

[数据挖掘技术]

数据挖掘技术在水库梯级调度中的研究与应用

文/张 军 张河宜 张 勇 张 鑫
国网神农架供电公司

人工智能 (Artificial Intelligence) 与数据挖掘 (Data Mining) 是当前研究的热门话题。现有水库调度算法对水库历史数据学习能力有限, 特别当维数大时, 模型求解时间长, 数据存储量大, 同时易产生“维数灾”问题。在充分研究现有水库各种调度算法后, 采用数据挖掘与深度学习相结合的方法, 研究梯级水库基于大数据的调度问题, 提出了一种基于人工神经网络改进型算法, 有效解决了现有算法模型求解慢、数据量大的问题, 在梯级水库调度、预测、决策和评估上有较好的应用价值。

数据挖掘和深度学习

1. 数据挖掘 (Data Mining)

传统的数据处理和多机分析处理侧重于查询和统计分析, 数据挖掘是挖掘大量数据之间潜在的模式和关联属性, 而这个模式和关联属性直接影响决策效果。数据挖掘技术可用于辅助决策过程, 也可以帮组专家学者对已有的知识系统进行修正或构建新的知识体系。

数据挖掘关联学科如图1所示。

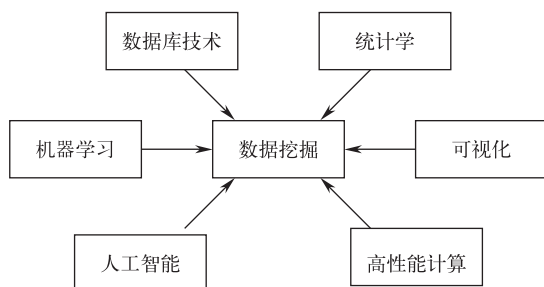


图1 数据挖掘关联学科

2. 数据挖掘的方法和发展趋势

数据挖掘的实现方法和技术手段分为聚合类分析、关联分析、回归分析和时间序列分析等。

(1) 聚合类分析

聚合类分析是把一个物理或抽象的数据集合按照某种特定数据关联分组的过程。根据数据使用场景不同, 并按相似性原则将一组个体归入一种类别, 因此一组数据因使用目的不同, 可存在多种聚合类, 这也是聚合类同传统单一属性分类法的不同之处。聚合类分析是在没有任何指导规则并有效避免人为干扰下而进行的分类聚合, 分类精细度随着计算量增加而增加。目前, 聚合类分析主要应用在模式辨别、数据分析、图像处理及市场预测等领域。通过对海量数据进行聚合类分析, 得出形成特定簇的相应数据分布。不仅可以进一步分析这些簇的特性, 还可以重新处理聚合类分析的结果, 作为输入的其他算法。聚合类分析的算法有很多, 分析不同数据在数据集中的相异度, 形成聚合类分析的标准化度量值是这些算法的主要思路。

(2) 关联分析

数据关联分析是数据挖掘研究的一个重要分支。最早是由Agrawal等人在20世纪90年代分析市场营销时提出, 旨在发现顾客购买模式同商品销

量的内在关联。关联规则挖掘可以通过数据库的逻辑操作或数学统计的方法实现。经分析得到与这些关系相吻合的规律是存在某种内在联系并非偶然突发。一个水库调度关联规律的例证是：分析统计某一流域的历史来水量数据，若发现该流域上游一级水库出现超警戒水位，则该流域下游二级站未来一周内超警戒水位的可能性大于80%。

(3) 回归分析

回归分析是针对连续型输出变量进行预测，通过学习这种关系来预测未知的数据，一般通过大量的运算得出能满足最大范围样本数据的关系。通过自变量和因变量来拟合一条直线，使得训练数据与拟合直线之间的距离最短，最常用的距离是采用平均二次方距离。通过对训练数据的分析可以得到这条直线的斜率和截距，从而可以对于未知数据进行预测，大致可分为逻辑回归和线性回归。

(4) 时间序列分析

时域数据库一般是通过人为规定的时间间隔来采样的序列值或事件组成的集合。时间序列分析是对时序数据库进行分析，从时序中抽取时序内部的规律，对时序的数值大小、周期波动、规律倾向等进行总结分析。时间序列分析主要包括规律分析、相近度搜索、序列模式挖掘和周期模式挖掘等与时间相关的数据分析。

数据挖掘的一般流程为：数据转换→数据挖掘→模式评估→知识呈现。未来将广泛应用于视音频信息挖掘、统计分析数据挖掘、可扩展的数据挖掘、复杂数据类型的挖掘、Web挖掘、隐私保护和数字确权等领域。

数据挖掘过程如图2所示。

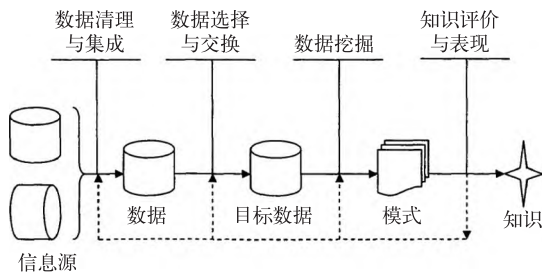


图2 数据挖掘过程

3. 深度学习

人工智能、机器学习、深度学习及数据挖

掘等都是当前的热点，其中深度学习（Deep Learning）与机器学习（Machine Learning）的联系是相当紧密的。机器学习通过分析大量数据进行学习，已经成为人工智能的一个分支，通过自学习算法，发现和挖掘数据潜在的规律，从而对未知的数据进行预测，机器学习图谱如图3所示。深度学习神经网络包含多个隐藏层的数据感知分类器，主要采用各种基于人工神经网络的方法实现机器学习的技术，目前已成功应用于无人驾驶视觉跟踪、自然语言翻译和语义挖掘及图像处理等方面。

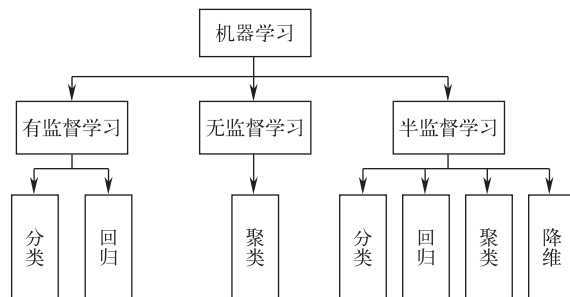


图3 机器学习图谱

目前使用最广的深度学习训练模型包括卷积神经网络（Convolutinal Neural Networks, CNN）、堆栈自动编码网络(Stacked Auto-Encoder, SAE)及递归神经网络等。深度学习网络与单一隐藏层神经网络的区别在于学习的多维度，即数据在模式识别的多步流程中所经过的节点层数。传统机器学习系统的网络为“浅层网络”，主要由1个输入层和1个输出层组成，并在两层之间添加一个隐藏计算层。传统的人工神经网络（Artificial Neural Network）模型一般由输入层、隐藏层及输出层三层组成，如图4所示。而深度学习模型的非线性运算模型由1个输入层、5个隐藏层和1个输出层等三层以上的网络层组成，如图5所示。

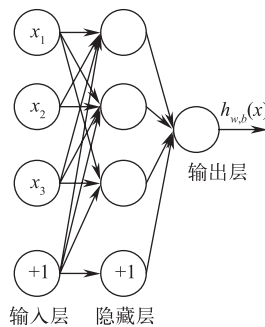


图4 三层神经网络

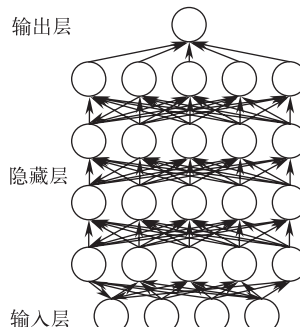


图5 多层神经网络

深度学习神经网络可以自动提取数据属性和特征,因此,使用深度神经网络既可以提取状态特征,又易于训练模型,实现近似的表示和策略自我优化。

一个简单的神经网络工作机制可以抽象为:将初始一元函数的输出作为下一个一元函数的输入。

第一次: $Y_1 = A_1(W_1X + b_1)$, 其中 X 是原始数据的输入, A_1 是激活函数, Y_1 是第一次的输出。

第二次: $Y_2 = A_2(W_2X + b_2)$, 其中 A_2 是激活函数, 参数 W_1 、 W_2 、 b_1 和 b_2 原则上各不相同。

神经网络学习中模型训练评价指标称为损失函数 (Loss Function), 一般采用均方误差和交叉熵误差函数。损失函数是评价神经网络优劣程度的指标, 即表示当前的神经网络输出在多大程度上不拟合监督数据。

均方误差为

$$E = \frac{1}{2} \sum_k (y_k - t_k)^2 \quad (1)$$

神经网络的激活功能使神经网络能够处理非线性问题。常用的激活功能有双曲正切函数

$$\tanh(z) \quad (2)$$

或sigmoid函数

$$g(z) = \frac{1}{1 + e^{-z}} \quad (3)$$

数据挖掘在水库梯级调度中的研究现状

1. 背景

水库的优化调度策略包含控洪防汛、发电生产、农业灌溉、水文预报、生态保护、降雨水量及电网调度等。现有水库调度存在的主要问题有: 一是对水库历史调度的数据学习能力有限; 二是当维数大时, 模型求解时间会很长, 收敛性差, 数据存储量巨大, 产生“维数灾”。

针对预测或者优化问题, 传统的预测方法有趋势外推法、时间序列法和回归分析法等。这些算法的原理简单, 理论相对成熟, 能用标准的技术分析模型中的参数。近年随着人工智能和深度学习的高速发展, 预测算法的精准度也得到了提高, 出现了不少新的预测算法: 误差反向传播 (Back Propagation) 神经网络、人工神经网络预测方法、模糊集预测方法、小波分析预测方法及优选组合预测方法等。

2. 水库中常用的随机优化调度模型

(1) 描述水库的入库径流

1) 水库的入库径流频率曲线。水电站的入库径流是一个周期性的连续性随机过程, 径流值是一个服从皮尔逊 (Pilson) III型分布的随机数值。根据水库历史的入库流量, 求出皮尔逊III型概率密度曲线的三个重要参数, 即入库流量的均值 $\bar{Q}$ 、变差系数 C_{vQ_i} 和偏差系数 C_{sQ_i} 。

① 均值。设某水文变量的观测系列 (样本) 为 $x_1, x_2, \dots, x_n$, 则均值为

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (4)$$

均值表示样本集中样本的平均分布情况, 能显示出该样本集大致水平。

令 $K_i = \frac{x_i}{\bar{x}}$, 则

$$\bar{K} = \frac{1}{n} \sum_{i=1}^n K_i = 1 \quad (5)$$

式中, K_i 为模比系数 (通过适线法确定), 一个均值等于1的由模比系数组成的系列。

② 均方差。均方差的开方表示样本离散的大小程度, 一般选取 $(x_i - \bar{x})^2$ 的平均值表示, 即

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \quad (6)$$

如果系列的均值相等, 则 σ 越大系列越离散, σ 越小系列越集中。

③ 变差系数。两个不同均值的系列不能直接用均方差比较离散程度。在水文计算中, 用均方差与均值的比值衡量系列的相对离散程度, 称为变差系数, 用 C_{vQ_i} 表示, 计算式为

$$C_{vQ_i} = \frac{\sigma}{\bar{x}} = \sqrt{\frac{\sum_{i=1}^n (K_i - 1)^2}{n}} \quad (7)$$

变差系数 C_{vQ_i} 越大, 系列的离散程度越大; 变差系数 C_{vQ_i} 越小, 系列的离散程度越小。

④ 偏差系数。由变差系数 C_{vQ_i} 乘以模比系数得到, 计算式为

$$C_{sQ_i} = K C_{vQ_i} \quad (8)$$

2) 马尔可夫决策过程。学习系统与环境相互作用如图6所示。适用于马尔可夫决策流程的系统主要有三个特点: 状态转移没有效率; 状态的调动存在不确定性; 可充分观察到智能体所处的每一步状

态。整个过程包括马尔可夫性、马尔可夫过程和马尔可夫决策过程。

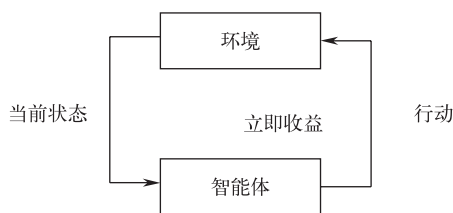


图6 学习系统与环境交互的框图

由马尔可夫性、马尔可夫过程和马尔可夫决策过程可建立马尔可夫决策过程（MDPS）模型，这是一种在强化学习中广泛采用的环境与环境交互模型。

马尔可夫性指下一种状态只和现在的动作状态联系在一起。一旦明确了现在的状态，历史数据就被遗弃。而马尔可夫决策过程与马尔可夫过程不同，马尔可夫决策过程的状态转移矩阵概率包含了动作。

3) 概率转移矩阵。如前所述，每个时段径流都有 N 组数据，每组径流数据对应一个径流状态，状态间的转移存在概率转移矩阵，在求取概率转移矩阵之前要搞清楚径流状态转移过程是独立过程、简单的马尔可夫过程还是复杂的马尔可夫过程，径流相关系数是径流相关性的度量，对径流过程类型的判断是有依据的。以实际观测数据计算相关系数

$$r_{t,t+1} = \frac{\sum_{i=1}^n (Q_t^i - \bar{Q}_t)(Q_{t+1}^i - \bar{Q}_{t+1})}{n\sigma_t\sigma_{t+1}} \quad (9)$$

式中， $r_{t,t+1}$ 为 t 时段转移到 $t+1$ 时段的相关系数； Q_t^i 、 Q_{t+1}^i 分别为 t 时段和 $t+1$ 时段的实际入库径流值； $\bar{Q}_t$ 、 $\bar{Q}_{t+1}$ 分别为 t 时段和 $t+1$ 时段径流值的均值； σ_t 、 σ_{t+1} 分别为 t 时段和 $t+1$ 时段径流值的均方差； n 为统计样本年数。

为实现随机动态规划，需进一步求解转移概率矩阵。 $P(X_{t+1}|X_t)$ 为简单马尔可夫链从 t 时段到 $t+1$ 时段的转移概率，有 $p_{jk}=P(X_{t+1}=j|X_t=k)$ ，第 t 个时段的概率转移矩阵形式为

$$P = \begin{pmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & & \vdots \\ P_{m1} & \cdots & P_{mn} \end{pmatrix} \quad (10)$$

由式（10）即可得到状态转移概率矩阵中的各

元素，从而求得状态转移概率矩阵为

$$p_{jk} = \frac{f_{jk}}{\sum_{k=1}^N f_{jk}} \quad (11)$$

(2) 目标函数

以水库最大年发电量期望值为随机调度模型的目标函数，在一个调度周期内（一般为一年）将各时段发电量的期望值逐次累加，可求解总发电量的期望值，表达式为

$$\begin{aligned} E^* &= \max \sum_{i=1}^T R_i[V_t, Q_{fd}] \\ &= \max \sum_{i=1}^T \sum_{j=1}^N p_{jk} E_i[V_t, Q_{rk,j}, Q_{fd}] \end{aligned} \quad (12)$$

式中， T 为调度周期内的总时段数； R_i 为 t 时段的指标函数； $V_i(i=1 \sim M)$ 为从小到大均匀离散的库容值； Q_{fd} 为发电流量， m^3/s ； $Q_{rk,j}$ 按皮尔逊Ⅲ型分布法计算得到；频率离散为从小到大的水库各时段天然随机来水的拟合值，相应概率为 $p_{jk}(j=1 \sim M)$ ， $k \in (1 \sim N)$ ； E_i 为 t 时段发电量，表达式为

$$\begin{cases} E_i[V_t, Q_{rk,j}, Q_{fd}] = K Q_{fd} H_j \Delta t_i \\ R_i[V_t, Q_{fd}] = \sum_{j=1}^N p_{jk} E_i[V_t, Q_{rk,j}, Q_{fd}] \end{cases} \quad (13)$$

式中， $H_j = Z_{sy} - Z_{xy} = (Z_t - Z_{t+1})/2 - Z_{xy}$ 。

(3) 约束函数

1) 水量平衡方程。公式为

$$V_{t+1} = V_t + (Q_{rk,t} - Q_{fd,t}) \Delta t_i \quad (14)$$

2) 发电流量约束式为

$$Q_{fd,t,\min} \leq Q_{fd,t} \leq Q_{fd,t,\max} \quad (15)$$

3) 水库出力约束式为

$$P_{t,\min} \leq P_t \leq P_{t,\max} \quad (16)$$

④ 水库库容约束式为

$$V_{t,\min} \leq V_t \leq V_{t,\max} \quad (17)$$

⑤ 上下游水位约束式为

$$Z_{t,\min} \leq Z_t \leq Z_{t,\max} \quad (18)$$

式中， V_t 、 V_{t+1} 分别为 t 时段的初、末蓄水量； $Q_{rk,t}$ 为 t 时段的入库流量； $Q_{fd,t}$ 为 t 时段的发电流量； $Q_{fd,t,\min}$ 、 $Q_{fd,t,\max}$ 分别为 t 时段水库的最小、最大发电流量； P_t 为 t 时段水库的实际出力； $P_{t,\min}$ 、 $P_{t,\max}$ 分别为 t 时段水库最小、最大出力约束； V_t 为 t 时段的水库库容； $V_{t,\min}$ 、 $V_{t,\max}$ 分别为 t 时段的最小、最大库容； Z_t 为 t 时段的水库上游水位； $Z_{t,\min}$ 、 $Z_{t,\max}$ 分别为 t 时段

的最低、最高上游水位。

(4) 水库中深度学习和数据挖掘

利用深度学习神经网络对梯级水库历史数据进行学习、训练，验证神经网络和建模，产生知识的流程如图7所示。

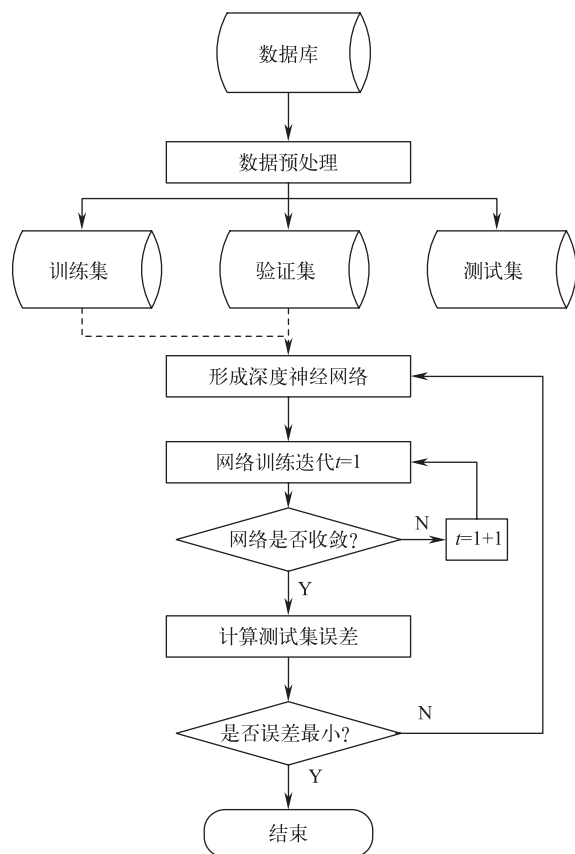


图7 深度学习神经网络模型训练图

结束语

1) 随着水文信息化程度越来越高，水文历史资料积累较多；水库调度领域的热点之一是如何充分利用这些数据建立预报和调度模型。

2) 以聚合类分析、分类预测及关联分析等多种算法，从水文资料角度对水文现象中隐含的规律进行挖掘，在水文预测和水库调度中已成为重要组成技术之一。

3) 常用的误差反向传播神经网络（Back

Propagation）模型是在水文水资源系统分析预报中使用的非线性数据挖掘技术，学习能力极强。然而，运用BP神经网络模型进行水文预报时，需要选择合适的预报因子（特征值），隐藏层和参数的设定是难点，并且目前深度学习中，神经网络训练模型的稳定性、适应性及泛化能力还不够好。在人工智能和大数据时代，对调度规则的智能自动提取的学习和研究还有待继续深入。

4) 随着计算机软件 and 硬件性能的不断提高，处理器由最开始的CPU发展为现在的GPU，甚至是TPU，使得计算机能够处理大数据时代庞杂的任务。而且随着人工智能、机器学习及深度学习等的高速发展，未来在处理大数据、水库优化调度及电力系统优化等方面数据挖掘和深度学习的结合将更加紧密，也将各个学科领域产生巨大影响。EA

(收稿日期：2023-02-17)

参考文献

- [1] 许惠君，李彩林，刘晓安．数据挖掘技术在水库调度中的研究与应用[J]．计算机与数字工程，2006，34(9)：61-63．
- [2] 张弛．数据挖掘技术在水文预报与水库调度中的应用研究[D]．大连：大连理工大学，2006．
- [3] 张弛，周惠成，王本德．决策树技术在水库兴利调度中的应用研究[J]．哈尔滨工业大学学报，2007(8)：1314-1318．
- [4] 徐锦．数据挖掘技术在水资源规划中的应用研究[D]．兰州：兰州大学，2010．
- [5] 李宏伟．基于关联规则的数据挖掘技术在中长期水文预报中的应用[J]．人民珠江，2013(6)：21-25．
- [6] 杨楠，叶迪，林杰，等．基于数据驱动具有自我学习能力的机组组合智能决策方法研究[J]．中国电机工程学报，2019，39(10)：2934-2946．
- [7] 孔晨华，吉龙军，张建军，等．基于人工智能技术电力调度智能代理操作方法研究[J]．电气应用，2021，40(2)：65-70．