

基于PCA主成分分析和BP神经网络企业库存预测的研究

腾杨刚, 陈劲杰, 葛桂林

(上海理工大学机械工程学院, 上海 200093)

摘要:近年来,人力资源和物流及仓储成本的不断攀升,导致零件制造成本不断上升,而准确的库存预测有助于企业据此调整生产计划,降低制造成本,有助于实现企业利润最大化。本文通过PCA主成分分析方法确定影响企业库存的因素,编写python代码分析出影响库存的主要因素,包括订单、当月销量等因素,提出JIT即零库存作为企业库存管理的发展方向。随后选取影响库存的因素,分析并计算相关网络参数,建立BP神经网络,用MATLAB编写预测算法,预测9月的库存,确认预测的合理性,验证了算法的有效性。

关键词: PCA主成分分析; BP神经网络; 库存预测; Python; 机器学习

中图分类号: TP311 **文献标识码:** A

Research on Inventory Prediction Based on Principal Component Analysis and BP Neural Network

TENG Yanggang, CHEN Jinjie, GE Guilin

(School of Machine Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: In recent years, the ever-increasing costs of human resources, logistics, and warehousing have led to the rise of component manufacturing costs. Accurate inventory forecasts can help enterprises adjust production plans accordingly, reduce manufacturing costs, and help maximize profits. This paper adopts the PCA method to determine the factors that affect the inventory, writes python codes to analyze the main factors affecting the inventory, including orders, monthly sales and other factors, puts forward the JIT zero inventory as the development direction of enterprise inventory management. Then the factors that affect the inventory are selected, analyzing and calculating relevant network parameters, establishing BP neural network, using MATLAB to write the prediction algorithm, predicting the inventory in September, confirming the rationality of the prediction, and verifying the effectiveness of the algorithm.

Keywords: PCA (Principal Component Analysis); BP neural network; inventory prediction; Python; machine learning

1 引言(Introduction)

近年来,制造业的人力和土地成本不断攀升,给企业利润造成较大的压力,而根据企业现有的库存所受的制约状况,分析企业库存的主要影响因素,合理的预测出企业某一阶段的库存,可以让企业提前做好生产储备计划,不仅能够为企业生产减少库存面积,还能为企业节省大量的流动资金、人力资源等,为企业增大利润空间。随着神经网络^[1]和相关数学分析方法的不断完善,以及相关语言算法的不断发展,这为预测企业的某一阶段库存提供了可能。

2 PCA主成分分析(Principal component analysis)

在实际的生产实践中,影响库存的因素极其复杂且不确定,甚至是天气的异常变化都会对库存产生影响,但是在大多数情况下,库存的主要影响因素相对固定,表1中选取主要

影响因素并对其进行PCA(Principal Component Analysis)降维^[2]。

表1 库存相关影响因素

Tab.1 Inventory related factors

月份	库存	剩余 订单	当月 销量	在制 零件	采购 周期	到货 周期	供应商	临采	退货
1	2348	2727	2034	698	2	3	5	786	38
2	2041	2530	1961	603	1.5	4	3	814	40
3	2142	2512	2004	649	1	3.5	5	831	21
4	2076	2552	1998	687	2	3	4	939	29
5	2741	2213	2061	624	2	2.5	4	570	23
6	2456	2588	2098	684	1.5	4	5	850	18
7	2875	2829	1921	706	2.5	3	4	534	0
8	2668	2142	1835	663	2	2.5	4	938	8
9		2406	2131	682	1.5	3	4	734	7

在表1中，列出由8个指标组成的8维随机向量，数据矩阵为

$$X^0 = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{18} \\ x_{21} & x_{22} & \cdots & x_{28} \\ \vdots & \vdots & \ddots & \vdots \\ x_{81} & x_{82} & \cdots & x_{88} \end{bmatrix} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$$

故综合指标为

$$\begin{cases} y_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{81}x_8 \\ y_2 = a_{12}x_1 + a_{22}x_2 + \cdots + a_{82}x_8 \\ \vdots \\ y_8 = a_{18}x_1 + a_{28}x_2 + \cdots + a_{88}x_8 \end{cases}$$

系数矩阵定义如下：

$$R_{ij} = \frac{\sum_{m=1}^n (x_{mj} - \bar{x}_i)(x_{mj} - \bar{x}_j)}{\sqrt{\sum_{m=1}^n (x_{mj} - \bar{x}_i)^2 \sum_{m=1}^n (x_{mj} - \bar{x}_j)^2}}$$

对于 $k=1, \dots, 8$ ，有

$$\sum_{j=1}^p a_{jk}^2 = 1$$

变量 y_1 是线性组合中方差最大的，其他依次递减，协方差 $D(X)$ 的特征值依次递减，则第 k 个主成分的贡献率为

$$\frac{\lambda_k}{\sum_{j=1}^8 \lambda_j} \geq \beta \quad (\beta \geq 80\%)$$

如果库存影响因素的前 n 个主成分累计贡献率越大，则可以认为前 n 个主成分可以表征数据，当累计贡献率超过80%时，即可认为所选的主成分可以代表整组数据，下面利用上述计算方法，将认为影响库存的八个因素进行主成分分析，比较这八个因素可否代表库存的主要影响因素。

调用Numpy、matplotlib.pyplot、PCA、pandas、python经典库计算数组的特征值及协方差矩阵等。

```
def pca(dataMat,percentage=0.8): #百分比
    meanVals=mean(dataMat,axis=0) #平均值
    meanRemoved=dataMat-meanVals #移除均值
    covMat=cov(meanRemoved,rowvar=0) #求协方差
    eigVals,eigVects=linalg.eig(mat(covMat))
    k=eigValPct(eigVals,percentage)
    eigValInd=argsort(eigVals) #排序
    eigValInd=eigValInd[:-(k+1):-1] #主成分
    redEigVects=eigVects[:,eigValInd]
    lowDDDataMat=meanRemoved*redEigVects #降维
    lowDDDataMat reconMat=(lowDDDataMat*redEigVects.
T)+meanVals
# 重构数据
```

在CMD运行python程序，计算库存影响因素相关性矩阵(表2)、总方差(表3)并绘出碎石图(图1)。

表2 库存影响因素的相关性矩阵

Tab.2 Correlation matrix of inventory influencing factors

相关性矩阵A								
相关性	订单	当月销量	在制零件	采购周期	到货周期	供应商数	临采	退货
订单	1.000	0.233	0.573	0.148	0.443	0.237	-0.241	0.073
当月销量	0.233	1.000	-0.009	-0.335	0.374	0.478	-0.152	0.443
在制零件	0.573	-0.009	1.000	0.484	-0.186	0.549	0.008	-0.399
采购周期	0.148	-0.335	0.484	1.000	-0.640	-0.276	-0.453	-0.381
到货周期	0.443	0.374	-0.186	-0.640	1.000	0.043	0.238	0.346
供应商数	0.237	0.478	0.549	-0.276	0.043	1.000	0.116	-0.106
临采	-0.241	-0.152	0.008	-0.453	0.238	0.116	1.000	0.304
退货	0.073	0.443	-0.399	-0.381	0.346	-0.106	0.304	1.000

表3 总方差解释

Tab.3 Interpretation of total variance

总方差解释			
成分	总计	方差百分比	累积 %
1	2.616	32.701	32.701
2	2.135	26.684	59.385
3	1.251	15.631	75.017
4	1.001	12.513	87.529
5	0.738	9.22	96.749
6	0.226	2.822	99.571
7	0.034	0.429	100
8	-3.50E-17	-4.37E-16	100

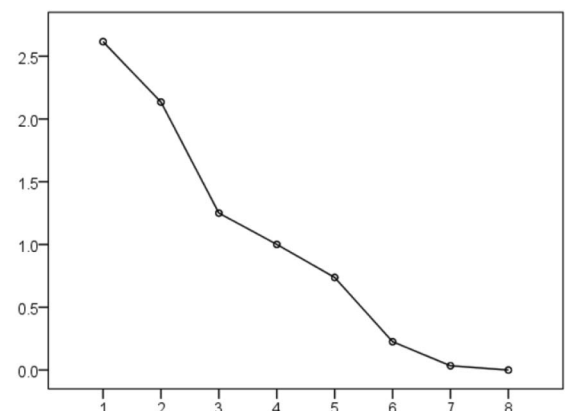


图1 碎石图

Fig.1 Gravel map

在表3中可以算出, 订单、当月销量、在制零件、采购周期四个因素累计贡献率为87.529%, 大于80%, 再结合图1碎石图, 前四个成分下降最为明显, 可以认为这四个因素是库存影响的主要因素, 提取四个主成分, 将主成分系数矩阵除以特征根得主成分函数表达式:

$$F_1 = -0.042 * Z_{X_1} + 0.377 * Z_{X_2} - 0.530 * Z_{X_3} + 0.862 * Z_{X_4} + 0.749 * Z_{X_5} + 0.084 * Z_{X_6} + 0.499 * Z_{X_7} + 0.65 * Z_{X_8};$$

$$F_2 = 0.759 * Z_{X_1} + 0.488 * Z_{X_2} + 0.746 * Z_{X_3} - 0.047 * Z_{X_4} + 0.324 * Z_{X_5} + 0.737 * Z_{X_6} - 0.095 * Z_{X_7} - 0.082 * Z_{X_8};$$

$$F_3 = 0.561 * Z_{X_1} - 0.019 * Z_{X_2} - 0.216 * Z_{X_3} + 0.294 * Z_{X_4} + 0.352 * Z_{X_5} - 0.514 * Z_{X_6} - 0.525 * Z_{X_7} + 0.321 * Z_{X_8};$$

$$F_4 = 0.291 * Z_{X_1} - 0.680 * Z_{X_2} + 0.242 * Z_{X_3} + 0.210 * Z_{X_4} + 0.152 * Z_{X_5} - 0.041 * Z_{X_6} + 0.635 * Z_{X_7} + 0.229 * Z_{X_8}$$

由综合主成分公式 $F = 0.3270 * F_1 + 0.2668 * F_2 + 0.1563 * F_3 + 0.1253 * F_4$, 前四个成分得分依次为: 3.0213、2.5447、1.5178、1.4115, 该得分证明认为所列库存影响因素可以代表库存的主要影响因素。

3 数据整理(Data organization)

对库存数据进行整理, 得训练集P、测试集T两个矩阵如下:

P=[2727 2530 2512 2552 2213 2588 2829 2142; 2034 1961 2004 1998 2061 2098 1921 1835; 698 603 649 687 624 684 706 663; 2 1.5 1 2 2 1.5 2.5 2; 3 4 3.5 3 2.5 4 3 2.5; 5 3 5 4 4 5 4 4; 786 814 831 939 570 850 534 938; 38 40 21 29 23 18 0 8];

T=[2348 2041 2142 2076 2741 2456 2875 2668]

4 BP神经网络(Back-Propation neural networks)

BP(Back Propagation)神经网络^[1,2]由输入(input layer)、隐含(hidden layer)和输出(output layer)三层构成, 图2是三层神经网络结构模型, 库存预测模型选用S型传递函数

$f(x) = \frac{1}{1+e^{-x}}$, 通过输出层误差调整函数: $E = \frac{1}{2} \sum_i (t_i + O_i)^2$, 隐含层误差函数: $E = \frac{1}{2} \sum_i [d_k - f(\sum_j^m w_{jk} y_j)]^2$, 输入层误差函数: $E = \frac{1}{2} \sum_i d_k - f[\sum_j^m w_{jk} f(\sum_j^n v_{ij} x_i)]^2$, 通过偏导求出权重:

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}} \quad (j = 0, 1, 2, \dots, m; i = 1, 2, \dots, l)$$

$$\Delta v_{ij} = -\eta \frac{\partial E}{\partial v_{ij}} \quad (i = 0, 1, 2, \dots, n; j = 1, 2, \dots, m)$$

不断调节权值和阈值使E至最小^[3]。

5 网络结构设计(Networks design)

(1)输入输出层的设计

订单、当月销量、在制零件、采购周期、到货周期、供

应商数、临采、退货作为输入, 以库存作为输出, 故输入层节点为8, 输出层的节点为1。

(2)隐含层设计

首先根据经验公式来确定神经元的个数^[3,4], 公式为 $l = \sqrt{n+m} + a$, 由8个库存影响因素得输入层的神经元个数为8, 输出层的神经元个数为1, $a \in [1, \dots, 10]$, 即 $l = \sqrt{8+1} + [1, 10]$, 所以, $l \in [4, 13]$, 故库存预测模型选择6个隐含层的神经元。

网络结构示意图如图2和图3所示。

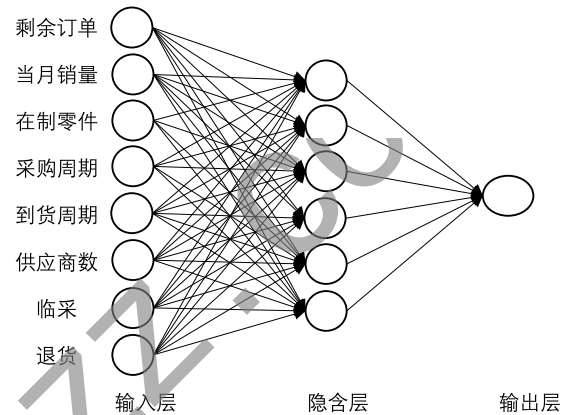


图2 神经网络结构图

Fig.2 Neural network structure

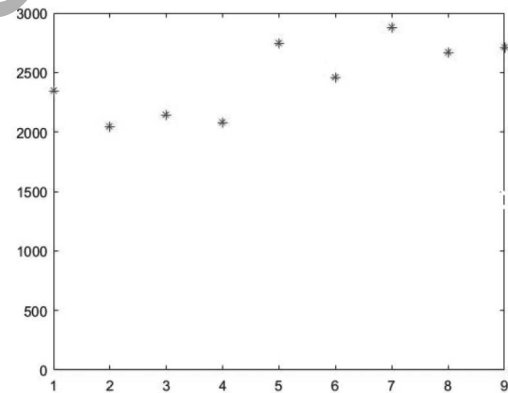


图3 程序运行结果

Fig.3 Program operation result

6 激励函数的选取(Select incentive functions)

一般情况下, 激励函数选用Sigmoid。隐含层的激励函数选择tansig。而因网络的输出在 $[-1, 1]$, 本库存预测模型输出层的激励函数选取S型对数函数logsig^[5]。

7 模型的实现(Model achievement)

构建库存预测模型的步骤如下:

将影响库存的数据归一化后输入构建的训练网络, 隐含和输出层分别为tansig和logsig激励函数, 隐含层神经元个数为6。设定相关网络参数: 网络的epochs为5000次, 期望的误差 (下转第13页)