基于数据挖掘的销售预测研究

李雅莉

(宝鸡文理学院 电子电气工程系，陕西 宝鸡 721007)

摘 要： 在研究数据挖掘预测算法时间序列AR模型的基础上，提出了将影响销售预测的因素与时间序列预测结合在一起的BP神经网络销售预测模型，该模型通过数据仓库获取销售历史数据。实例验证表明：BP神经网络销售预测模型比时间序列AR销售预测模型精度高。

关键词： 数据挖掘；销售预测；AR模型；BP神经网络

中图分类号： TP311 **文献标识码：** A

Research of sale forecast based on data mining

LI Ya Li

(Department of Electronic and Electricity Engineering, Baoji University of Arts and Sciences, Baoji 721007,China)

Abstract: On the basis of researching time series AR model for forecasting of data mining algorithm, a model of BP neural network sale forecast combined influence factors of sale forecast with time series forecasting was provided. The model obtained the historic sale data from data warehouse, and examples verified that it is a higher accuracy BP neural network model of sale forecast than time series AR model.

Key words: data mining; sale forecast; AR model; BP neural network

销售预测是企业市场营销管理中最重要的因素之一，也是企业供应链的关键环节。销售预测是在对影响市场供求变化的诸多因素及过去和现在的销售资料进行分析、研究的基础上，运用科学的方法，对未来市场产品的供求发展趋势进行估计和推测。根据销售预测结果，企业可以科学合理地制定采购计划、生产计划、库存计划及营销计划。

随着计算机技术、网络技术、通信技术和Internet技术的发展和各个业务操作流程的自动化，企业产生了数以几十或上百GB的销售历史数据，面对这些海量数据，传统的预测系统越来越不适应新的预测要求，主要表现在：预测涉及海量数据的处理，传统的方法无法满足运行效率、计算性能、准确率及存储空间的要求；预测所需的数据含有大量不完整(缺少属性值或仅包含聚集数据)、含噪声(错误或存在偏离期望的孤立点值)、不一致的内容(来源于多个数据源或编码存在差异)，导致预测陷入混乱[1]。在这种情况下，一个新的研究领域——数据挖掘DM(Data Mining)出现了。

数据挖掘是由计算机自动从已有的大量数据中提取隐含的、未知的、具有潜在应用价值的信息或模式的过程。常见的用于销售预测的数据挖掘算法有：(1)统计分析方法，如时间序列分析、线性回归模型分析、非线性回归模型分析、灰色系统模型分析、马尔可夫分析法等，统计分析法是目前最成熟的数据挖掘技术[1]；(2)仿生物方法，如人工神经网络、遗传算法等，这是数据挖掘算法研究的新方向。

本文主要在数据挖掘预测算法时间序列AR模型分析的基础上，试图建立将影响销售预测的因素与时间序列预测结合在一起的BP神经网络销售预测模型，以提高销售预测的准确性。该模型通过数据仓库获取销售历史数据，并运用实例对这两种算法进行了验证对比。

1 销售预测数据挖掘模型的建立

1.1 时间序列AR模型预测

时间序列分析是根据已知时间序列中的销售数据的变化特征和趋势，预测未来销售值。在时间序列模型中，自回归模型AR是应用最广的一种预测模型。

AR(n)模型的一般形式为：

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_n x_{t-n} + \alpha_t$$

式中：$t=1,2,\cdots N$，时间序列$\{x_t\}$为已知数据，并假定是平稳随机序列；$\phi_1,\phi_2,\cdots,\phi_n$为模型参数；$n$为模型阶数；$\alpha_t$为

测量误差，并假定是白噪声序列。

建立AR模型的一般步骤为[2]：

(1)对时间序列进行平稳性处理。由于AR模型适用于平稳时间序列，因此，建立模型之前要对销售序列数据进行平稳化预处理，通常采用零均值处理。

(2)AR模型阶数的确定。根据准则函数定阶法来确定模型的最优阶数。确定AR模型阶数的准则包括：FPE准则、AIC准则、BIC准则,其函数表达式分别表示为:

FPE准则函数 $FPE(n)=\frac{N+n}{N-n}\sigma_{\alpha}^{2}$

AIC准则函数 $AIC(n)=N\ln\sigma_{\alpha}^{2}+2n$

BIC准则函数 $BIC(n)=N\ln\sigma_{\alpha}^{2}+n\ln N$

式中：N为样本量；σ_{α}^{2}为模型残差均方差。在各自准则函数取得最小值时的阶数为模型的最优阶数，在最优阶数下所建立的模型就是最适用的模型。

(3)AR模型的参数估计。当模型阶数固定时，用普通最小二乘法可对模型参数进行估计：$\phi=(X'X)^{-1}X'Y$

(4)AR模型的预报方程。AR模型的l步预报值为：

$$\begin{cases}\hat{x}_t(l)=\sum_{i=1}^{n}\phi_i x_{t+1-i} & (l=1)\\ \hat{x}_t(l)=\sum_{i=1}^{l-1}\phi_i x_i(l-i)+\sum_{i=1}^{n}\phi_i x_{t+1-i} & (1<l\leqslant n)\\ \hat{x}_t(l)=\sum_{i=1}^{n}\phi_i x_i(l-i) & (l>n)\end{cases}$$

式中：$\hat{x}_t(l)$为在t时刻根据t时刻及之前的数据基于AR模型预测第$t+l$时刻的值。

(5)对预测值进行还原。由于对原始数据进行了平稳性处理，因此，必须对该预测值进行还原，得到实际销售预测值。

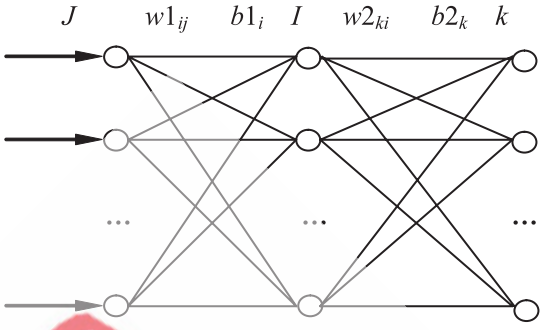
通过AR模型建立销售预测模型，就是根据已知时间序列中的销售数据的变化特征和趋势，预测未来销售值。在历史销售值与预测销售值之间建立线性关系，预测时，输入预测时间前n个销售值，便可根据预测模型计算出预测时间的销售值。

由于产品的需求往往是由许多因素综合决定的，而且影响需求的各种因素之间存在着各种错综复杂的相互作用，具有非线性的特征。根据统计分析方法建立的AR模型无法表达这种相互作用。

1.2 BP网络预测

神经网络作为一种非线性自适应系统，具有通过自学习提取信息内部特征的优点，非常适合解决销售数据中的数据挖掘问题。BP网络是目前应用最为广泛的一种神经网络，具有很强的映射能力，可以实现输入和输出间的任意非线性映射。

BP网络一般由一个输入层、一个或多个隐含层以及一个输出层组成，是通过误差反向传播学习算法来修正网络的权值和阈值。图1是一个典型的三层BP网络模型，其中r为输入层神经元数，$s1$为隐层神经元数，$s2$为输出层神经元数，f_1和f_2为传递函数，$W1_{ij}$和$b1_i$是输入层到隐层的权值和阈值，$W2_{ki}$和$b2_k$是隐层到输出层的权值和阈值。



$j=1,2,\ldots,r$； $i=1,2,\ldots,s1$； $k=1,2,\ldots,s2$

图1 三层BP网络

BP算法由信息的正向传递与误差的反向传播两部分组成。在正向传递过程中，输入信息从输入层经隐含层计算传向输出层，如果在输出层没有得到期望的输出，则计算输出层的误差变化值，然后反向传播，通过网络将误差信号沿原来的连接通路反传回来修改各层神经元的权值和阈值直至达到期望目标。

其算法流程如下[3]：

(1)信息的正向传递

①隐层第i个神经元的输出为：

$$a1_i=f_1\left(\sum_{j=1}^{r}w1_{ij}p_j+b1_i\right)$$

②输出层第k个神经元的输出为：

$$a2_k=f_2\left(\sum_{i=1}^{s1}w2_{ki}a1_i+b2_k\right)$$

③定义的误差函数为：$e=\frac{1}{2}\sum_{k=1}^{s2}(t_k-a2_k)^2$

(2)误差的反向传播

①输出层权值、阈值的调节公式可分别表示为：

$$w2_{ki}(t+1)=w2_{kj}(t)+\eta\sum_{m=1}^{Q}\delta 2_{km}a1_{im};$$

$$b2_k(t+1)=b2_k(t)+\eta\sum_{m=1}^{Q}\delta 2_{km}$$

②隐含层权值、阈值的调节公式分别表示为：

$$w1_{ij}(t+1)=w1_{ij}(t)+\eta\sum_{m=1}^{Q}\delta 1_{im}P_{jm};$$

$$b1_i(t+1)=b1_i(t)+\eta\sum_{m=1}^{Q}\delta 1_{im}$$

式中：P为输入样本；Q为输入样本个数；t_k为网络的期望输出；η为学习步长，$0<\eta<1$；$\delta 1$为隐含层误差传输

项；$\delta 2$为输出层误差传输项。

由于BP算法采用梯度下降法来收敛实际输出与理想输出之间误差，网络有可能陷入局部极小值，采用附加动量与自适应学习速率相结合的方法来改进算法。

通过BP网络建立销售预测模型，就是将影响销售预测的因素和销售历史数据作为输入参数，要预测的销售数据作为输出参数，建立网络模型。预测时，根据网络模型输入预测时间即可计算出要预测时间的销售值。

1.3 预测评估标准

为了比较AR模型和加入影响销售因素后的BP模型的预测能力，采用了平均绝对百分比误差*MAPE*来评估预测的精确性。

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-f_i}{y_i}\times 100\%\right|$$

式中：y_i为第i期实际值；f_i为第i期预测值。

2 实例分析

为了验证AR模型和加入影响销售因素后的BP模型的运行效果，将其应用于某一零售企业决策支持系统的销售量预测中，对决策支持系统中数据仓库的销售数据进行挖掘处理。

2.1 问题分析

对决策支持系统中数据仓库的销售数据进行挖掘处理，以预测商品的月销售量作为时间序列，建立AR模型；以销售时间、商品的月平均价格、购买此类商品的顾客平均收入作为影响因素，与销售时间序列一起作为BP网络的输入端，建立BP网络模型。当用户输入预测时间和商品类型时，系统就能通过这些模型得到销售量。

2.2 数据准备

系统按照预测要求从数据仓库中提取销售预测所需的数据，AR模型预测要求序列是平稳序列，BP网络输入样本如果属于不同的量纲，为了避免量级上的差别影响网络的识别精度，在训练前对数据进行归一化处理。对这些数据按照算法要求进行处理后存放在临时数据库中。

2.3 运行效果

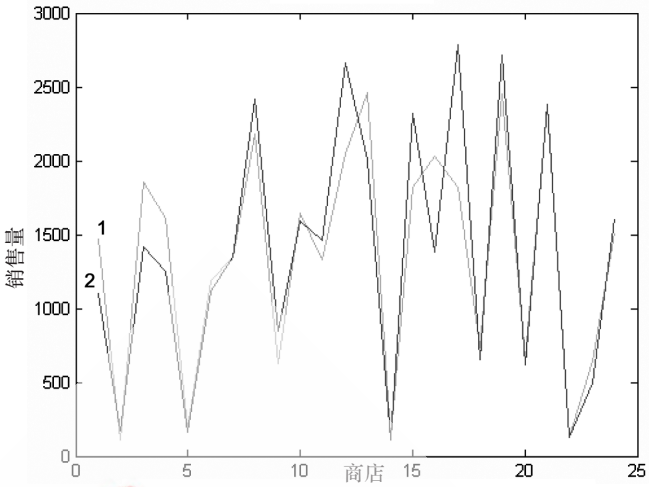
对这两种模型分别在Matlab下编程，数据选取企业1998年食品类商品按月汇总后在不同地区的销售信息，取前10个月数据作为训练样本，第11个月数据作为验证样本。

一个评价预测精度的参考标准认为，平均绝对百分比误差在20％～50％之间的为可行预测，低于20％的为良好预测[4]。如图2和图3所示的两种模型的预测结果，和表1所列的预测精度比结果看，BP模型预测精度比AR模型较高，但两个模型预测都是良好预测。

通过以上对两种模型的实例分析，得出以下结论：

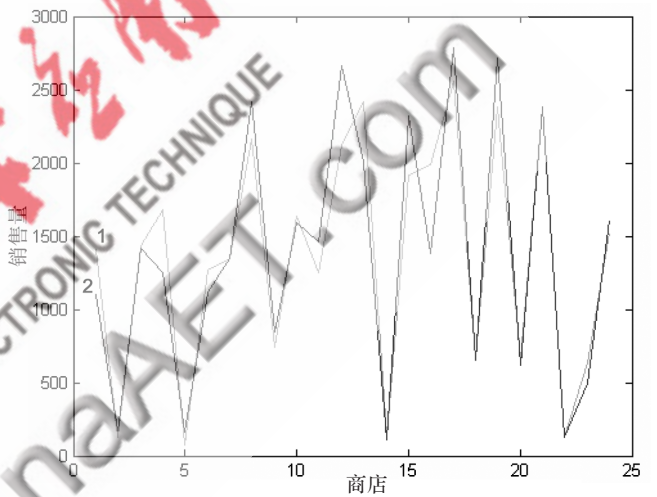
(1)由于销售预测是一个非线性问题，所以BP网络预测精度较AR模型高。而AR模型考虑到时间序列的随机特性和统计特性，也能达到令人满意的预测精度。

(2)销售量除了与销售的时间序列有关以外，还受许多综合因素的影响，像商品的质量、价格、销售的时间、地区、顾客的购买力、气候、促销方式、市场竞争力等，由于BP模型可以是多输入的网络结构，因而可以方便地利用它来考虑其他因素对销售量的影响，在数据完备的情况下，建立起销售量的BP网络预测模型，更能全面地反映出销售量与其他因素的关系。



灰色1—预测销售量，黑色2—实际销售量

图2 AR模型预测结果



灰色1—预测销售量，黑色2—实际销售量

图3 BP网络预测结果

表1 预测精度对比

评估标准	AR模型	BP模型
平均绝对百分比误差/%	18.3	16.27

参考文献

[1] 刘玲梅，孔志周.数据挖掘在销售预测中的应用[J].商业时代(理论版)，2004，23(17)：8-9.

[2] 陈玉祥,张汉亚. 预测技术与应用[M]. 北京：机械工业出版社, 1985.

[3] 陈祥光,裴旭东. 人工神经网络技术及应用[M]. 北京：中国电力出版社，2003.

[4] 王玉荣. 商务预测方法[M]. 北京：对外经济贸易大学出版社, 2003.

（收稿日期：2009-01-11）