基于MLP神经网络模型

对棉花期价运行方向模拟预测的研究

■ 姚禹

多层感知器作为神经网络的一种，属于前馈人工神经网络模型，能够处理非线性可分离的问题，其表现形式是通过输入层，把多个数据集利用隐藏层的数据转换，将结果映射到单一的输出数据集中。本文通过选取棉花期货盘面的相关数据，利用多层感知器神经网络模型，对郑棉主力合约价格运行方向进行预测研究，整体拟合效果较好。

研究背景及概念

神经网络是由大量简单的处理单元广泛地互相连接而形成的复杂网络系统。神经网络作为一种高度复杂的非线性动力学习系统，具有大规模并行、分布式存储和处理、自组织、自适应和自学能力，特别适合需要同时考虑诸多因素及模糊的信息处理问题。

对期货绝对价格运动路径的研究，需要考虑各种因子的影响，难以准确模拟预测期货绝对价格的运行水平。不过，若进行降维考虑，利用现有影响期价运行的因子，去模拟预测期价的相对运行方向则相对容易。

多层感知器（MLP）作为神经网络的一种，属于前馈人工神经网络模型，能够处理非线性可分离的问题，其表现形式是通过输入层，把多个数据集利用隐藏层的数据转换，将结果映射到单一的输出数据集中。本文基于MLP模型，对棉花期价运行方向进行模拟预测研究。

模型搭建及分析

本文以郑棉主力合约价格运行方向为研究预测对象，样本选取时间为2012年1月2日—2022年9月30日，共2613个交易日数据。因研究对象属于微观短期现象，故假设价格变动已充分反映基本面因素。为便于量化，本文将影响棉花期价的因素分为期货市场基本行情指标与期货市场技术指标两类，具体指标构成如表1所示。

项目	指标名称
期货市场基本行情指标	开盘价
	最高价
	最低价
	收盘价
	结算价
	成交量
	成交额
	持仓量
期货市场技术指标	指数平滑异同平均（MACD）
	十日简单移动平均（MA10）
	心理指标（PSY）
	相对强弱指标（RSI）
	变动速率（ROC）
	威廉指标（LWR）
	人气意愿指标（ARBR）
	乖离率（BIAS）

表1为影响棉花期价的因素指标

所选取的指标既有绝对数指标又有相对数指标，但绝对数指标对预测价格的影响程度明显高于相对数指标，若不对原始数据进行处理，则会对价格预测的准确性产生影响。为方便计算，在构建MLP神经模型之前，需要对原始数据进行归一化处理，将各原始指标间的差异尽可能缩小。这不仅能在进行梯度计算时，避免绝对数值过度拟合的问题，而且在MLP神经模拟中，归一化处理后的数据收敛速度相较原始数据更快，同时能提升拟合的精确度。

	最小值	最大值	平均数	标准偏差
开盘价	10000.00	22810.00	15865.33	3011.39
最高价	10105.00	22960.00	15980.22	3012.17
最低价	9890.00	22345.00	15746.82	3005.70
收盘价	9990.00	22855.00	15861.53	3010.84
结算价	9995.00	22535.00	15865.1072	3007.54125
成交量	1814.00	2864938.00	275281.93	267531.13
成交额	17843.15	25513810.48	2088554.30	2059820.94
持仓量	41538.00	1166786.00	339646.36	156146.80
MACD	−1290.03	1175.51	−16.19	246.39
MA	10195.00	22099.50	15873.95	3005.35
PSY	8.33	88.89	49.20	12.59
RSI	5.71	93.37	49.10	14.39
ROC	−19.12	33.23	−0.08	4.57
LWR	5.03	97.45	51.38	22.79
ARBR	17.41	672.22	111.24	55.96
BIAS	−14.29	15.19	−0.06	2.15

表2为原始指标描述统计特征

本文利用标准差标准化法对数据进行标准化处理，计算公式为X=（X−变量的平均值）/变量的标准差。经过标准差标准化处理的数据符合正态分布特征。本文对原始数据按照上述公式进行调整后的描述性统计特征表3所示。根据表3中的数据可以明显看出，经调整后的指标差异明显缩小，避免因数值过大而造成对价格预测过度拟合。

	最小值	最大值	平均数	标准偏差
Z（开盘价）	−1.95	2.31	0.00	1.00
Z（最高价）	−1.95	2.32	0.00	1.00
Z（最低价）	−1.95	2.20	0.00	1.00
Z（收盘价）	−1.95	2.32	0.00	1.00
Z（成交量）	−1.02	9.68	0.00	1.00
Z（成交额）	−1.01	11.37	0.00	1.00
Z（持仓量）	−1.91	5.30	0.00	1.00
Z（MACD）	−5.17	4.84	0.00	1.00
Z（MA）	−1.89	2.07	0.00	1.00
Z（PSY）	−3.25	3.15	0.00	1.00
Z（RSI）	−3.01	3.08	0.00	1.00
Z（ROC）	−4.16	7.29	0.00	1.00
Z（LWR）	−2.03	2.02	0.00	1.00
Z（ARBR）	−1.68	10.02	0.00	1.00
Z（BIAS）	−6.6	7.09	0.00	1.00

表3为经调整后的指标描述统计特征

接下来，为分析所选指标与郑棉主力合约结算价之间的联动性特征，笔者对各指标与结算价进行相关性分析，所得出的结果如表4所示。

由表4中的相关系数结果可以看出，开盘价、最高价、最低价、收盘价与MA5个指标及结算价的相关性较强，可以把这5个指标直接选入价格预测模型中。对于剩余指标，由于其相关系数较低，不适宜直接纳入预测模型体系，需要对剩余指标进行进一步处理。

指标	相关系数
开盘价	0.999
最高价	1.000
最低价	1.000
收盘价	1.000
成交量	−0.237
成交额	−0.075
持仓量	−0.395
MACD	0.185
MA	0.994
PSY	0.119
RSI	0.139
ROC	0.094
LWR	−0.114
ARBR	0.103
BIAS	0.064

表4为各指标与郑棉结算价的相关系数

对于剩余10组指标，本文选用主成分分析法对数据进行降维处理，保证本文预测指标间相互独立的同时，消除多重共线性，剔除噪声对模型的干扰。为检验样本数据是否适用主成分分析法，需先对样本数据进行KMO球形检验，检验结果如表5所示。

KMO取样适切性量数		0.767
Bartlett球形检验	近似卡方	23631.775
	自由度	45
	显著性	0.000

表5为KMO球形检验

根据表5的数据可知，KMO统计值为0.767，大于0.5，表明指标间的相关程度无太大差异，适合进行主成分分析。主成分分析法的成分提取原则有两种：第一种是提取主成分对应的特征值大于1的前m个特征值；第二种是前m个主成分累积贡献率大于85%。本文为保证所选主成分最大程度反映指标信息，故以第二种提取原则为准。由表6可知，前5个主成分的累积贡献率为89.951%，对10个指标的整体解释能力较强。

成分	初始特征值			提取载荷平方和		
	总计	方差百分比	累积%	总计	方差百分比	累积%
1	4.789	47.889	47.889	4.789	47.889	47.889
2	2.241	22.410	70.300	2.241	22.410	70.300
3	0.801	8.009	78.308	0.801	8.009	78.308
4	0.629	6.294	84.602	0.629	6.294	84.602
5	0.535	5.349	89.951	0.535	5.349	89.951
6	0.469	4.690	94.640			
7	0.271	2.712	97.353			
8	0.158	1.576	98.929			
9	0.083	0.831	99.759			
10	0.024	0.241	100.000			

表6为总方差解释

进一步对所划分的5组主成分进行分析。由表7可知，在第一主成分中，RSI、ROC、BIAS数额较大，说明第一主成分主要反映这三个指标信息；在第二主成分中，成交量、成交额、持仓量数额较大，说明第二主成分主要反映了这三个指标信息。以此类推，第三主成分主要反映MACD与PSY这两组指标；第四主成分主要反映ARBR指标；第五主成分主要反映LWR指标信息。

	成分				
	1	2	3	4	5
Zscore(成交量)	0.438	0.850	−0.095	0.036	0.241
Zscore(成交额)	0.459	0.835	−0.054	0.010	0.272
Zscore(持仓量)	0.267	0.728	−0.101	0.069	−0.611
Zscore(MACD)	0.737	0.007	0.489	−0.125	0.007
Zscore(PSY)	0.664	−0.106	0.474	0.457	−0.029
Zscore(RSI)	0.920	−0.203	−0.045	−0.077	−0.069
Zscore(ROC)	0.897	−0.126	0.092	−0.189	0.054
Zscore(LWR)	−0.819	0.278	0.175	0.034	0.113
Zscore(ARBR)	0.583	−0.319	−0.452	0.496	0.092
Zscore(BIAS)	0.818	−0.210	−0.265	−0.331	0.000

表7为成分矩阵

A1	A2	A3	A4	A5
0.2001	0.5678	−0.1061	0.0454	0.3295
0.2097	0.5578	−0.0603	0.0126	0.3719
0.122	0.4863	−0.1129	0.087	−0.8353
0.3368	0.0047	0.5464	−0.1576	0.0096
0.3034	−0.0708	0.5296	0.5762	−0.0396
0.4204	−0.1356	−0.0503	−0.0971	−0.0943
0.4099	−0.0842	0.1028	−0.2383	0.0738
−0.3742	0.1857	0.1955	0.0429	0.1545
0.2664	−0.2131	−0.505	0.6254	0.1258
0.3738	−0.1403	−0.2961	−0.4174	0

表8为主成分指标对应系数

为简便表示，本文以A1、A2、A3、A4、A5分别表示第一、第二、第三、第四、第五主成分特征向量，X_1表示成交量，X_2表示成交额，X_3表示持仓量，X_4表示MACD，X_5表示PSY，X_6表示RSI，X_7表示ROC，X_8表示WR，X_9表示ARBR，X_{10}表示BIAS。笔者根据成分矩阵的相关数据以及主成分对应的特征值，能够得到5个主成分中每个指标所对应的系数。主成分指标对应系数如表8所示。

根据表8的结果，将得到的特征向量与标准化的数据相乘，就可以得出主成分的线性组合。本文用Comp.1、Comp.2、Comp.3、Comp.4、Comp.5分别表示5个主成分的线性组合，表达式如下所示：

$$Comp.1=0.2001X_1+0.2097X_2+0.122X_3+0.3368X_4-0.3034X_5+0.4204X_6+0.4099X_7-0.3742X_8+0.2664X_9+0.3738X_{10}$$

$$Comp.2=0.5678X_1+0.5578X_2+0.4863X_3+0.0047X_4-0.0708X_5-0.1356X_6-0.0842X_7+0.1857X_8-0.2131X_9-0.1403X_{10}$$

$$Comp.3=-0.1061X_1-0.0603X_2-0.1129X_3+0.5464X_4+0.5296X_5-0.0503X_6+0.1028X_7+0.1955X_8-0.505X_9-0.2961X_{10}$$

$$Comp.4=0.0454X_1+0.0126X_2+0.087X_3-0.1576X_4+0.5762X_5-0.0971X_6-0.2383X_7+0.0429X_8+0.6254X_9-0.4174X_{10}$$

$$Comp.5=0.3295X_1+0.3719X_2-0.8353X_3+0.0096X_4-0.0396X_5-0.0943X_6+0.0738X_7+0.1545X_8+0.1258X_9+0X_{10}$$

通过预测指标的调整选取，本文共选取开盘价、最高价、最低价、收盘价、结算价、MA、Comp.1、Comp.2、Comp.3、Comp.4、Comp.5共11个指标构成预测因子代入价格预测模型进行研究。

本文通过MLP多层感知机神经网络模型，对棉花期价运行方向进行预测，对棉价上行用1表示，对棉价下行用2表示。本文设置的训练集与检验集如表9所示，按照7∶3的比例对样本进行训练集与检验集划分。其中，训练集共1852组数据，检验集共761组数据，对总共2613组有效数据进行训练。通过学习拟合模型，检验集761组数据用于对该模型的测试，验证MLP模型的结果。

		N	百分比
样本	训练	1852	70.9%
	检验	761	29.1%
有效		2613	100.0%
已排除		0	
总计		2613	

表9为个案处理摘要

在神经网络隐藏层的激活函数选取中，本文采用双曲正切函数，即$\tanh(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$。该函数的优点在于平滑性与可求导性，同时解决了Sigmoid函数不以零为中心的输出问题。在日常操作中，一般选择多层网络层数来尽可能模拟样本运行路径，而层数并不与预测效果成正比关系。随着网络层数层层递进，模型的网络结构复杂程度会加速上涨，从而容易衰减网络收敛速度，甚至加大输出结果的误差，最终产生过拟合现象，弱化神经网络模型的泛化能力。

输入层	协变量	1	开盘价
		2	最高价
		3	最低价
		4	收盘价
		5	MA
		6	Comp.1
		7	Comp.2
		8	Comp.3
		9	Comp.4
		10	Comp.5
		11	结算价
	单元数		11
	协变量的重新标度方法		标准化
隐藏层	隐藏层数		1
	隐藏层1中的单元数		6
	激活函数		双曲正切
输出层	因变量	1	涨跌
	单元数		2
	激活函数		Softmax
	误差函数		交叉熵

表10为神经网络信息

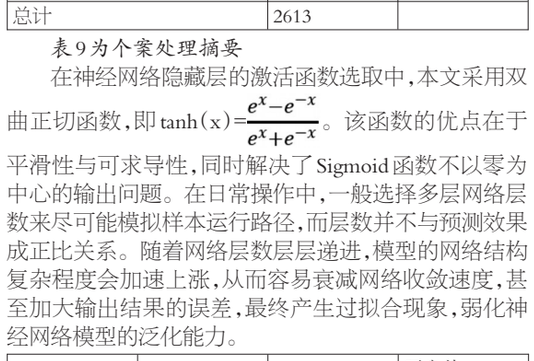

图1为MLP神经网络

由于本文所研究的对象为棉花期价相较上一个交易日的运行方向，若划分层数过多，会产生过度拟合问题。因此，本文神经网络的层数选取3层，即神经网络的输入层为1层，输入的神经变量为11个；隐含层为1层，所含神经元变量为6个，额外加一个偏差项；输出层为1层，所含神经元变量有2个。具体信息如表10与图1所示，图1表示MLP模型运行后的神经网络图，线条的粗细代表了权重的大小。

在表11模型摘要数据中，我们能够看到，模型误差在一个步骤中未出现优化减少的情况，说明模型按预定模式中止运行。在训练集中，模型的相对错误百分比为21.1%；在检验集中，模型的相对错误百分比为21.7%。由此证明，不论是训练集还是检验集，该模型判断郑棉运行方向的准确概率处于78.3%—78.9%之间。

训练	交叉熵误差	818.710
	不正确预测百分比	21.1%
	使用的中止规则	误差在一个步骤中没有减小
	训练时间	0:00:00.52
检验	交叉熵误差	341.673
	不正确预测百分比	21.7%
因变量：涨跌		
a.误差计算基于检验样本。		

表11为模型摘要

表12输出是模型对样本预测的分类结果。由表中数据可以看出，在训练集中，模型对郑棉期价向下运动的预测准确率为82%，对郑棉期价向上运行的预测准确率为75.5%。在检验集中，模型对郑棉期价向下运动的预测准确率为78.6%，对郑棉期价向上运行的预测准确率为78.1%，表明该模型对郑棉价格向下运动预测准确率略高于向上运动的准确率，模型的整体准确率处于可接受的水平。

样本	观察值	预测值		
		0	1	正确百分比
训练	0	793	174	82.0%
	1	217	668	75.5%
	总体百分比	54.5%	45.5%	78.9%
检验	0	315	86	78.6%
	1	79	281	78.1%
	总体百分比	51.8%	48.2%	78.3%
因变量：涨跌				

表12为分类结果

图2表示自变量正态化后重要性分布，按照重要性值降序分布。由图2可知，开盘价以及收盘价指标的重要性占据前两位。从市场实际运行情况来看，开盘价与收盘价指标反映短期内市场情绪走向，对下一交易日棉花期价运行方向具有重要指导性意义，说明该模型具有实际的参考性。

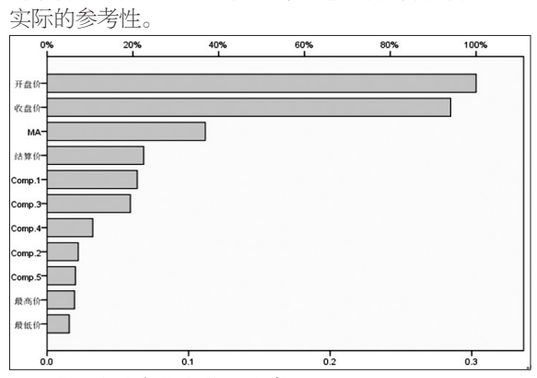

图2为正态化重要性分布

从图3预测箱形可以看出，黑色箱体表示下跌，呈现出左高右低状态，灰色箱体表示上涨，呈现左低右高状态。表示模型在实际向下运动的过程中，判断下行的概率高于其判断上行的概率；在实际向上运行的过程中，判断上行的概率高于其判断下行的概率，证明其拟合程度较高。

在图4的ROC曲线中，曲线下面积是AUC可判断试验的诊断价值，ROC曲线下的面积值在1.0和0.5之间。在AUC＞0.5的情况下，AUC越接近于1，说明诊断效果越好。AUC在0.5—0.7时准确性低。本模型的AUC=0.876，说明诊断效果较好。

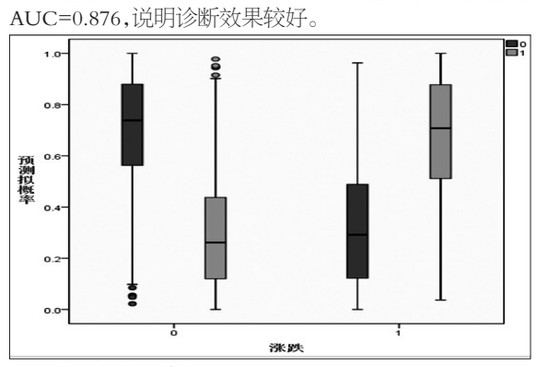

图3为预测箱形

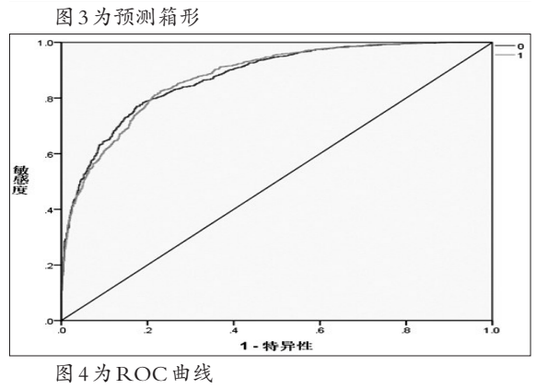

图4为ROC曲线

研究结论及建议

通过选取棉花期货盘面的相关数据，利用MLP神经网络模型对郑棉主力合约价格运行方向进行预测研究，整体拟合效果较好。不过，棉花期价变动受诸多因素影响，本文只考虑了期货市场内部的影响因素，未考虑外部基本面因素如产量、天气升贴水、需求、政策、宏观环境等，在一定程度上造成模型的脆弱性，这也是该模型未来需要改进的部分。笔者建议，可将基本面影响因素添加至模型中，充实棉花期价预测因素，进一步提高模型对期价运动方向预测的准确性。

（作者单位：华安期货）