，最低负荷平均相对误差≤3.172%。

通过对模型的检验可知，预测结果的精度较高，达到工业技术要求。

(3) 决策树预测历史数据较预测新数据有更高的精度，对保定2004年1月至2004年12月的最大、最小及平均负荷进行预测，则一年平均最大值预测精度可达94.3%，平均值预测精度可达94.86%，最小值预测精度达到96.7%。而且预测短时间范围的数据较预测长时间范围数据有更高的精确度。用优化决策树预测负荷时，应尽量采用附近时期的1～2个月的训练集来预测短时间(7天)范围的数据可获得较好的效果。

5 结 论

气象因子与负荷有着密切的关系，通过对多组训练集生成决策树的预测值与真实值的比较可知，气象类的8个因子(月份、日期、最高气温、最低温度、平均水汽压、平均风速、星期、节日类型)对电力负荷具有较高并且较稳定的量化影响，用这8个气象因子来对短期日电力负荷的最高最低值进行预测得到的结果具有较高的精确度。其中，将所有气象属性都处理为基值和变化值[4]。月份、日期、星期、节日类型等属性可直接实现，其余的气象因子通过该算法进行模型训练，形成24个或96个预测模型，则可进行日24点或96点负荷预测。

决策树算法可以揭示分类属性对决策属性的相对重要性。作为决策树算法的一种，从ID3算法生成的负荷预测决策树中我们还可以观察到：气温和气温变化值对所在地区的日最大、最小负荷变化率具有相对较大的影响，而湿度、湿度变化值及星期因素对负荷最大、最小变化率的影响相对较弱；对日平均负荷变化率影响较大的则有气温、星期和气温变化率因素。总体看来，气温及其变化值对负荷影响较大，而湿度及其变化值对负荷影响较小。

参考文献

[1] 牛东晓. 电力负荷预测技术及其应用[M]. 北京:中国电力出版社.

[2] 胡江林,陈正洪,洪斌,等. 华中电网日负荷与气象因子的关系[J]. 气象,28(3):14-18.

[3] Yang Hongtzer, HuangChao-Ming. A New Short-Term Load Forecasting Approach Using Self-Organizing Fuzzy ARMAX Models[J]. IEEE PWRS, 1998, 13(1):464-473.

[4] RahmanS, BhatnagarR. An Expert System Base Algorithm for Short-Term Load Forecasting[J]. IEEE PWRS, 1998, 3(2).

[5] Hok-L, HsuYY, LeeCE, et al. Short-Term Load Forecasting of TaiWan Power System Using a Knowledge-Based Expert System[J]. IEEE PWRS, 1990, 5(4).

[6] T M Peng, N F Hubele, G G Karady. Advancement in the Application of Neural Networks for Short-Term Load Forecasting[J]. IEEE PWRS, 1992, 7(1):427-435.

[7] HoK-L. Short-Term Load Forecasting Using Multi-Layer Neural Network with an Adaptive Learning Algorithm[J]. IEEE PWRS, 1992, 7(1).

[8] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, (1): 81-106.

[9] 李雄飞,李军. 数据挖掘与知识发现[M]. 北京:高等教育出版社,2003.

[10] 朱六璋,袁林,黄太贵. 短期负荷预测的实用数据挖掘模型[J]. 电力系统自动化,2004,28(3):49-52.

[11] 汪峰,于尔铿. 基于因素影响的电力系统短期负荷预报方法的研究[J]. 中国电机工程学报,1999,19(8):54-57.

[12] Ying-Hwa Kuo, Donall, Shapiro M A. Feasibility of short-range numerical weather prediction using observations from a network of profilers[J]. Mon. Wea. Rev, 1987,115:2402-2427.