

基于电力大数据的地区E-GDP值预测

田世明¹, 龚桃荣¹, 黄小庆², 于文龙²

(1. 中国电力科学研究院有限公司 需求侧多能互补优化与供需互动技术北京市重点实验室, 北京 100192;

2. 湖南大学 电气与信息工程学院, 湖南 长沙 410082)

摘要:为通过电力发展和使用数据评估一个地区的经济发展水平,提出一种表征地区国内生产总值(GDP)发展趋势的类GDP值(E-GDP)的预测方法。该方法基于多源电力大数据和动态贝叶斯网络(DBN)机器学习,采用灰色关联分析法筛选出与GDP变化趋势关联度较大的关键电力数据。利用格兰杰因果分析确定与GDP变化具有因果关联关系的电力指标,并确定各电力指标间的因果关系。进一步运用所得出的因果关系建立DBN预测获得E-GDP。最后将所提方法应用于上海市E-GDP预测,算例结果表明所提方法可以准确地预测地区E-GDP值,同时还可预测得出GDP的概率分布情况。

关键词:灰色关联分析;格兰杰因果分析;动态贝叶斯网络;机器学习;GDP;电力大数据

中图分类号:TM 73;F 416.61

文献标志码:A

DOI:10.16081/j.epae.201911028

0 引言

作为国民经济的基础产业,电力与各行各业的发展息息相关。由于电能无法大规模储存,实时的电力数据犹如支持经济社会有序运行的“血液”,时刻反映经济社会的发展状态^[1]。进行电力与经济关系研究对于把握国家与地区经济发展趋势具有重要意义。

目前,在电力与经济关系的定量分析方面,文献[2-3]提出了电力与经济关系的误差修正模型;文献[4]采用多元谱分析方法确定电力与经济发展周期的关系;文献[5]采用时间序列分析方法中的协整理论分析电力消费与经济增长关系。这些方法建模复杂、实用性差,尤其是协整理论和向量误差修正模型所得出的结果与所研究的时间序列自身的特点有关,不能适用于任意时间序列^[6],普适性差、实用性弱。

同时,以上研究所用到的电力数据局限于用电量与发电设备装机容量等少数电力数据,数据支撑不够。尤其是当前我国经济正在经历发展模式转型,环保压力日益增大,节能增效和电能替代技术的应用范围逐渐扩大,出现国内生产总值(GDP)增长但用电量下降的情况,导致电力数据与GDP之间的关系日益复杂。如2012—2016年,中国全社会用电量增速分别低于GDP增速2.3%、0.2%、3.2%、5.9%和1.8%^[7];湖南省2014年和2015年的用电量增速分别低于GDP增速9.0%、7.4%,单位GDP电耗降幅均超过6.0%,远大于历史同期平均值,且在极端炎热

的2016年和2017年,单位GDP的电耗值仍有所下降^[8]。由此可见,单一数据不能准确反映GDP变化趋势。

本文结合当前多源化电力大数据背景^[9-10],从电力大数据出发,挖掘电力数据与经济之间的相关关系,应用机器学习算法,建立E-GDP预测新方法。在此,将通过电力数据预测的类GDP值称为E-GDP。由于电力大数据类型多、参考应用价值高、数量规模庞大并且贯穿于社会生产生活的各个环节,识别、筛选与经济相关性较强的电力数据是进行E-GDP预测的首要问题。

实际中,能进行变量选择的算法较多。如文献[11]采用的前进逐步选择与逆向逐步选择法,方法大体一致,只是运行方向相反,该方法节省计算时间但结果仅接近于最优而非最优;文献[12]采用的分阶段选择法,该方法得出的结果最优但计算时间过长;文献[13]采用的lasso因子选择法较上述3种方法加入了对复共线性数据的处理,使得计算结果更为准确。但以上提到的变量选取方法多用于线性回归预测。而本文选择变量因子的目的在于分析后续因果关系,以形成因果网络并运用动态贝叶斯网络DBN(Dynamic Bayesian Network)进行推理预测,因此上述方法并不适用。文献[14]采用一种多因素统计的灰色关联分析法,与相关分析法、回归分析法等传统的多因素分析方法相比,其具有对数据要求较低且计算量较小、建模仿真容易实现等优势。因此,本文利用灰色关联分析法进行电力指标筛选。

在筛选出与GDP变化关联度较大的电力数据后,本文对电力数据与GDP进行格兰杰因果分析,确定主要电力指标及其与GDP之间的因果关系。然后,本文运用上述因果关系建立DBN,构造GDP预测模型,并进行算例仿真。

收稿日期:2019-03-11;修回日期:2019-09-29

基金项目:国家电网公司科技项目(520940180016)

Project supported by the Science and Technology Project of SGCC(520940180016)

1 电力经济指标的选择与因果分析

选择合适、严谨科学的电力指标是进行 GDP 预测的前提,选取指标应考虑的因素包括指标重要程度及指标数量等。另外对于初选指标,首先应尽量全面选取,随后在指标体系优化时进一步筛选初选指标。

1.1 电力经济指标初选及其优化

指标选取具有明确的流程,可概括为 3 个步骤。

①指标初选。基于调研和电力经济关系分析,初步确定能够表征 GDP 变化趋势的主要电力指标,形成初选的电力经济指标。②将初选的电力指标制成调查问卷,运用德菲尔法即一种匿名的专家问卷调查法,结合专家意见进行指标的增删修改。③运用灰色关联分析法计算各个指标相对于 GDP 变化的灰色关联度,删除与 GDP 变化趋势关联度小于 0.7 的电力经济指标,保留关联度大于等于 0.7 的指标作为优选的电力经济指标。其中,步骤①和②分析文献[5-6]中电力与经济之间的关系,结合调研和专家意见,初步选定电力经济指标为地区年发电量 λ_1 、地区年用电量 λ_2 、地区年末发电设备容量 λ_3 、地区公用变电容量 λ_4 、地区架空线长度 λ_5 、地区电缆长度 λ_6 和地区年最大负荷 λ_7 。

1.2 基于灰色关联分析法的指标优选

1.2.1 灰色关联分析法原理

灰色关联分析法将各分析因素的样本数据作为分析的参考依据,通过计算灰色关联度表征各个分析因素之间关联关系的大小、强弱与次序^[14]。若 2 个因素之间的灰色关联度较大,则表示 2 个因素之间变化态势(指 2 个因素之间的大小、速度、方向等)基本一致;反之,则表示 2 个因素之间变化态势一致性较弱,即 2 个因素之间关联关系弱。总体而言,该方法的核心是因素之间灰色关联度的计算。

1.2.2 基于灰色关联分析法的指标优选步骤

基于灰色关联分析法的指标优选包括 5 个步骤。

(1)确定分析序列。

将地区 GDP 数据作为分析的参考数据序列 A'_0 ,其他的电力数据则形成相应分析序列 $A'_i(i=1,2,\dots,n)$ 。 $n+1$ 列数据序列构成矩阵,如式(1)所示。

$$[A'_0 \ A'_1 \ \cdots \ A'_n] = \begin{bmatrix} a'_0(1) & a'_1(1) & \cdots & a'_n(1) \\ a'_0(2) & a'_1(2) & \cdots & a'_n(2) \\ \vdots & \vdots & & \vdots \\ a'_0(N) & a'_1(N) & \cdots & a'_n(N) \end{bmatrix}_{N \times (n+1)} \quad (1)$$

其中, N 为序列变量的长度; $a'_0(l)$ 为参考变量序列数据, $l=1,2,\dots,N$; $a'_i(l)$ 为变量序列数据, $i=1,2,\dots,n$ 。

(2)变量的无量纲化处理。

电力数据与 GDP 数据在数量级大小规模以及量纲方面均具有较大的差异,所以为了得到准确的分析结论,需要对变量序列进行无量纲化处理。处理后的分析序列如下:

$$[A_0 \ A_1 \ \cdots \ A_n] = \begin{bmatrix} a_0(1) & a_1(1) & \cdots & a_n(1) \\ a_0(2) & a_1(2) & \cdots & a_n(2) \\ \vdots & \vdots & & \vdots \\ a_0(N) & a_1(N) & \cdots & a_n(N) \end{bmatrix}_{N \times (n+1)} \quad (2)$$

其中, $a_i(l)$ 为处理后的变量序列数据, $i=0,1,\dots,n$ 。

采用均值法对变量序列数据进行无量纲化处理,即:

$$a_i(k) = a'_i(l) / \left(\frac{1}{N} \sum_{l=1}^N a'_i(l) \right) \quad i=0,1,\dots,n \quad (3)$$

(3)最大差、差序列以及最小差的计算。

计算式(2)中第一列(参考序列)与其余各列(分析比较序列)对应期差值的绝对值,所形成的绝对差值矩阵为:

$$\Delta = \begin{bmatrix} \Delta_{01}(1) & \Delta_{02}(1) & \cdots & \Delta_{0n}(1) \\ \Delta_{01}(2) & \Delta_{02}(2) & \cdots & \Delta_{0n}(2) \\ \vdots & \vdots & & \vdots \\ \Delta_{01}(N) & \Delta_{02}(N) & \cdots & \Delta_{0n}(N) \end{bmatrix}_{N \times n} \quad (4)$$

其中, $\Delta_{0i}(l) = |a_0(l) - a_i(l)|, i=1,2,\dots,n$ 。

(4)计算关联系数。

对绝对差值阵中数据作如下变换:

$$\xi_{0i}(l) = \frac{\Delta_{\min} + \rho \Delta_{\max}}{\Delta_{0i}(l) + \rho \Delta_{\max}} \quad (5)$$

其中, $i=1,2,\dots,n; l=1,2,\dots,N$ 。 $\Delta_{\max}$ 和 $\Delta_{\min}$ 分别为绝对差值矩阵中的最大数和最小数,即最大差与最小差; ρ 为分辨系数,在(0,1)内取值,一般取 0.1~0.5, ρ 值越小越能提高关联系数间的差异; $\xi_{0i}(l)$ 为关联系数,为不超过 1 的正数,反映第 i 个比较序列 A_i 与参考序列 A_0 在第 l 期的关联程度。

关联系数矩阵为:

$$\xi = \begin{bmatrix} \xi_{01}(1) & \xi_{02}(1) & \cdots & \xi_{0n}(1) \\ \xi_{01}(2) & \xi_{02}(2) & \cdots & \xi_{0n}(2) \\ \vdots & \vdots & & \vdots \\ \xi_{01}(N) & \xi_{02}(N) & \cdots & \xi_{0n}(N) \end{bmatrix}_{N \times n} \quad (6)$$

(5)灰色关联度的计算。

比较序列 A_i 与参考序列 A_0 之间的灰色关联度通过计算 N 个关联系数的平均值得到, A_i 与 A_0 的关联度为:

$$r_{0i} = \frac{1}{N} \sum_{l=1}^N \xi_{0i}(l) \quad i=1,2,\dots,n \quad (7)$$

根据关联度进行电力经济指标的筛选,以 $r_{0i} \geq 0.7$ 为标准指标预测地区GDP。

1.3 基于格兰杰因果分析的指标因果关系分析

1.3.1 格兰杰因果分析方法的基本思想

格兰杰因果分析法基本思想的数学描述为:假设变量 x 的变化引起变量 y 变化,则 x 的变化相较于 y 应该具有时间先行性。即 x 变化在时间上先于 y ,并且运用 x 的过去观测值进行 y 的预测时具有显著性,可以极大地提高 y 的预测精度^[8-9]。因而格兰杰因果分析法对于研究变量的时序因果关系具有重要作用。

1.3.2 格兰杰因果分析法步骤

(1) 指标平稳性的检验。

采用单位根方法进行平稳性检验。单位根方法的检验主要包括3个模型。模型2在模型1的基础上加入截距项,模型3在模型1的基础上加入了趋势项与截距项。3个模型对应的表达式分别为^[10-11]:

$$\Delta h_{t1} = \delta h_{t-1} + \sum_{k=1}^m \varphi_k \Delta h_{t-k} + \varepsilon_t \quad (8)$$

$$\Delta h_{t2} = \mu + \delta h_{t-1} + \sum_{k=1}^m \varphi_k \Delta h_{t-k} + \varepsilon_t \quad (9)$$

$$\Delta h_{t3} = \mu + \omega t + \delta h_{t-1} + \sum_{k=1}^m \varphi_k \Delta h_{t-k} + \varepsilon_t \quad (10)$$

其中, ε_t 为随机误差项; $\mu, \varphi_k, \delta, \omega$ 为系数; $\Delta h = \{\Delta h_1, \Delta h_2, \dots, \Delta h_t\}$ 为时间序列 $h = \{h_1, h_2, \dots, h_t\}$ 经过差分计算后得到的差分序列; m 为差分序列的最优滞后阶数; k 为差分序列的滞后阶数; t 为时间序列的时间标度。

对于时间序列进行平稳检验,将其分别构造为式(8)~(10)的模型形式,利用单位根临界值表检验零假设 $H_0: \delta=0$ 。如果3个模型中任意一个模型的检验结果拒绝零假设,则表明原时间序列拒绝存在单位根的原假设,判定该序列为平稳时间序列。反之,判定时间序列非平稳^[2,15]。

(2) 协整校验。

对于2个时间序列数据,其存在协整关系的前提是:2个时间序列的单整阶数必须相同(如果一个时间序列原序列不稳定,但经过 d 次差分后成为稳定序列,则该序列被称为 d 阶单整);序列协整关系的检验主要采用Johansen协整校验^[2-3],检验过程中依据赤池信息准则来确定滞后阶数,如果相应的迹统计量小于临界值则拒绝原假设,证明存在协整关系。

(3) 格兰杰因果关系校验。

当证明时间序列之间具有协整关系后,才能将时间序列进行格兰杰因果关系校验,检验模型为:

$$w_t = c + \sum_{j=1}^p \alpha_j \Delta w_{t-j} + \sum_{k=1}^q \beta_k \Delta h_{t-k} + \varepsilon_t \quad (11)$$

其中, q, p 为最优滞后阶数; j 为滞后阶数; c 为系数;

Δh_{t-k} 为差分序列 Δh 在时间标度为 $t-k$ 时所对应的滞后项变量数值; Δw_{t-j} 为差分序列 Δw 在时间标度为 $t-j$ 时所对应的滞后项变量数值; β_k 为差分序列 Δh 在时间标度为 $t-k$ 时所对应的滞后项协整系数; α_j 为差分序列 Δw 在时间标度为 $t-j$ 时所对应的滞后项协整系数; w_t 为时间序列 $\{w_1, w_2, \dots, w_t\}$ 在时间标度为 t 时所对应的变量数值。

检验的零假设 H_0 为不存在 $x \rightarrow y$ 的格兰杰原因,即 $H_0: \beta_1 = \beta_2 = \dots = \beta_q = 0$ 。若零假设 H_0 成立,则有:

$$w_t = c + \sum_{j=1}^p \alpha_j \Delta w_{t-j} + \varepsilon_{t0} \quad (12)$$

其中, ε_{t0} 为对应于零假设成立的随机误差项。

$$F = \frac{(S_1 - S_0)/q}{S_0/(T - q - p - 1)}$$

为服从自由度 $(q, T - p - q - 1)$

的 F 分布, S_1 为式(11)的残差平方和, S_0 为式(12)的残差平方和, T 为样本容量。滞后阶数可根据赤池信息准则确定。比较 F 统计量与临界值的大小即可得检验结果。当 F 统计量大于等于临界值时,零假设 H_0 不成立,可判定 x 是 y 的格兰杰原因;反之可判定 x 不是 y 的格兰杰原因。

2 基于DBN建立GDP预测概率模型

由计量经济学分析可知,电力指标对于经济变化具有先行性的作用,即二者的因果关系表现为:当前时期经济的变化不仅受到当前时期电力指标的影响,而且受到前一时期电力指标的影响;与此同时,经济变化同时受到国际金融环境、期货政策等不确定因素的影响。DBN可充分体现变量间因果关系时序变化特点,是一种基于不确定性的概率推理网络算法,因而适用于本文中所描述的电力经济指标因果关系,并根据因果关系进行GDP预测。

2.1 DBN理论

贝叶斯网络理论形成于1988年^[16-17],由国外学者Judea Pearl率先提出,经过不断完善,是贝叶斯方法与图形理论的有机结合。该网络以概率理论知识及图理论为基础,将网络变量间的因果关系以有向无环图和条件概率的形式表达,是解决工程实践中复杂问题以及具有不确定性问题的有效方法,目前广泛应用于人工智能方面。

贝叶斯网络整体而言,主要分为DBN与静态贝叶斯网络SBN(Static Bayesian Network)^[18]。两者的共同组成部分为网络结构和网络参数:①网络结构的形式主要为非循环有向图,也称为有向无环图,是网络模型中变量之间因果关系的最直观的体现,主要由节点变量以及有向边构成,其中,节点变量表示网络中的随机变量,有向边表示节点变量间的因果关系;②网络参数主要以节点变量间条件概率表的

形式表示,以条件概率数值的大小表示节点变量间因果关系的相互依赖程度。

当网络参数和网络结构确定时,贝叶斯网络可以结合给定的初始条件,利用概率理论推理得到各个网络节点变量的概率分布,该过程推导详见文献[19],在此不再赘述。

DBN 是将 SBN 在时序轴上展开而形成随机概率图模型。DBN 将不同时间片上的 SBN 按照时间顺序以有向边的形式依次连接,则 DBN 不仅包含 SBN 描述系统不确定性因果关系的优点,还在模型中加入时序因素。因此 DBN 是处理具有时序信息相关因果关系的概率不确定性推理方法。

2.2 利用 DBN 构造 GDP 概率预测模型

利用 DBN 建立 GDP 概率预测模型,首先应选择节点变量集,其次形成节点变量状态域,而后构建对应的网络结构。归结起来可分为如下步骤:①选取合适的节点变量,确定节点变量集;②分析节点变量间因果关系;③划分变量状态,形成节点变量状态域;④形成 SBN;⑤形成 DBN。

参考第 1 节指标筛选中灰色关联分析法和格兰杰因果关系检验法的分析结果可完成步骤①和②。

对于步骤③,由于各个变量之间的量纲不同,为便于后续网络参数的学习与计算,需要将网络变量进行状态划分,形成变量的状态域,具体做法为:参考给定的变量历史数据的取值变化情况,结合实际,将其划分形成各自的状态区间,并赋予状态值。

根据步骤①—③的分析结果,将变量间的因果关系形成 SBN,如图 1 所示。图中, $X_a^b(a=t, t+1; b=1, 2, 3)$ 为变量 X_b 在时间标度 a 下的状态值。

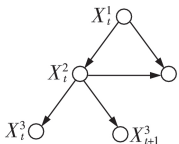


图 1 SBN 结构

Fig.1 Structure diagram of SBN

将 SBN 在时间轴上展开,即加入时间信息形成 DBN,以全面描述网络节点变量在同一时刻以及不同时刻之间的因果关系,如图 2 所示。图中变量定义参考图 1。

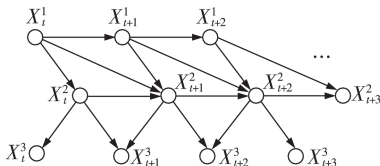


图 2 DBN 结构

Fig.2 Structure diagram of DBN

3 GDP 预测概率分布的计算

3.1 GDP 预测概率模型节点参数的计算

已知 GDP 概率预测模型结构计算节点参数,通过对已知的大量历史数据进行训练学习,计算得出反映节点变量间因果关系依赖强弱的条件概率。

结合本文电力大数据的应用背景,文中所用到的历史数据均是完整的,节点参数的计算方法采用极大似然估计法^[17]。该方法具体推导见附录 A。

3.2 预测分布的概率推理

在网络结构和网络节点参数已知的条件下,可通过 DBN 概率推理算法计算得到预测时刻 GDP 的数值概率。常用的网络推理算法主要有 2 种^[17,19-20]:①精确推理算法,该算法主要应用于网络规模较小,即网络结构较为简单的情况,计算结果较为准确;②近似推理算法,该算法主要应用于网络规模较大,即网络拓扑结构复杂、网络节点变量较多的情况,以解决计算复杂度高、计算量过大的问题,计算结果精确度稍差。本文采用精确推理算法中较为常用的联合树推理算法计算预测时刻 GDP 数值的概率分布。

4 算例分析

根据调研数据及上海市统计年鉴数据,可得 2002—2018 年上海市各项电力数据及 GDP 数据。

以 GDP 数据为参考序列进行灰色关联分析,计算各电力指标相对于 GDP 的灰色关联度: λ_1 的灰色关联度为 0.677; λ_2 的灰色关联度为 0.713; λ_3 的灰色关联度为 0.8596; λ_4 的灰色关联度为 0.8849; λ_5 的灰色关联度为 0.7842; λ_6 的灰色关联度为 0.6916; λ_7 的灰色关联度为 0.7965。

根据灰色关联度的大小,筛选出与 GDP 变化趋势较为接近的电力指标。

将筛选出的电力指标与 GDP 进行格兰杰因果关系检验,为排除异方差性,检验时对电力指标与 GDP 的时序变量取自然对数^[2]。

首先运用 Eviews6.0 软件对数据进行单位根检验,检验结果表明 GDP 取值 γ_{GDP} 与各电力指标数据原对数序列均不稳定。将各个序列进行一阶差分运算后经过单位根检验,均通过平稳性检验,则各个序列为单整序列;对各个序列进行格兰杰因果关系检验,校验时采用 VAR 模型进行最优滞后期的确定,最终得到的因果关系如表 1 所示。表中仅列出格兰杰因果关系成立的情况;“*”、“**”分别代表 5%、10% 的置信水平,如伴随概率为 0.024 8**表示在 5% 的置信水平下原假设成立,证明有 95% 的可能拒绝原假设,即 $\ln \lambda_4$ 是 $\ln \gamma_{GDP}$ 的格兰杰原因。

根据表 1 中的校验结果并结合格兰杰因果校验的基本思想,同时考虑到时间序列数据均具有自相关性的特点^[2,5](如经济时间序列 GDP 在当前时刻的

值与其前一时刻的值相关),建立如图3所示的GDP预测DBN结构。图中, $\gamma_{GDP,t}$ 为在时间标度 t 时所对应的地区GDP值。后续的网络参数学习、GDP预测的计算过程利用FULLBNT工具箱实现。

表1 格兰杰因果校验结果

Table1 Results of Granger causal verification

零假设	滞后阶数	F 检验值	伴随概率
$\ln\lambda_4$ 不影响 $\ln\gamma_{GDP}$	4	39.490 70	0.024 8**
$\ln\gamma_{GDP}$ 不影响 $\ln\lambda_4$	4	4.970 60	0.174 4
$\ln\lambda_7$ 不影响 $\ln\gamma_{GDP}$	4	10.676 80	0.087 5*
$\ln\gamma_{GDP}$ 不影响 $\ln\lambda_7$	4	4.660 35	0.184 4
$\ln\lambda_4$ 不影响 $\ln\lambda_2$	1	2.219 57	0.164 4
$\ln\lambda_2$ 不影响 $\ln\lambda_4$	1	3.320 96	0.095 1*

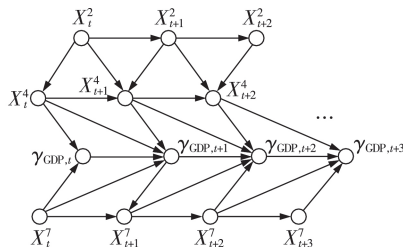


图3 用于GDP预测的DBN结构

Fig.3 DBN structure for GDP prediction

BNT是由国外学者 Kevin P. Murphy 在 MATLAB 环境下开发的用于贝叶斯网络学习的工具箱^[21],其基本计算过程主要包含结构学习、参数学习以及概率推理3个部分,并分别有相对应的函数,可进行SBN和DBN的分析计算。利用BNT工具箱进行的DBN推理计算流程见附录B,推理流程图见附录B中图B1所示。

根据上述流程,以图3中所示电力指标以及GDP的2002—2015年年增长率数据为模型参数学习样本,分别预测得到2016—2018年上海市GDP年增长率的概率分布如图4所示。

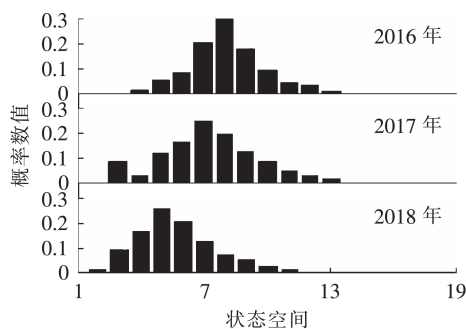


图4 2016—2018年GDP增长率概率分布

Fig.4 Probability distribution of GDP growth rate from 2016 to 2018

根据历史数据,将GDP年增长率在2%~21%的范围内,以1%为间隔,连续离散为19个状态,分别以编号1—19表示。

将图4中的分布概率最大值所对应的状态区

间的中间值取为GDP增长率的预测值,可以得到2016年上海市GDP年增长率为9.5%,即GDP数值为28 079.60亿元,与实际统计数据28 178.65亿元仅有0.4%的误差,同理得到2017年与2018年上海市GDP年增长率分别为8.5%与6.5%,继而得到2017年GDP数值为30 466.37亿元,与实际统计数据30 632.99亿元仅有0.54%的误差;2018年GDP预测值为32 446.68亿元,与实际统计数据32 679.87亿元有0.71%的误差。验证了该方法的正确性。

将BP神经网络法、灰色预测法、灰色神经网络与本文应用的DBN方法的预测结果进行比较,2018年GDP预测分析如表2所示。

表2 DBN与智能算法比较

Table 2 Comparison between DBN and intelligence algorithms

方法	预测结果/亿元	实际数据/亿元	相对误差/%
DBN	32 446.68	32 679.87	0.71
BP神经网络	31 464.18	32 679.87	3.72
灰色预测	31 402.09	32 679.87	3.91
灰色神经网络	31 843.27	32 679.87	2.56

由表2可知,DBN相对于智能算法以及组合智能算法,其预测结果较为准确。这是因为智能算法都是确定性的预测方法,即在某一时刻只能够预测得出确定值,同时当数据出现较为明显的波动时,利用确定性方法得出的预测结果会出现明显较大的误差;而DBN中的节点参数是通过大量的历史数据学习获得的,所以出现上述数据波动情况时,DBN会根据历史数据的网络学习情况,预测得出基本与实际情况相符合的概率分布。由此可知,即使不确定因素导致预测存在偏差,得到的预测结果区间并不是实际的预测区间,实际的状态区间也会在预测分布中以概率形式表现出来,这是DBN最大的优点。

5 结论

由计量经济学分析可知,GDP待预测年的经济指标不仅与同时期的电力指标紧密相关,还与前一时期的电力指标直接相关。而DBN可充分体现变量间因果关系时序变化情况的特点,是一种基于不确定性的概率推理网络算法。本文通过灰色关联分析法分析电力数据与GDP变化趋势的关联性,以此筛选关键电力指标;而后通过格兰杰因果分析关键电力指标间的时序因果关系;提出了用于E-GDP预测的DBN概率模型。算例对上海地区实际数据进行分析计算,结果表明该E-GDP预测方法正确有效。相比于其他常规机器学习方法,DBN可预测出下一预测时刻全部可能的GDP变化值及其相对应的概率,所包含的预测信息更为全面可信。

附录见本刊网络版(<http://www.epae.cn>)。

参考文献:

- [1] 许子智,曾鸣. 美国电力市场发展分析及对我国电力市场建设的启示[J]. 电网技术,2011,35(6):161-166.
XU Zizhi,ZENG Ming. Analysis on electricity market development in US and its inspiration to electricity market construction in China[J]. Power System Technology,2011,35(6):161-166.
- [2] 高铁梅. 计量经济分析方法与建模:EVIEWS应用及实例[M]. 北京:清华大学出版社,2009:7-20.
- [3] 张兴平,牛玉琴,赵旭. 中国电力消费协整关系模型[J]. 中国电机工程学报,2008,28(13):114-119.
ZHANG Xingping,NIU Yuqin,ZHAO Xu. Cointegration analysis of electricity consumption in China[J]. Proceedings of the CSEE,2008,28(13):114-119.
- [4] 袁家海,丁伟,胡兆光. 电力消费与中国经济发展的协整与波动分析[J]. 电网技术,2006,30(9):10-15.
YUAN Jiahai,DING Wei,HU Zhaoguang. Analysis on cointegration and comovement of electricity consumption and economic growth in China[J]. Power System Technology,2006,30(9):10-15.
- [5] 陈银峰. 电力需求与经济发展的关联关系研究[D]. 上海:上海交通大学,2010.
CHEN Yinfeng. Relationship between electricity demand and economic growths[D]. Shanghai:Shanghai Jiaotong University,2010.
- [6] 徐青山,王文帝,林章岁,等. 面向行业大数据特征挖掘的电力经理指数指标体系的建立与应用[J]. 电力自动化设备,2015,35(7):15-21.
XU Qingshan,WANG Wendi,LIN Zhangsui,et al. Establishment and application of EMI indicator system orienting to massive industrial data mining[J]. Electric Power Automation Equipment,2015,35(7):15-21.
- [7] 谭显东,单葆国,吴姗姗,等. 经济新常态下中国全社会用电量与GDP增速差距研究[J]. 中国电力,2017,50(12):5-9.
TAN Xiandong,SHAN Baoguo,WU Shanshan,et al. Study on the difference between China's total electricity consumption and economic growth under the new normal[J]. Electric Power,2017,50(12):5-9.
- [8] 陈火焱,陈佳,李凯,等. 湖南省电力消费强度下降原因解析[J]. 中国电力,2018,51(5):172-178.
CHEN Huoyan,CHEN Jia,LI Kai,et al. Analysis on the causes for the decline of electricity consumption intensity in Hunan Province[J]. Electric Power,2018,51(5):172-178.
- [9] WU X,ZHU X,WU G Q,et al. Data mining with big data[J]. IEEE Transactions on Knowledge and Data Engineering,2014,26(1):97-107.
- [10] 郝然,艾芊,肖斐. 基于多元大数据平台的用电行为分析构架研究[J]. 电力自动化设备,2017,37(8):20-27.
HAO Ran,AI Qian,XIAO Fei. Architecture based on multi-variate big data platform for analyzing electricity consumption behavior[J]. Electric Power Automation Equipment,2017,37(8):20-27.
- [11] 周雄斌. 函数线性回归模型中的变量选择问题研究[D]. 重庆:重庆大学,2017.
ZHOU Xiongbin. Study on variable selection in functional linear regression models[D]. Chongqing:Chongqing University,2017.
- [12] 苏莉. 回归模型中变量选择的若干问题研究[D]. 兰州:兰州交通大学,2017.
SU Li. Research on some problems of variable selection in regression model[D]. Lanzhou:Lanzhou Jiaotong University,2017.
- [13] 刘挺坚,刘友波,刁塑,等. 连锁故障中负荷损失数值特征的非参关联分析[J]. 电力自动化设备,2018,38(7):148-154,161.
LIU Tingjian,LIU Youbo,DIAO Su,et al. Nonparametric correlation analysis of numerical feature for load loss in cascading failure[J]. Electric Power Automation Equipment,2018,38(7):148-154,161.
- [14] 胡钊,徐则诚,金哲,等. 基于灰色关联模型的输电线路舞动预警方法[J]. 电力自动化设备,2018,38(2):117-122.
HU Po,XU Zecheng,JIN Zhe,et al. Early warning method of transmission line galloping based on grey relational model[J]. Electric Power Automation Equipment,2018,38(2):117-122.
- [15] HAREDA SHT F N,GHASSEMI F,MORADI M H. Causal inference of gene expression data using a clustering-based extension of Kernel-Granger causality[C]//Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering(ICBME). Tehran,Iran:IEEE,2016:84-88.
- [16] JONGSAWAT N,DECHAROENCHITPONG J. Creating behavior-based rules for snort based on Bayesian network learning algorithms[C]//IEEE International Conference on Science and Technology(TICST). Pathum Thani,Thailand:IEEE,2015:267-270.
- [17] 肖秦琨,高嵩,高晓光. 动态贝叶斯网络推理学习理论及应用[M]. 北京:国防工业出版社,2007:60-90.
- [18] 董雷,周文萍,张沛,等. 基于动态贝叶斯网络的光伏发电短期概率预测[J]. 中国电机工程学报,2013,33(增刊):38-45.
DONG Lei,ZHOU Wenping,ZHANG Pei,et al. Short-term photovoltaic output forecast based on dynamic Bayesian network theory[J]. Proceedings of the CSEE,2013,33(Supplement):38-45.
- [19] 沈斌,温涛. 基于贝叶斯最大似然估计的金融预测[J]. 统计与决策,2018,34(7):85-88.
SHEN Bin,WEN Tao. Financial forecast based on Bayesian maximum likelihood estimation[J]. Statistics and Decision-making,2018,34(7):85-88.
- [20] 陈国胜. 基于动态贝叶斯网络的战场信息预测与评估[D]. 南京:南京理工大学,2013.
CHEN Gusheng. Battlefield information prediction and evaluation based on dynamic Bayesian network[D]. Nanjing:Nanjing University of Science & Technology,2013.
- [21] WU H,JIANG D,ZHAO Y,et al. Dimensional emotion driven facial expression synthesis based on the multi-stream DBN model[C]//The 2012 Asia Pacific Signal & Information Processing Association Annual Summit and Conference(APSIPA ASC). Hollywood,USA:IEEE,2012:1-6.

作者简介:



田世明

田世明(1965—),男,湖南凤凰人,教授级高级工程师,硕士,主要研究方向为智能电网、需求侧管理、大数据分析等(**E-mail**: laotian@epri.sgcc.com.cn);

龚桃荣(1982—),女,湖北汉川人,博士,通信作者,主要研究方向为需求侧管理、大数据分析等(**E-mail**: gongtr@163.com);

黄小庆(1981—),女,湖南岳阳人,副教授,博士,主要研究方向为电气信息技术、电力系统分析(**E-mail**: misshuang504@126.com);

于文龙(1994—),男,河南洛阳人,硕士研究生,主要研究方向为电力系统分析(**E-mail**: 15136360489@163.com)。

Forecasting regional E-GDP value using power big data

TIAN Shiming¹, GONG Taorong¹, HUANG Xiaoqing², YU Wenlong²

(1. Beijing Key Laboratory of Demand Side Multi-Energy Carriers Optimization and Interaction Technique,
China Electric Power Research Institute Co., Ltd., Beijing 100192, China;

2. College of Electrical and Information Engineering, Hunan University, Changsha 410082, China)

Abstract: In order to evaluate the economic development level of a region by using data of power development and utilization, a prediction method of E-GDP (E-Gross Domestic Product) which represents the development trend of regional GDP is proposed. Based on multi-source power big data and DBN (Dynamic Bayesian Network) machine learning, this method can screen out the key power data with a large correlation with the GDP change trend by gray correlation analysis method. Then, Granger causal analysis is used to determine the power indicators that have a causal relationship with GDP changes, and to determine the causal relationship among the various power indicators. Furthermore, the resulting causal relationship is used to establish a DBN to predict E-GDP. Finally, the proposed method is applied to the prediction of Shanghai E-GDP value. The example shows that the proposed method can accurately predict regional E-GDP value, and can also measure the probability distribution of GDP.

Key words: gray correlation analysis; Granger causal analysis; DBN; machine learning; GDP; power big data

(上接第173页 continued from page 173)

Modeling and invulnerability analysis of power communication interdependent network based on community overlapping

LONG Qinfei¹, WANG Tao¹, GU Xueping¹, WANG Tieqiang²

(1. School of Electrical and Electronic Engineering, North China Electric Power University, Baoding 071003, China;

2. Hebei Power Dispatch and Communication Center, Shijiazhuang 050021, China)

Abstract: In order to better reveal the characteristics of the power communication network, a more practical modeling method of power communication network is proposed based on the theory of community overlapping and the model of interdependent network. The characteristics of the actual communication network structure are analyzed, the power grid are divided into communities by point-to-edge community transformation and Markov clustering algorithm, and then the community overlapping nodes are determined. In accordance with the community overlapping nodes and the interdependent network model, a communication network model is established and then the hierarchical topology structure and coupling relationship of the power communication interdependent network system are constructed. A corresponding power communication network is set up for IEEE standard examples, and an improved failure model of the power communication network is established on the basis of information attacks. The invulnerability analysis is performed based on the node loss ratio, load loss ratio, and network relative efficiency index. The results show that the damage to the system caused by the attack is worse with the improvement of the attack method, and the expansion of the network scale can significantly improve the network's tolerance to random attacks. Compared with the traditional model, the proposed model demonstrates the anti-destructive ability of the actual power communication network with high accuracy.

Key words: cyber-physical power system; hybrid system model; overlapping community; interdependent network; invulnerability analysis

附录 A

设 $Q = \{Q_1, Q_2, \dots, Q_z\}$ 这一数据集合表示历史数据，可构造极大似然函数^[19]为：

$$L = \sum_{i=1}^n \sum_{z=1}^M \log_2 P(X_i | P_a(X_i), Q_z) \quad (A1)$$

其中， n 为历史数据中变量总时间跨度； i 为历史数据中变量的时间标度；下标 z 为历史数据组的组别； M 为历史数据组的总数； $P_a(X_i)$ 为变量 X_i 的父节点，在动态贝叶斯网络中可理解为，存在一条由 $P_a(X_i)$ 指向 X_i 的有向线段。

若出现网络节点是离散变量时，定义 $\theta_{i,j,f} = P(X_i = f | P_a(X_i) = j)$ ， $\theta_{i,j,f}$ 为所求的节点条件概率， f, j 为节点 X_i 与其父节点 $P_a(X_i)$ 的不同状态值（该状态值来源于 2.2 节步骤③的状态划分）。则可将式（14）改写为：

$$L = \sum_i \sum_z \log_2 \prod_{j,l} \theta_{i,j,f}^{I_{i,j,f,z}} = \sum_i \sum_z \sum_{j,f} I_{i,j,f,z} \log_2 \theta_{i,j,f} = \sum_{i,j,f} \sum_z I(X_i = f, P_a(X_i) = j | Q_z) \log_2 \theta_{i,j,f} \quad (A2)$$

$$I_{i,j,f,z} = \begin{cases} 1, & \text{在 } Q_z \text{ 中 } (X_i = f, P_a(X_i) = j) \\ 0, & \text{其它} \end{cases} \quad (A3)$$

$\sum_z I(X_i = f, P_a(X_i) = j | Q_z)$ 为计算得出历史数据中情况 $(X_i = f, P_a(X_i) = j)$ 出现次数的统计值。而后 $\theta_{i,j,f}$ 的极大估计值 $\hat{\theta}_{i,j,f}$ 可以采用拉格朗日乘法计算，即：

$$\hat{\theta}_{i,j,f} = \frac{\sum_z I(X_i = f, P_a(X_i) = j | Q_z)}{\sum_z \sum_f I(X_i = f, P_a(X_i) = j | Q_z)} \quad (A4)$$

附录 B

利用 BNT 工具箱进行的 DBN 推理计算流程。

（1）总体而言，常用的网络结构确定方法有 2 种：①基于 BNT 工具箱可以通过对历史数据学习获得网络结构，但该方法对于数据质量要求严格，计算过程耗时较长；②可以根据经验知识的方法确定网络结构，因而本文在专家调研的基础上，采用灰色关联分析法与格兰杰因果关系检验确定网络结构，如图 3 所示。而后采用函数 `mk_dbn(dag,ns)` 形成网络结构。

（2）参数学习，主要计算网络变量间的条件概率表，主要采用 MLE，其在工具箱中相对应的函数为：`Learn_params_dbn()`。

（3）经过步骤（1）、（2）已经完成了 DBN 的构建与参数学习，采用图 B1 中流程及函数进行概率推理：

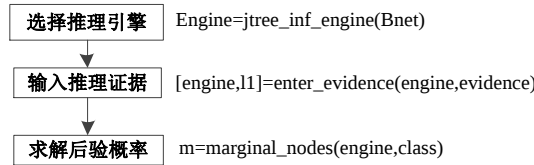


图 B1 DBN 概率推理流程

Fig.B1 DBN probabilistic inference mechanism

如图 B1 所示，网络推理流程主要包括 3 个模块：选择推理引擎模块，文中采用联合树推理算法；输入推理证据模块，即给出与 GDP 变化关联度较大的节点变量在当前时刻的状态值；求解后验概率模块，即由当前时刻状态值预测得到下一时刻的 GDP 增长率状态值。