基于时延网络的流程供应链时序数据挖掘

彭晨[1],岳东[1],许世范[2]

(1.南京师范大学控制科学与工程系,210042,南京)
(2.中国矿业大学信息与电气工程学院,221008,徐州)

[摘要] 针对连续流程供应链中存在的大量时序数据的知识发现问题,提出了基于信息集成基础上的应用时延神经网络处理时序数据的知识发现方案,解决时序数据膨胀而知识匮乏的问题.在网络数据集成的前提下,采用VB提取数据集市中的综合数据,作为时延神经网络模型的输入序列.利用MATLAB神经网络工具编写知识发现模型来发现时序数据中蕴含的知识,构成知识链并作为供应链优化运行的支撑链.以某选煤厂时序数列分析为例,上述方法在应用中取得较好的效果.

[关键词] 时延神经网络,流程供应链,数据挖掘,知识链
[中图分类号]TP274,TD94, [文献标识码]B, [文章编号]1672-1292(2002)04-0034-05

在煤炭供应链运行过程中存在大量的时间序列,如矿井原煤生产量、选煤厂入洗量、产品销售量、物资消耗量等指标数据.这些指标数据是轻度综合性数据,是在事务性日常生产经营数据记录基础上的综合性统计性数据.从数据集市中可以选择查询维,确定查询面则可取得这些时间序列数据.这些时间序列不仅反映了过去即发事实数据的大小,更重要的是蕴含了生产过程及外部市场变化的内在信息.它能反映生产销售原料消耗以及经营过程中的物流、信息流、资金流等的相互联系和动态产业化过程.供应链决策层在计划制定、战略规划时必须了解过去的情况并分析将来趋势,才能做出理性决策.供应链中的信息集成提供了时间序列综合性数据,而如何分析和处理这些时间序列,正确认识生产经营等的动态特性,对供应链整体优化运营是至关重要的.

1 时序数据特征

时间序列的数据库内某个字段是随着时间而不断变化的,例如股票价格每天的涨跌、浏览网页的次序、产品的销售等.时间序列模式根据数据随时间变化的趋势可以预测将来的值.考虑到时间的特殊性质,像一些周期性的时间定义如星期、月、季节、年等,从不同时间维角度,利用现有数据随时间变化的一系列序列值,才能更好地预测将来.模式按功能可分为两大类:预测型(predictive)和描述型(descriptive).预测型模式可以根据数据项的值精确确定某种结果的模式,而挖掘预测模式所使用的数据都可以明确知道结果;描述型模式则对数据中存在的规则做一种描述,或者根据数据的相似性把数据分组,但不能直接用于预测.

由于外界及内在条件的变化,基于传统回归方法及其参数模型法基础之上的分析方法,存在参数求解困难、假设因素太多的缺点.而神经网络方法具有非线性数据的快速拟合能力,因此在数据采掘过程中,神经网络是聚类的有力工具,在事务数据库的分析和建模方面应用广泛[1].

收稿日期:2002-12-16.
资金项目:国家"八六三"资助项目(863-511-9601-1301)和教育部资助项目([2002]247).
作者简介:彭晨,1973-,在职博士后,南京师范大学信息与控制研究中心讲师,主要从事流程供应链建模及时序数列的知识发现研究.

2 时延序列网络基本原理

神经网络建立在可以自学习的数学模型基础上,它可以对大量复杂的数据进行分析,并可以完成对人脑或计算机来说极为复杂的模式抽取及趋势分析.由于神经元网络是非线性模型,一般认为它比传统的统计学工具更为理想.人工神经网络最大的长处是可以自动地从数据中学习,从而形成知识,因此它具有较大的创新性.

时延序列网络是指具有时间序列输入的神经网络,网络具有时延及反馈存在,网络中神经元的输入—输出不是简单的输入输出映射关系,对神经元网络的训练及输出具有复杂网络的动态特征.时延序列网络的输入输出的关系可用如下的差分方程描述[3].

$$\begin{cases} Y(K) = F[u(k), u(k-1), \cdots, u(k-n)] \\ u(k = (x_{k,1}, x_{k,2}, \cdots, x_{k,j})^T \end{cases}$$

式中 J 为已知序列中因变量个数;F 函数为输入输出间的映射关系;K 为当前要预测的时间标识;n 表示 $Y(k)$ 与 $u(k)$ 的 n 个时刻有关;$u(k)$ 为输入序列,可由 j 个单变量组成.

由于时序关系,这种神经网络由输入延迟形成动态过程.因为输入是动态时间变化序列,所以在有足够输入序列情况下,可直接使用输出序列作为预测序列,由历史序列 $\sum u(k-n)$ 及当前序列 $u(k)$ 构成输入输出序列动态调整预测序列.可知时延序列的特征为:输入序列随时间变化,由前序列构成输入输出序列对网络进行训练.用训练好的网络可对下一时序进行预测.所以本质上时延序列网络是一个动态时序训练网络.精确度与训练样本及要求的误差有关[3,5].

3 基于 DB 的时序神经网络算法结构

针对流程供应链时序数列特征,设计的基于 DB 的神经网络知识发现方法由三阶段构成:综合性数据获取,网络构造训练,规则抽取.

3.1 综合性数据获取

供应链链节中有许多时间维度的时序数列,根据决策需求可从不同时间粒度对时序数据进行规律抽取、数据预测,根据流程行业数据分布特点采取了虚拟集市的数据构建方式[2],对于要处理的综合性数据获取方式,本文采取如下做法:

(1)确定主题域,选择查询面,确定查询维,从供应链数据集市中抽取综合性数据放在多维数组中;

(2)前台 VB 程序一方面通过 ODBC 与 SQL-Server 数据库通讯并从中抽取数据,另一方面通过 VB 的 MATLAB 机制与 MATLAB 建立联系,调用已编好的 *.M 程序,利用 VB 从数据集市中抽取的数据作为神经网络的输入.

3.2 网络构造和训练

根据综合数值序列的维度、时序维数确定网络结构,一是输入结点数,二是输出结点数,三是中间层结点数.对此有两种可能数据输入方式:①前后序列仅作为训练样本值,数值序列本身中含有输入输出对.例如对成本结构进行预测,时间序列中已包含不同成本构成要素值及成本值,不同时间序列仅作为训练样本进行训练,不同序列间没有前后关系.②后一序列作为前一序列的输出,前一序列作为后一序列的输入,时间序列之间存在耦合关系.例如对销售总量进行预测时,根据前几年的综合性数据构造时序输入序列及推后一时间单位的输出序列进行输入输出的示教训练.

接 3.1 节步骤,对训练网络:

(3)VB 调用 MATLAB 中的神经网络工具箱进行神经网络训练,已抽取的综合时序数列作为输入输出值;

(4)编制的 MATLAB 应用程序根据输入输出值序列智能判断采取合适的训练算法.

在第二类输入输出中通过实际仿真发现采用 BP 算法不如采用径向基函数算法快速.因为 BP 采用全局逼近神经网络,网络中的一个或多个数值自适应可调参数在输入的每一点对任何一个输出都有影响,对于每一个输入输出数据对,网络的权值都需要调整,从而导致全局逼近网络学习速度很慢.因学习时间过长,存在较大时滞,这个缺点对实时控制来说常常是不可忽视的.而局部逼近网络对于输入输出数据对,只有少量的权值需要进行调整,从而使局部逼近网络具有学习速度快的优点.根据上述分析在训练中采用 RBF 算法,在准确性、及时性方面取得较好的效果.

3.3 训练结果使用

利用网络训练结果,输入要预测的序列值,将输出序列作为决策等的依据.对 3.2 节的第二种时序输入形式而言,将当前输出序列与上一时刻序列构成新的输入序列作为神经网络模型输入即可得下一个序列的预测值.为保持连贯性,接 3.2 节这一步骤为:

(5)选取输入值(从数据集市中或从已抽取的数组中选取);

(6)利用训练好的网络,输入已知序列值,输入要求的预测值;

(7)显示或输出误差曲线或趋势曲线.

4 时延神经网络挖掘算法步骤

时延反馈型网络又称输出时延反馈网络,是由多层前馈网与输出时延反馈两部分组成,典型结构如图 1 所示.前馈网输入由两部分组成:输出反馈和它们拍延迟线,Z^{-1}表示一步延迟.在文中时延网络采用 RBF 算法调整权值.下面说明时延网络的基本步骤.

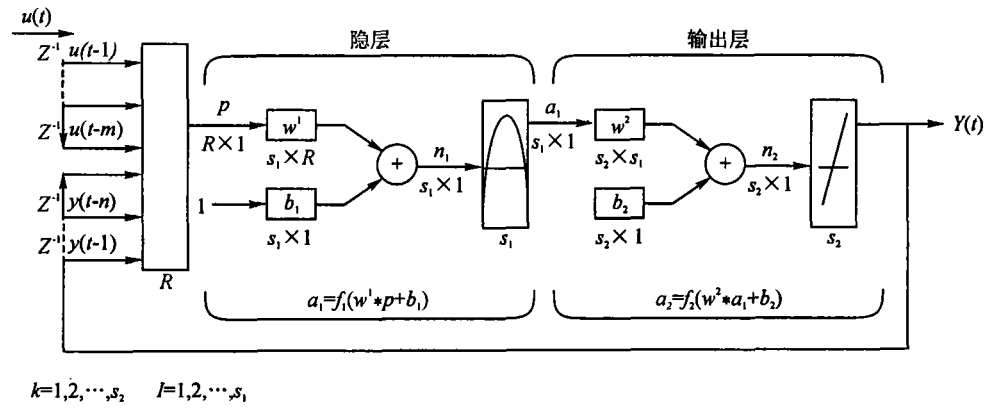

图 1　三层时延神经结构

4.1 在时刻 t,多层前馈网的输入 / 输出样本对

$u(t-1),u(t-2),\cdots,u(t-m),y(t-1,),y(t-2),\cdots,y(t-n)/y_d(t)(m+n=R$ 输入)

4.2 网络输出

$$y(t)=f(y(t-1),y(t-2),\cdots,y(t-n),\ u(t-1),\ u(t-2),\cdots,u(t-m);w)$$

$$=f[x_i]=f[\sum w_{ij}(t)I_i]$$

式中:I_i 为样本输入时,节点 i 的第 j 个输入;$f(\cdot)$ 取可微型 S 作用函数;w_{ij} 为权值系数.

4.3 目标函数

$$J(t)=\frac{1}{2}\|y_d(\tau)-y(\tau)\|^2=\sum_{\tau}E(\tau)=\frac{1}{2}\sum_{\tau}(y_d(\tau)-y(\tau))^2$$

式中:$\tau=1,2,\ldots\ L$. L 为样本长度.

4.4 训练网络

网络训练过程就是调整隐层及输入层的权值及阀值,使训练输出结果与样本的误差小于指定的 ε 或在指定的训练次数不能得出满意的结果后,要求增加训练次数或改变学习率等. 由输出层,据目标函数 J,按梯度下降法反向计算逐层调整权值.取步长为常值,可得到神经元 j 到神经元 i 的联接权 $t+1$ 次调整算式:

$$w_{ij}(t+1) = w_{ij}(t) - \eta \partial J(t)/\partial w_{ij}(t) = w_{ij}(t) + \Delta w_{ij}(t)$$

5 时延神经网络算法运用

运用上述时延神经网络算法,前台利用 VB 程序通过 ODBC 编制数据库调用程序,后台在确定查询主题、查询面、查询维的情况下,利用 SQL-Server 语法编写综合数据抽取存储(store)过程,查询结果放在 VB 定义的多维数组中,之后,利用 VB 与 MATLAB 的通讯机制,编写 MATLAB 调用程序,调用在 MATLAB 下已编好的具有输入输出接口的 *.M 程序. MATLAB 中的 M 程序以 VB 多维查询结果数据作为输入,经过 NNET 中 BP、ART 等进行神经元网络训练.如果训练结果在给定误差范围内,训练结果作为预测的处理函数.若网络训练次数超过给定最大次数且不能达到误差要求,则需调整参数,重新训练网络.挖掘过程结构流程如图 2 所示.

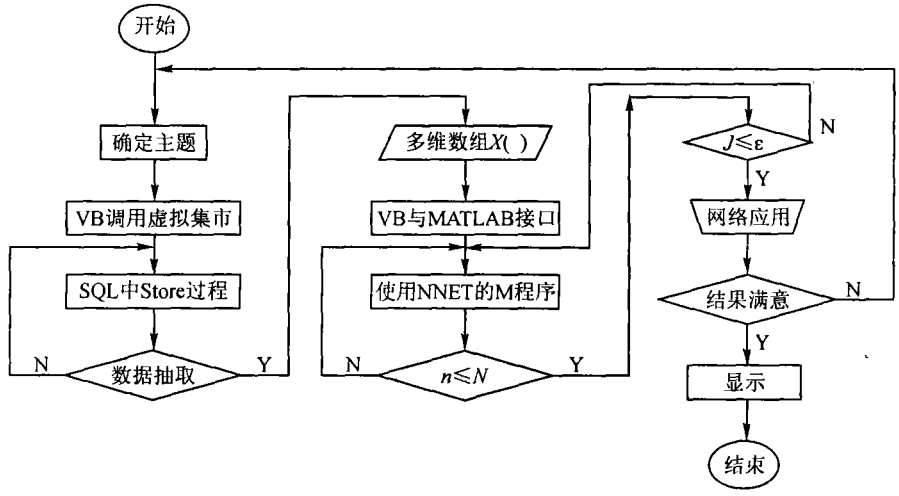

图 2 时延神经网络应用流程

以某选煤厂 1992～2000 年销售记录的精煤量和出口煤量为例,利用上述算法及调用编好的 VB 程序及 MATLAB 神经元网络挖掘程序,选择年销售量作为查询面,经过综合数据的抽取后,对数据进行时延格式输入,挖掘结果如图 3 所示.其中"*"为实际序列,"+"为预测序列.可以看出其中"+"型曲线和"*"型曲线相比,横坐标提前一单位坐标,纵坐标基本相同.上述时延神经网络数据挖掘方法,较好地运用了时序数列的时延性,预测了年销售数量,且与实际销售数量存在较小误差.拟合结果误差如图 3 右图所示.可以看出时延网络的神经元挖掘方法可以较好地解决时序数据的时延问题,为流程供应链中的时序数据的分析提供了一个较好的分析挖掘方法.

6 结论

连续流程供应链中存在的大量时序数据问题,在供应链信息集成的基础上,应用数据挖掘方法采掘海量数据中蕴藏的知识,指导供应链优化运行,并将知识发现机制引入流程供应链的决策环节,使供应链运营过程中的数据流转化为辅助供应链决策过程的知识流.但目前 KDD 研究还主要局限在离散企业如零售 POS 信息、银行欺骗诊断、销售菜篮子分析等领域.将 KDD 方法与流程供应链相结合,将

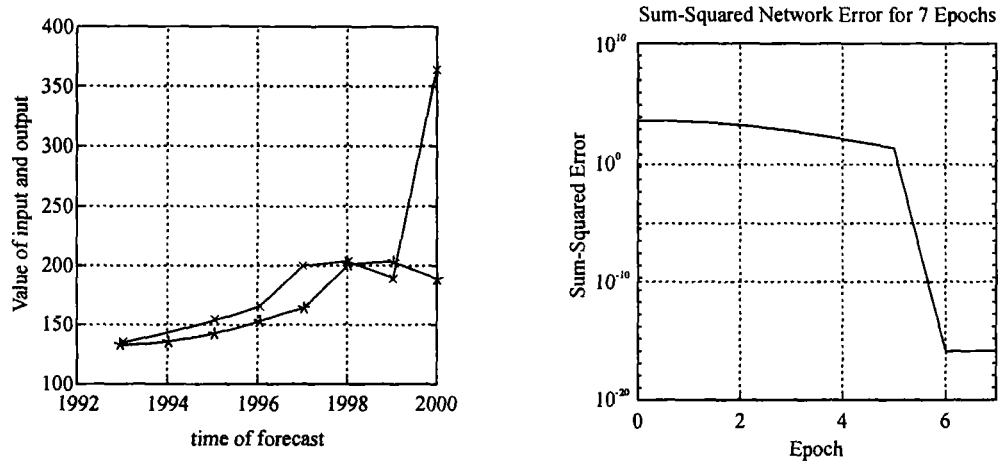

图3　时延网络挖掘结果

KDD发现的知识转化为决策者的知识,形成知识网链且作为供应链运行的支撑链是需进一步研究的问题.

[参考文献]

[1] 高俊波,陆勤,蔡庆生,等.时序数据的模式发现算法研究[J].计算机工程,2002,26(8):15～18.

[2] McDaniel K H, Wallace K G. Real time mine ventilation simulation[J]. Mining Engineering, 1997,49(8):71～75.

[3] 董卫军,高元宝.基于神经网络的生产指标预测和分析方法研究[J].黄金,1999,20(11):14～17.

[4] T Munakata. Knowledge Discovery[J]. Communications of the ACM,1999,42(11):26～29.

[5] J Han, W Gong, Y Yin. Mining Segment-Wise Periodic Patterns in Time-Related Databases[A]. Proc of 1998 Intl Conf on Knowledge Discovery and Data Mining (KDD'98)[C]. New York City,1998.214～218.

The Data Mining of Time-Series of Flow Supply Chain Based on Time-Delayed NNET

Peng Chen[1], Yue Dong[1], Xu Shifan[2]

(1. Department of Control Science and Engineering, Nanjing Normal University, 210042, Nanjing, PRC)

(2. College of Information and Electrical Engineering, CUMT, 221008, Xuzhou, PRC)

Abstract: Aiming at question of knowledge discovery of the large qualities of time-series data in continuous flow supply chain, the project of KDD based on the integrated information is presented and the time-delayed NNET is applied. The question of abundant data and absent knowledge is resolved in time-series data. In the premise of integration of the network data, visual basic is used to extract data as the input sequence of the model of time-delayed NNET. NNET tools of MATLAB are utilized to compile the model of KDD to find the pattern of time-series, forecast the rules of data, form the knowledge chain and use it as the support chain of optimize operation of supply chain. The above methods have been used successfully to some coal preparation.

Key words: time-delayed NNET, flow supply chain, data mining, knowledge chain

[责任编辑:严海琳]