数据挖掘和动态神经网络在负荷预测中的研究与应用

李晓峰[1] 黄国兴[2] 高巍巍[3] 丁树春[4]

(北京理工大学计算机科学与技术学院 北京 100081)[1] (华东师范大学软件学院 上海 200062)[2]
(黑龙江外国语学院信息科学系 哈尔滨 150025)[3] (黑龙江大学电子工程学院 哈尔滨 150080)[4]

摘 要 中长期负荷变化规律与社会经济指标的关系很难用一个准确的数学模型来表达。将数据挖掘技术应用到全社会用电量增长的关联分析中,从2000年以来的社会经济指标中选取多项组成相关因素数据库,对缺失数据进行了补全,使用聚类分析挖掘出与全社会用电量关系密切的若干指标,并对指标中的失真数据进行修正,构建了更加科学的负荷预测模型。通过时间序列的动态神经网络,对负荷预测模型进行了测试和验证,结果表明该预测模型具有很好的收敛性,效果令人满意。

关键词 年度负荷预测,数据挖掘,数据补全,数据修正,NARX神经网络

中图法分类号 TP39 **文献标识码** A

Research and Application of Data Mining and Dynamic Neural Networks in Load Forecasting

LI Xiao-feng[1] HUANG Guo-xing[2] GAO Wei-wei[3] DING Shu-chun[4]

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)[1]
(Software Engineering Institute, East China Normal University, Shanghai 200062, China)[2]
(Department of Informatics and Science, Heilongjiang International University, Harbin 150025, China)[3]
(School of Electronic Engineering, Heilongjiang University, Harbin 150080, China)[4]

Abstract Dependence of medium and long-term load variation on socio-economic indicators is difficult to express by an accurate mathematical model. This paper applied data mining techniques to the association analysis of the total electricity consumption growth. Multiple indicators were selected from the socio-economic indicators since 2000 to compose relevant factors database, and the missing data were completed. Several indicators closely related to total electricity consumption were mined using cluster analysis, and the distortion data were corrected, thus a more scientific load forecasting model was constructed. Through time series of dynamic neural network, the load forecasting model was tested and validated. The results show that the prediction model has good convergence and satisfactory effect.

Keywords Annual load forecast, Data mining, Data completion, Data correction, NARX neural network

1 引言

电力系统负荷预测是指在考虑一些重要的系统运行特性、增容决策和自然条件下,利用一套系统的处理过去和未来负荷的方法,在一定精度意义下,决定未来某特定时刻或某些特定时刻的负荷值。负荷预测是电力系统规划、设计、调度的基础,其重要性早已被人们所认识。按预测期限的不同,负荷预测可分为年度预测、月度预测、日预测、小时预测。年度负荷预测的方法除专家调查法、产值单耗法、弹性系数法等传统预测方法外,还有回归分析、灰色预测以及神经网络等预测方法[1]。进行中长期负荷预测时,需要引入社会经济指标,但几乎所有的社会经济指标都对电力负荷有着或大或小的影响[2],如果引入全部的社会经济指标,产生的工作量和计算量将很难接受。国内外研究人员在负荷预测中通常会基于经验引入若干项社会经济指标,如国内生产总值(GDP)、各个产业的总值、总人口等。但选取相关指标还没有统一的理论依据。

随着数据挖掘技术的发展和应用,其分析思路和计算算法在电力系统的预测、分类、聚类分析、关联分析等领域都得到有效应用,提高了研究人员处理历史数据的能力[3,4]。本文将数据挖掘技术应用到中长期负荷预测中,从2000年至2010年全国经济发展的数据中,选取多个社会经济发展指标作为候选指标,通过计算全社会用电量与所选社会经济发展指标间的关联程度,挖掘出与全社会用电量增长关联密切的指标,并结合时间序列的动态神经网络,对所选指标进行了验证。研究结果显示了全社会用电量增长的一般性规律。

2 相关工作

2.1 数据补全

中长期负荷预测的历史数据易缺失,如果抛弃含有缺失数据的社会经济指标,就会降低模型的预测性能[5]。因此,对

到稿日期:2013-05-25 返修日期:2013-07-11 本文受国家自然科学基金(61102071),教师科研专项基金(CTF120771)资助。

李晓峰(1978—),男,博士生,副教授,CCF高级会员,主要研究方向为数据挖掘、文本挖掘、智能算法,E-mail:mberse@126.com;黄国兴(1946—),男,教授,博士生导师,主要研究方向为数据挖掘、信息安全;高巍巍(1976—),女,硕士,副教授,CCF会员,主要研究方向为数据库、数据挖掘;丁树春(1967—),男,硕士,副教授,主要研究方向为无线电通信与信息安全。

于缺失数据应采取合理的填补措施补充完整。如果缺失首、末端数据，可以采用趋势比例计算进行数据的补全，例如数列

$$\{X\}=[\phi(1),X(2),X(3),\cdots,\phi(n)] \tag{1}$$

数列中 $\phi(1)$和 $\phi(n)$为缺失数据，则对其补全的公式为：

$$\phi(1)=[X(2)]^2/X(3) \tag{2}$$

$$\phi(n)=[X(n-1)]^2/X(n-2) \tag{3}$$

如果是中间数据空缺，一般可使用非邻均值生成法，例如数列

$$\{X\}=[X(1),X(2),\cdots,X(k-1),\phi(k),X(K+1),\cdots,X(n)] \tag{4}$$

式中，$\phi(k)$为缺失数据，则对其补全的公式为：

$$\phi(k)=0.5X(k-1)+0.5X(k+1) \tag{5}$$

2.2 数据标准化方法

从年鉴中选出的候选指标具有不同的量纲，而且不同指标的数量级差别也很大。神经网络进行学习训练以及预测负荷时，必须先对输入输出数据进行标准化，使各项指标间具有可比性[6]。数据标准化的方法主要有以下两种：

(1)极值法。利用指标的极大值和极小值，对候选指标进行无量纲转化，常用的公式有：

$$x_i'=\frac{x_i-\min x_i}{\max x_i-\min x_i} \tag{6}$$

式中，$\min x_i$ 表示指标的极小值，$\max x_i$ 表示指标的极大值，x_i'表示标准化后的值，结果在 0～1 之间。

(2)标准差法。标准差法的计算公式为：

$$x_i'=\frac{x_i-\bar{x}}{s} \tag{7}$$

式中，$\bar{x}$ 表示指标的均值，x_i'表示标准化后的值，结果会超出 0～1 区间。

2.3 聚类分析

数据挖掘就是使用模式识别、统计和数学技术，从大量的数据中发现有意义的新关系、模式和趋势的过程。也就是从海量数据中挖掘出可能有潜在价值的信息技术[7]。数据挖掘能够对数据进行分类、聚类、关联性分析以及预测。

聚类分析能够识别出数据间的关联规则，并把这些数据分成若干类。通常聚类分析分为 Q 型聚类和 R 型聚类。其中 Q 型聚类分析是以距离来度量样品之间的相似程度。当使用 P 个指标描述每个样品不同方面的性质时，就形成了一个 P 维的向量。再把 N 个样品看成 P 维向量空间中的 N 个点，则两个样品间的相似程度就可用 P 维空间中的两点距离公式来度量。令 d_{ij} 表示样品 X_i 与 X_j 的距离，常用的距离公式有以下几种。

(1)闵可夫斯基距离。闵氏距离的公式如下：

$$d_{ij}(q)=\left(\sum_{k=1}^{p}|X_{ik}-X_{jk}|^q\right)^{1/q} \tag{8}$$

式(4)中 q 是一个可变参数。当 $q=1$ 时，就是曼哈顿距离；当 $q=2$时，就是欧氏距离；当 $q=\infty$时，就是切比雪夫距离。

在遇到多元数据分析的问题时，以上几种算法存在两个问题：1)没有考虑到总体变异对“距离”远近的影响；2)闵氏距离受到变量量纲的影响，这对多元数据的处理是不利的。

(2)标准化欧氏距离

标准化欧氏距离是针对欧氏距离的缺点而作的一种改进方案。先将各个分量都标准化到均值、方差相等，标准化变量的数学期望为 0，方差为 1。标准化过程用公式描述为：

$$X_i^*=\frac{X_i-m}{s} \tag{9}$$

式中，X_i^* 为标准化后的值，X_i 为标准化前的值，m 为分量的均值，s 为分量的标准差。经过推导可以得到两个向量 X_i 与 X_j 间的标准化欧氏距离的公式：

$$d_{ij}=\sqrt{\sum_{k=1}^{n}\left(\frac{X_{ik}-X_{jk}}{s_k}\right)^2} \tag{10}$$

2.4 失真数据的修正

对选出的社会经济指标进行初步分析，画出其动态折线图，从图形中观察数据变动的轨迹，特别是异常数值和转折点，分析其原因。对于查出的异常数值，若记历史数据序列为 $\{X\}$，其表达式为：

$$\{X\}=[X(1),X(2),\cdots,X(n)] \tag{11}$$

设 $X_0=[\sum_{i=1}^{n}X(i)]/n$，如果

$$X(i)>X_0\times 1.2 \tag{12}$$

或者

$$X(i)<X_0\times 0.8 \tag{13}$$

则认为 $X(i)$为失真数据，此时可将失真数据视为缺失数据，利用数据补全的方法进行补全。

2.5 动态神经网络

神经网络分为两大类：静态神经网络和动态神经网络，动态神经网络又分为有反馈和无反馈两种。无反馈的动态神经网络是指网络的输出不仅依赖于当前的输入，而且依赖于前面的输入。有反馈的动态神经网络是指网络的输出不仅依赖于当前和之前的输入，也依赖于之前的输出[8]。动态神经网络最主要的应用就是时间序列的预测。

时间序列预测是通过对预测目标时间序列的处理，来研究其变化趋势。Matlab 的时间序列预测工具箱提供 3 种网络，分别用来解决 3 种不同的非线性时间序列问题[9,10]：外部输入的非线性自回归 NARX(nonlinear autoregressive with external input)、非线性自回归 NAR(nonlinear autoregressive)和非线性输入输出(nonlinear input-output)。

本文采用第一种 NARX 网络(其是一种有反馈的动态神经网络，也可以看作是有时延输入的 BP 神经网络)，再加上输出到输入的时延反馈网络。本文使用的 NARX 网络结构如图 1 所示。

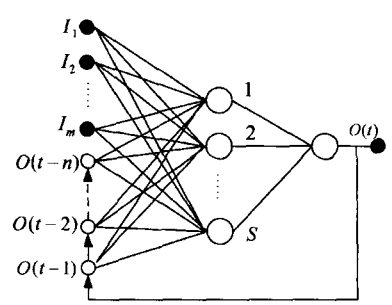

图 1 NARX 递归网络结构

从图 1 可知，NARX 网络的输入由两部分组成，一部分为外部输入，另一部分为输出到输入的时延反馈。假设 $I(t)$和 $O(t)$分别为网络在 t 时刻的外部输入和输出，输出反馈的时延长度为 n_o，外部输入的时延长度为 n_i，则 NARX 网络的输出为：

$$O(t)=f(O(t-1),O(t-2),\cdots,O(t-n_o),I(t-1),I(t-2),\cdots,I(t-n_i)) \quad (14)$$

此时，NARX 网络的输出不仅依赖于当前和之前的输入，同时也依赖于之前的输出。

3 实验分析与结果

3.1 数据收集

本文研究的数据选自中国统计年鉴及历年电力年鉴，其中全社会用电量是本文研究的核心对象。年度负荷预测的历史数据一般以 5～10 年为宜，故本文选取 2000 年至 2010 年历史数据。从统计年鉴中选取能够表征经济发展的 8 个方面，选取 22 项指标作为候选指标，组成图 2 所示的相关因素数据库。在收集数据时，由于中国统计年鉴中工业用电量只统计到 2009 年，使用式(3)对 2010 年的数据进行了补全。

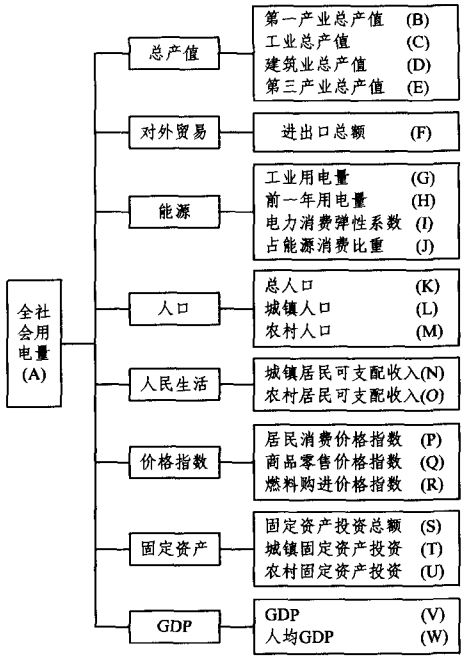

图 2 本文选择的候选指标

3.2 数据标准化

极值法和标准差法各有特点：极值法对指标数据个数和分布要求不是很严格，转化后的数据在 0～1 之间，便于做进一步的处理；而标准差法一般在指标的数据呈正态分布时使用，转化结果会超出 0～1 区间，有时会影响进一步的数据处理。考虑到标准化后的数据受实际值、最大值和最小值的影响，本文在数据标准化时使用的是极值法，选用式(1)、式(2)进行数据处理。

3.3 聚类分析

根据式(8)—式(10)，以及标准化后的候选指标数据，计算出全社会用电量与其他候选指标之间的几种类间距离，计算结果如表 1 所列。类间距离越小，说明该指标与全社会用电量的关系越密切，选用它可以较好地反映全社会用电量增长的一般性规律。

从表 1 的计算结果可以看出，电力消费弹性系数、总人口、农村人口以及价格指数的类间距离值较大，说明这些指标的变化与全社会用电量的增长关联性相对较小。而工业产业总值、工业用电量、前年用电量以及 GDP 的类间距离值较小，说明这些指标的变化与全社会用电量的增长关联性较大。同时可以看出，3 种类间距离的计算结果基本一致。

表 1 全社会用电量与其他指标之间的类间距离

候选指标	Euclidean	Chebychev	Manhattan
B	0.2769	0.1823	0.6922
C	0.2786	0.1126	0.8138
D	0.5083	0.2109	1.5188
E	0.5676	0.2859	1.6031
F	0.3018	0.1404	0.8332
G	0.0321	0.0221	0.0791
H	0.1025	0.0720	0.1950
I	1.1939	0.5504	3.5853
J	0.9385	0.5226	2.4022
K	1.3676	0.6212	3.9540
L	0.8545	0.3681	2.4942
M	1.3518	0.6760	3.7209
N	0.1393	0.0759	0.3629
O	0.0995	0.0601	0.2481
P	1.3502	0.6241	3.8834
Q	1.3203	0.6062	3.7974
R	1.2940	0.6195	3.6814
S	0.7411	0.2816	2.2615
T	0.7669	0.2889	2.3429
U	0.5722	0.2341	1.7253
V	0.3097	0.1363	0.9079
W	0.2748	0.1253	0.7951

综合表 1 的结果，选出在 3 种类间距离中都最小的 9 个指标，分别是第一产业总产值、工业总产值、工业用电量、进出口总额、前一年的用电量、城镇居民可支配收入、农村居民可支配收入、GDP、人均 GDP。

3.4 失真数据的修正

对选出的 9 个指标以及全社会用电量的数据关系进行进一步分析，画出动态折线图，如图 3 所示，从图 3 中可看出 F 指标(进出口总额)的数据出现较大转折，需进一步处理。

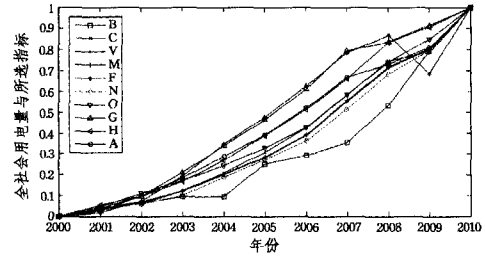

图 3 修正前

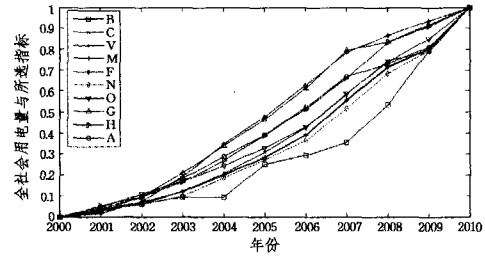

图 4 修正后

利用式(12)和式(13)，计算出 F 指标在 10 点处是失真数据。确定 F 指标在 10 点处是失真数据后，将其看成缺失数据，利用式(5)对数据进行修正补全。补全后重新画出动态折线图，如图 4 所示。此时可以将 9 个指标作为神经网络的训练数据。

3.5 NARX递归网络预测

本文使用NARX神经网络进行预测。首先对所使用的隐含层神经元数、网络的输出反馈时延长度以及训练函数进行了优选训练。

(1)隐层数的确定。隐层数的多少决定了网络的误差,但过多的隐层数会导致网络复杂化,增加网络训练时间和过拟合的倾向。因此隐层数定为1。

(2)隐层节点数的确定。隐层节点数对神经网络的性能有很大影响,但目前将理论上科学的方法用于隐层节点数确定。一般采用以下经验公式确定隐层节点数 n:

$$n=\sqrt{l+m}+k \tag{15}$$

式中,l 为输入层节点数,m 为输出层节点数,k 为1～10的常数。对应到预测模型中,输入层节点数为9,输出层节点数为1,因此,隐层节点数 n 的取值范围为5～14。

(3)各层函数的确定。单隐层BP神经网络输入层和隐层传递函数一般使用tansig函数,输出层传递函数使用purelin函数。隐层训练函数以及隐层节点数需要通过多组实验进行优选。

优选实验结果如表2所列。其中SSE(Sum Squared Error)为误差平方和,是拟合数据和原始数据对应点的误差的平方和,SSE越接近于0,说明预测模型越好。

表2 优选实验结果

训练函数	隐层节点数	SSE
trainbr	12	7.7408e−009
trainrp	10	7.9882e−004
trainlm	10	6.4846e−004
traingdm	11	0.0071
traingdx	11	0.0054

从表2中可以看出,通过优选实验,训练函数trainbr,隐含层节点数为12时,神经网络的误差达到令人满意的效果。其网络结构如图5所示。

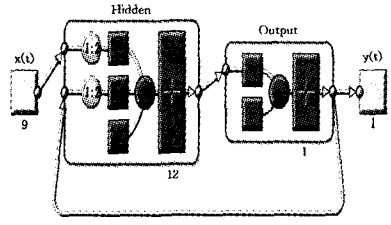

图5 NARX神经网络

使用时间序列神经网络进行预测时,有3个非常重要的参数,即误差自相关性、输入和误差间相关性以及时间序列响应。

对于一个完美的预测模型,误差自相关性应该是一个非零的自相关函数,只应该出现在零延迟处,这说明预测误差互不相关。除了零延迟这一点以外,误差自相关性应该落在95%的置信度内,如图6所示。

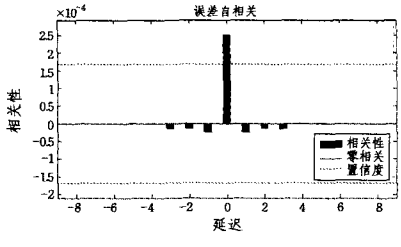

图6 误差自相关性

输入和误差互相关性证明误差与输入序列是如何相关的,对于一个完美的预测模型,此时所有的相关性都是零,一般相关性会落在零延迟附近的置信界限内,如图7所示。

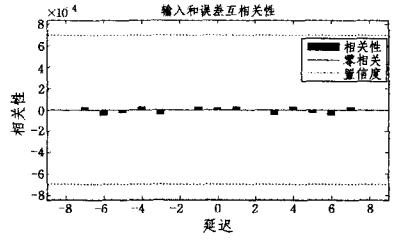

图7 输入和误差互相关性

时间序列响应显示了输入、期望输出和误差在时间轴上的变化情况,也指示出选择的用于训练、测试和验证的时间点,如图8所示。

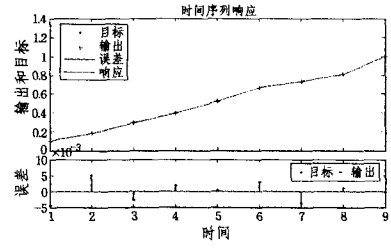

图8 时间序列响应

结束语 本文将数据挖掘的聚类分析算法和思路应用到用电量增长关联性分析中,从2000年后的年鉴中的历史数据中,发现了与全社会用电量增长关系密切的若干指标从而拓展了数据挖掘的应用领域,为长期负荷预测模型输入数据的选取提供了理论依据。该方法和思路也可用于分析电力系统其他类似问题。

参 考 文 献

[1] 康重庆,夏清,张伯明.电力系统负荷预测研究综述与发展方向的探讨[J].电力系统自动化,2004,28(17):1-11

[2] 李智勇,陈志刚,徐政,等.中国全社会用电量增长主导因素辨识[J].电力系统自动化,2010,34(23):30-35

[3] 张石,张瑞友,汪定伟.基于DPCA-BP神经网络的中长期电力负荷预测方法[J].东北大学学报:自然科学版,2010,31(4):483-485

[4] 刘瑾,杨海马,陈抱雪,等.神经网络在电力负荷预测中的应用[J].自动化仪表,2012,33(9):21-24

[5] 毛李帆.中长期负荷预测的异常数据辨识与缺失数据处理[J].电网技术,2010,34(7):148-152

[6] 程玉桂,黎明,林明玉.基于遗传算法和BP神经网络的城区中长期电力负荷预测与分析[J].计算机应用,2010,30(1):224-226

[7] Shiu A. LAM P L. Electricity consumption and economic growth in China[J]. Energy Policy,2004,32(1):47-54

[8] Cai L, Ma S Y, Cai H T, et al. Prediction of SYM-H index by NARX neural network from IMF and solar wind data[J]. Sci China Ser E-Tech Sci,2009,52(10):2877-2885

[9] 李艳红,雷金辉.电力负荷时间序列预测的应用与研究[J].科学技术与工程,2011,11(4):860-864

[10] 代小红,王光利.L-M优化BP算法在短期负荷预测中的应用[J].计算机科学,2011,38(7):265-267