我国通货膨胀风险的预测模型

——基于决策树 – BP 神经网络

薛　晔,蔺琦珠,任　耀

(太原理工大学 经济管理学院,山西　太原　030024)

摘　要: 选取我国 2010 年 1 月至 2015 年 3 月的月度相关数据,运用分布滞后模型对通货膨胀影响因素进行乘数效应分析,利用决策树算法对通货膨胀的影响指标进行筛选和优化,借助 BP 神经网络对通货膨胀风险等级进行预测。结果表明: 动态乘数效应显著,且通货膨胀受流动性过剩、产出缺口、国房景气指数、人民币兑美元实际汇率滞后一期的动态乘数效应较大; 所构建的决策树 – BP 神经网络模型比传统的 ARIMA 模型分类准确率高、均方误差小且对短期通货膨胀风险等级的预测效果较为理想,有望为基于大数据的宏观经济实时预测系统提供新的构建思路。

关键词: 分布滞后模型; 决策树; BP 神经网络; 通货膨胀; 风险预测

中图分类号: F821.5　　文献标识码: A　　文章编号: 1004 – 972X(2016)01 – 0082 – 08

DOI:10.16011/j.cnki.jjwt.2016.01.015

一、引言

目前,我国处于经济新常态时期,调结构稳增长成为经济活动的主要方向,在这样的经济环境下我国的通货膨胀总体处于较低的水平,但宏观经济周期的更替是经济运行中的一种客观规律,随着经济结构的调整和经济的复苏与繁荣,物价水平的上升成为必然。然而过低和过高的通货膨胀水平均不利于经济的稳定发展,物价水平的适当上涨即适度的通货膨胀会反过来促进经济的发展,但是,严重的通货膨胀将会导致人民实际购买力下降,降低人民生活水平,减少社会总需求,促使生产厂商缩小生产规模,减少对劳动力的需求,使得失业人数增加,影响社会安定,进而形成严重的社会问题。因此,我们需要提前准确预测未来的通货膨胀水平,并据此在财政政策或货币政策等方面提前做出调整,这样才能有效规避过低或过高的通货膨胀水平对经济社会带来的风险,保持社会经济的健康运行。

二、文献综述

目前国内外对通货膨胀的相关研究成果主要是从通货膨胀的影响因素及通货膨胀风险的预测等方面展开的。

(一) 通货膨胀影响因素的相关研究

Ray[1] 等运用 VAR 模型研究了金融资产价格(以股票为代表) 对印度通货膨胀的影响,结果发现,金融资产价格的变动是造成印度通货膨胀的格兰杰原因; 王晓芳等[2] 运用协整与误差修正模型研究我国股票收益与通货膨胀的关系发现,从短期来看,我国通货膨胀水平与股票市场收益呈现正相关关系,从长期来看,我国通货膨胀与股票市场收益之间存在长期均衡关系; 王维安等[3] 通过运用 VAR 模型研究发现我国房地产市场收益与通货膨胀之间存在正相关关系; Berger 等[4] 运用 B – VAR 模型研究了欧洲通货膨胀与货币供应量之间的关系,结果发现 20 世纪 80 年代之后,流动性过剩因素对欧洲通货膨胀的影响逐渐变弱; 项后军等[5] 运用非线性计量方法研究了汇率对于通货膨胀的传递效应,结果发现,汇率传递系数在低通货膨胀时期符号基本为负,在高通货膨胀时期符号基本为正; 肖争艳等[6]

收稿日期: 2015 – 10 – 21

基金项目: 国家自然科学青年基金项目(41101507); 山西省高等学校人文社会科学重点研究基地项目(2014314) 及山西省高等学校优秀青年学术带头人支持计划项目联合资助

作者简介: 薛　晔 (1974—),女,山西闻喜人,太原理工大学经济管理学院副教授,博士研究生导师,研究方向为模糊决策与优化、风险分析与风险管理、模糊信息优化处理;

任　耀 (1983—),男,山西交口人, 太原理工大学经济管理学院博士研究生,研究方向为创新管理。通讯作者。

运用 BVAR 模型研究了国际大宗商品与我国通货膨胀之间的关系，发现外部冲击对我国通货膨胀的影响仅存在于短期内。

由上述文献的梳理可知，对通货膨胀形成原因的检验已有研究成果所采用的研究方法基本是 VAR 模型及其特殊形式，但 VAR 模型是非结构化的、直接根据数据构建的模型，目的在于进行格兰杰因果关系检验、协整分析、脉冲响应及方差分解分析，缺乏相应的经济理论基础支撑，且每个解释变量对被解释变量的影响系数是没有经济意义的，故本文选择经典计量模型——分布滞后模型对我国通货膨胀的影响因素进行短期乘数、动态乘数和长期乘数效应分析，从理论上为本文所建模型提供具有经济意义的备选指标。

（二）通货膨胀风险预测的相关研究

早在 1982 年，Engle[7] 运用 ARCH(4) 模型对英国的通货膨胀水平进行了预测；Bos 等[8] 通过构建 ARIMA 和 ARFIMA 模型对美国的通货膨胀率进行了预测，结果表明 ARFIMA 模型的区间预测能力要高于 ARIMA 模型；夏荣尧[9] 运用 ARIMA 模型对我国的通货膨胀率进行了预测，认为针对通货膨胀的各项政策从实施到作用于实体经济具有一定的滞后性；张嘉为等[10] 首先运用因子预测、向量自回归等多种方法单独对我国通货膨胀率进行预测，然后将多种预测模型进行集成得到更为准确的预测结果。

由此可知，国内外对通货膨胀风险预测的研究成果主要是通过构建计量经济模型来实现的。但是，以 2013 年为元年的大数据时代已经到来，传统的计量经济模型面对实时可得的海量数据，在准确把握数据间的各种相关关系方面存在不足：(1) 传统的计量经济模型建立在抽样统计的基础上，而在大数据时代，信息化水平的提高使得海量的数据和信息实时可得，宏观经济分析已经从样本统计时代走向总体统计时代[11]。另外，随着经济的对外开放，影响宏观经济运行的指标变得纷繁复杂，经典的计量模型受到因果检验的限制导致探寻多个经济变量间复杂关系的可信度降低，而数据挖掘技术有能力通过探求相关关系而不是因果关系对大数据时代复杂的宏观经济问题进行研究；(2) 已有研究中仅考虑结构化的数据，其特点是时滞性严重，会导致对宏观经济的预测成为徒劳，而大数据中的非结构化数据如网络文本等具有很强的时效性，可以对经济形势做出最快速的预警。

总之，目前已有研究成果存在以下 3 个主要问题：(1) 通货膨胀影响因素分析方法缺乏相应的经济理论基础支撑，且每个解释变量对被解释变量的影响系数是没有经济意义的；(2) 所采用的数据大多是时滞性严重的结构化数据；(3) 通货膨胀风险的预测和评估方法受到因果检验的限制而导致可信度偏低。因此，本文首先运用由经济理论支撑的分布滞后模型对我国通货膨胀的影响因素进行乘数效应分析，结合时效性强的非结构化数据利用决策树和 BP 神经网络模型对我国短期通货膨胀风险进行预测，从而提高其可靠性。

三、相关理论及研究方法

（一）通货膨胀风险等级界定

为了确定通货膨胀所处的区间，根据通货膨胀率大小表示通货膨胀风险等级，以便于较准确地判断通货膨胀风险是否需要采取相应的控制措施。根据物价水平上涨幅度的不同，通货膨胀理论将通货膨胀分为恶性通货膨胀（15% 以上）、奔腾式通货膨胀（6% ~10%）、温和式通货膨胀（4% ~6%）和爬行式通货膨胀（2% ~3%）。殷波[12] 通过构建 DSGE 模型估算出在各种不同的货币政策规则下，中国经济应该选择的最优通货膨胀区间为 0.5% ~3%。因此结合通货膨胀理论和学者们的研究，将通货膨胀风险的第一个等级区间视为 0.5% ~3%，为无风险区 R_1；同时根据我国 2005 年以来的通货膨胀率波动情况和政府相应的调控措施可以看出，我国目前通货膨胀可容忍的上限为 5%，因此将通货膨胀风险的第二个等级区间定为 3% ~5%，为低风险区 R_2；在此基础上结合通货膨胀理论并考虑到训练样本数量应达到足够训练分类器的标准，将通货膨胀风险的第三个等级区间中风险区 R_3 为 5% ~7%，7% 以上视为第四个等级高风险区 R_4，具体见表 1。

表 1　　通货膨胀风险等级界定

风险等级 R	R_1	R_2	R_3	R_4
通货膨胀率 π	0.5% ~3%	3.1% ~5%	5.1% ~7%	7% 以上
风险描述	无风险	低风险	中风险	高风险

因为近年来尤其是 2010 年以来，我国基本没有超过 7% 的通货膨胀率，所以我国通货膨胀风险预警中需要重点关注的是低风险区和中风险区。

（二）决策树

决策树是一种分类模型，构造决策树最关键的是结点分裂属性的选择，而不同的决策树算法之间的区别是属性选择度量方法的不同。目前使用最为广泛的决策树算法包括两类：一类是基于基尼指数的 CART 算法，通过度量数据分区的不纯度来选择

分裂属性；另一类是基于信息增益的 ID3 算法，选择具有最高信息增益的属性作为该结点的分裂属性，但是 ID3 算法倾向于选择具有大量值的属性，所以 ID3 的后继算法 C4.5 使用一种称为增益率的信息增益扩充克服了 ID3 的这种缺陷。因为宏观经济分析需要大量的信息来做判断，所以本文选取基于信息增益的方法来构造决策树，同时又因为本文所选属性为连续值，所以最终选取经过改进的 C4.5 算法来构造决策树，具体如下：

设 D 为一个包含 $|D|$ 个数据样本的集合，类别属性有 m 个不同的值，对应于 m 个不同的类别集合 $C_i, i \in \{1,2,3,\cdots,m\}$，$|C_i|$ 是类别集合 C_i 中的样本个数，则要对一个给定数据对象进行分类所需要的信息熵为：

$$I(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

其中，$p_i = |C_i|/|D|$ 表示任意一个数据对象属于类别集合 C_i 的概率，使用以 2 为底的 log 函数是因为信息用二进制编码。

设一个属性 A 取 v 个不同的值 $\{a_1, a_2, \cdots, a_v\}$，利用属性 A 可以将集合 S 划分为 v 个子集 $\{D_1, D_2, \cdots, D_v\}$，其中集合 D_j 包含了数据集合 D 中属性 A 取 a_j 值的数据样本。若选择属性 A 对当前样本集进行划分，那么利用属性 A 划分当前样本集合所需要的信息熵计算公式为：

$$I_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} I(D_j) \tag{2}$$

其中，$|D_j|/|D|$ 表示 $\{D_1, D_2, \cdots, D_v\}$ 中第 j 个子集的权值，信息熵 $I_A(D)$ 的值越小，表示利用属性 A 进行子集划分的结果越“纯”（好）。

这样，利用属性 A 对当前分支结点进行相应子集划分所获得的信息增益为：

$$Gain(A) = I(D) - I_A(D) \tag{3}$$

其中，信息增益 $Gain(A)$ 表示根据属性 A 对当前结点进行子集划分所带来的所需信息熵的减少量。至此便为决策树 ID3 算法中信息增益的计算方法，而决策树 C4.5 算法中的信息增益率将属性值的数量考虑在内，它用“分裂信息”将信息增益规范化为：

$$SplitI_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{4}$$

信息增益率计算公式为：

$$GainRate(A) = \frac{Gain(A)}{SplitI_A(D)} \tag{5}$$

在每个分支结点上，决策树 C4.5 算法计算每个属性的信息增益率，从中选择信息增益率最大的属性作为在该结点上进行子集划分的属性，直到信息增益率低于某一特定阈值时停止决策树的构造。

因为决策树的构造不需要任何参数设置，适用于探测式知识发现，且容易转换成分类规则，可解释性很高，所以本文通过构造决策树对我国通货膨胀水平的影响因素进行属性选择，为下文通货膨胀风险预测模型的构模提供指标。

（三）BP 神经网络

BP 神经网络是一种基于后向传播算法的多层前馈神经网络，通过反向传播来不断调整网络的权值和阈值，使其误差平方和最小。

1. 对神经网络中所有的权值和阈值进行初始化，将它们设置为一个较小的随机数；在此基础上进行输入的正向传播，根据隐含层和输出层各单元输入的线性组合计算出相应各单元的输出，图 1 展示了隐含层和输出层的每个单元的输出的计算方法，首先计算该单元的纯输入，每个连接到该单元的输入乘以相应的权重并累加起来形成该单元的纯输入，给定隐含层或输出层中的一个单元 j，ω_{ij} 是前一层单元 i 到单元 j 的连接权重，O_i 是前一层单元 i 的输出，其纯输入 I_j 计算公式为：

$$I_j = \sum_i \omega_{ij} O_i + \theta_j \tag{6}$$

然后将激活函数 $f(I) = 1/(1+e^{-I})$ 作用于纯输入 I_j 形成该单元的输出：

$$O_j = \frac{1}{(1+e^{-I_j})} \tag{7}$$

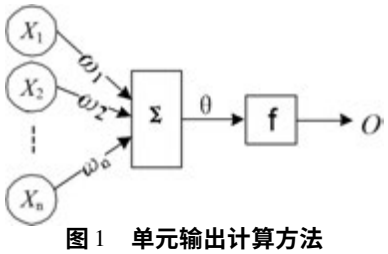

图 1　单元输出计算方法

2. 误差的后向传播及权值阈值的更新。神经网络的输出与实际输出之间的误差通过网络后向传播，并在此过程中对相应权值和偏差进行更新修改，使得整个网络的误差平方和达到最小。

输出层单元 j 误差计算：

$$Err_j = O_j(1-O_j)(T_j - O_j)$$

隐含层单元 i 误差计算：

$$Err_i = O_i(1-O_i)\sum_k Err_k \omega_{ik}$$

其中，T_j 为基于已知给定样本类别的实际输出，ω_{ik} 为单元 i 与前一层单元 k 之间的权值，Err_k 为

单元 k 的误差。在误差的传播过程中对权值和阈值进行更新以期能够较准确地反映出实际输出。

网络中权值的更新:

$\Delta\omega_{ij} = (l)\, Err_j O_j$;　$\omega_{ij} = \omega_{ij} + \Delta\omega_{ij}$

网络中阈值的更新:

$\Delta\theta_j = (l)\, Err_j$;　$v_j = \theta_j + \Delta\theta_j$

数学理论已证明三层的神经网络就能够以任意精度逼近任何复杂的非线性映射,且预测精测一般要高于其他的分类和预测方法,故本文选择 BP 神经网络来预测内部机制复杂的通货膨胀风险。但是由上述神经网络算法的处理过程可知,神经网络算法没有显式的描述输入—输出映射关系,无法对大量变量进行属性选择,因此本文先运用决策树方法进行属性选择,形成分类规则,再结合 BP 神经网络以较高的精度对我国通货膨胀风险进行预测。

四、我国通货膨胀风险预测

为了构建通货膨胀风险预测模型,首先,对我国通货膨胀影响指标进行选取,并收集整理相应指标 2010 年 1 月至 2015 年 3 月的月度数据(注: 本文用 t 表示月度,$t = 1,2,\cdots,63$);其次,利用分布滞后模型对我国通货膨胀影响因素进行乘数效应分析,包括短期乘数和动态乘数; 最后,构建决策树 - BP 神经网络模型并进行实证分析。

(一)指标选取与数据预处理

从货币供给、需求与供给、资产价格、外部冲击和通货膨胀预期五个方面来探讨我国通货膨胀的影响指标。

1. 货币供给(X_{1t})。货币学派认为货币供应量的变动是引起物价水平变动的根本原因,因此选取流动性过剩指标作为货币供给的代理指标,月度流动性过剩指标计算公式为 $X_{1t} = M_{2t}/GDP_t$,其中,M_{2t} 为第 t 个月的广义货币供应量,GDP 为第 t 个月的国内生产总值。

2. 需求与供给(X_{2t})。新凯恩斯主义学派认为通货膨胀是通货膨胀惯性、需求冲击与供给冲击的共同作用,产出水平不足导致需求过剩和物价水平的上涨。因此,本文选取产出缺口 $X_{2t} = \dfrac{(Y_t - Y_t^*)}{Y_t^*} \times 100$ 作为过度需求的衡量指标,其中 Y_t 为第 t 个月的实际国内生产总值,Y_t^* 为根据 Y_t 运用 HP 滤波法计算出的第 t 个月的潜在产出。

3. 资产价格(X_{3t}和 X_{4t})。资产价格通过不同渠道(如收入效应、替代效应等)会引起投资、消费等总需求的扩张,进而影响到通货膨胀,因此本文选取月度国房景气指数 X_{3t}和月度上证综合指数 X_{4t}作为我国资产价格的代理指标,二者均选取月末值,并对指标值作对数化处理。

4. 外部冲击(X_{5t}和 X_{6t})。在开放经济中,汇率变动会影响进口商品价格,进而间接影响一国物价水平,因此选取人民币兑美元实际汇率 X_{5t}来度量汇率水平。同时,国外原材料价格的变动会影响国内厂商的生产成本,进而对国内商品的价格水平产生影响,因此,文中选取国际油价 X_{6t}作为国外原材料价格的代理指标。

5. 通货膨胀预期(X_{7t})。姚余栋等[13]通过构建联立方程体系研究发现通货膨胀预期是影响当前我国通货膨胀的重要决定因素,因此选取通货膨胀预期为通货膨胀风险的影响指标。由于大数据时代非结构化数据实时可得,所以根据孙毅等[14]关于网络搜索行为与通货膨胀预期关联性的分析,本文选取 14 个通货膨胀核心关键词的百度指数进行主成分分析,得到月度通货膨胀预期指数 X_{7t}。

6. 通货膨胀率 π_t。通货膨胀率的计算步骤:(1)将月度 CPI 环比序列转换成以 2010 年 1 月为基期的月度 CPI 定基序列;(2)计算月度 CPI 的环比增长率;(3)得到月度通货膨胀率序列。

综上,通货膨胀风险的影响指标见表 2,所有指标均经过价格调整、季节调整处理,数据时段为 2010 年 1 月至 2015 年 3 月。M_2、GDP 和 CPI 数据来源于中国国家统计局网站,国房景气指数数据来源于中国金融信息网,上证综合指数来源于新浪财经网站,人民币兑美元汇率数据来源于国家外汇管理局网站,国际油价数据来源于美国能源信息署网站,通货膨胀核心关键词数据来源于百度指数网站。

表 2　　通货膨胀风险影响指标

指标	X_1	X_2	X_3	X_4	X_5	X_6	X_7
指标描述	流动性过剩	产出缺口	国房景气指数	上证综合指标	人民币兑美元实际汇率	国际油价	通货膨胀预期

(二)通货膨胀影响指标的乘数效应分析

宏观经济活动的构成极其复杂,而通货膨胀作为重要的宏观经济问题,在形成过程中,人们固有的行为习惯和心理思维模式、工业生产与流通中每一个环节的实现、相应的管理制度等往往存在滞后效应,均可能导致表 2 中的指标对通货膨胀风险产生动态乘数效应,而决策树 - BP 神经网络属于数据挖掘方法,它们是基于数据来进行学习训练的,缺乏相应的经济理论支撑,在分析影响因素的滞后性方面存在缺陷,因此在进行运用决策树 - BP 神经网络进行预测之前需要结合分布滞后模型来对通货膨胀影

响因素的滞后性进行检验。

1. ADF 平稳性检验与滞后期阶数的选择。为了避免伪回归的现象，先进行 ADF 平稳性检验，再根据 LogL、LR、FPE、AIC、SC、HQ 等判别准则选择最佳滞后阶数。借助 Eviews7. 2，得到平稳性检验和滞后期阶数结果见表 3。

表 3　　平稳性检验与滞后阶数选择

指标	滞后期	ADF 检验		结论
		T 统计量	P 值	
π		– 1. 118581	0. 0013	平稳
X_1	1	– 3. 800948	0. 0053	平稳
X_2	1	– 2. 991154	0. 0443	平稳
X_3	1	– 1. 869453	0. 0336	平稳
X_4	2	0. 733483	0. 0018	平稳
X_5	1	– 1. 077313	0. 0076	平稳
X_6	1	– 0. 247351	0. 0050	平稳
X_7	1	– 7. 604094	0. 0000	平稳

2. 分布滞后模型。由上文可得分布滞后模型如下:

$$\pi_t = \alpha + \sum_{\substack{i=1\\ i\neq 4}}^{\substack{j=1\\ i=7}} \beta_{ij} X_{i(t-j)} + \sum_{j=0}^{2} \beta_{4j} X_{4(t-j)} + \mu_t \qquad (8)$$

由表 3 知式(8) 中的滞后期阶数是确定的，故选用 Almon 多项式法对式(8) 中 β_{ij} 和 β_{4j} 进行估计，结果见表 4。

表 4　　分布滞后模型参数估计结果

变量	乘数	T 统计量	P 值(显著性水平 = 0. 1)
X_1	0. 135	0. 71	0. 2377
$X_1(-1)$	0. 187	0. 67	0. 0365
X_2	0. 075	2. 31	0. 1538
$X_2(-1)$	0. 161	2. 07	0. 0078
X_3	0. 189	2. 29	0. 3546
$X_3(-1)$	0. 264	2. 89	0. 0028
X_4	0. 015	0. 07	0. 1389
$X_4(-1)$	– 0. 023	– 0. 13	0. 0000
$X_{4(-2)}$	0. 006	– 0. 17	0. 0009
X_5	0. 16	2. 98	0. 3789
$X_5(-1)$	– 0. 29	– 3. 49	0. 0000
X_6	0. 005	1. 59	0. 0038
$X_6(-1)$	0. 038	1. 07	0. 0007
X_7	0. 045	8. 80	0. 1260
$X_7(-1)$	0. 091	7. 45	0. 0027
R^2	0. 9971	F	30. 02

由表 4 知，$F = 30.02$，表明方程整体是显著的。流动性过剩(X_1) 、产出缺口(X_2) 、国房景气指数(X_3) 、人民币兑美元实际汇率(X_5) 、通货膨胀预期(X_7) 的短期乘数 β_{10}、β_{20}、β_{30}、β_{50} 和 β_{70} 分别为 0. 135、0. 075、0. 189、0. 16 和 0. 045，但 P 值均大于 0. 05，故对通货膨胀风险的即期影响都不显著; 而它们滞后一期的动态乘数 β_{11}、β_{21}、β_{31}、β_{51} 和 β_{71} 分别为 0. 187、0. 161、0. 264、– 0. 29 和 0. 091，且 P 值均小于 0. 05，故对通货膨胀风险的延后影响都显著，表明这 5 个指标对通货膨胀风险的影响不是即期而是前一期; 上证综合指数(X_4) 的短期乘数 β_{40} 为 0. 015，对通货膨胀风险的影响为正效应，但 $P > 0.05$，故即期影响不显著; 而滞后一、二期的动态乘数 β_{41} 和 β_{42} 分别为 – 0. 023 和 – 0. 006，对通货膨胀风险的影响均为负效应，且 P 值均小于 0. 05，故动态效应显著且滞后一期的影响要比滞后二期的影响大得多; 国际油价(X_6) 的即期乘数 β_{60} 和滞后一期的动态乘数 β_{61} 分别为 0. 005 和 0. 038，对通货膨胀风险的影响均为正且都显著，但滞后一期的影响要远远大于即期国际油价对通货膨胀风险的影响。因此，本文选取各指标的滞后一期 $X_{1(t-1)}$、$X_{2(t-1)}$、$X_{3(t-1)}$、$X_{4(t-1)}$、$X_{5(t-1)}$、$X_{6(t-1)}$ 和 $X_{7(t-1)}$ 为下面所建决策树 – BP 神经网络模型的备选指标。

(三) 决策树 – BP 神经网络模型

为了提高通货膨胀风险预测模型的精度和可靠性，下面先用决策树方法对上面所选择的备选指标进行筛选和优化，再用 BP 神经网络模型对我国通货膨胀风险等级进行评估，最后与 ARIMA 模型进行比较分析以体现所建议的决策树 – BP 神经网络模型的优点。需要说明的是: (1) 下面的运算通过 SASEM 软件实现。(2) 本文按照分层抽样方法对原始数据集进行划分以保证训练集(目的是估计模型) 、验证集(目的是确定最优模型) 和测试集(目的是检验最终选择的最优模型的性能如何) 中通货膨胀风险等级结构相近。

1. 决策树。根据决策树算法对备选指标 $X_{1(t-1)}$、$X_{2(t-1)}$、$X_{3(t-1)}$、$X_{4(t-1)}$、$X_{5(t-1)}$ 和 $X_{6(t-1)}$ 进行筛选。

首先，根据 C4. 5 算法的信息增益率初步构造决策树，例如根结点分裂指标的选择方法如下: 根结点各指标的信息增益率见表 5，选择具有最高信息增益率的人民币兑美元实际汇率作为根结点的分裂指标，以此方法继续构造决策树的分支。

表 5　根结点各指标信息增益率(从大到小排列)

指标	$X_{5(t-1)}$	$X_{3(t-1)}$	$X_{1(t-1)}$	$X_{2(t-1)}$	$X_{7(t-1)}$	$X_{6(t-1)}$	$X_{4(t-1)}$
信息增益率	0. 4601	0. 2536	0. 2174	0. 1327	0. 0432	0. 0221	0. 0079

其次，对初步构造的决策树进行剪枝以删去不可靠的分支，由此产生一系列经过修剪的候选决策树，然后结合验证集选择误分类率(见图 2) 最小的决策树为最优模型(见图 3) 。由图 1 可知，当叶结点数量为 5 时验证集的误分类率为 0. 1473 达到最小，此时的决策树是最优模型图 3，即当选择人民币对美元实际汇率 $X_{5(t-1)}$、国房景气指数 $X_{3(t-1)}$、流动性过剩 $X_{1(t-1)}$ 和产出缺口 $X_{2(t-1)}$ 对月度通货膨胀风

险进行预测时，分类准确率最高。换句话说，这四个指标对月度通货膨胀风险的影响最为显著，这与分布滞后模型的估计结果表 3 一致，这样不仅表明决策树模型通过机器学习对数据进行训练来选取用于预测的指标结果是有效的，而且还体现了训练得到的最优决策树模型对通货膨胀风险影响较大的指标进行分析和控制的能力。例如，当人民币对美元实际汇率 $X_{5(t-1)} < 13.9282$ 且国房景气指数 $X_{3(t-1)} > 100.34$ 时，很可能发生等级 3 通货膨胀风险，因此当观测到 $X_{5(t-1)} < 13.9282$ 且 $X_{3(t-1)} > 100.34$ 时，则应对 X_5 和 X_3 进行重点观测，必要时及时采取预防措施，这样才能做到防患于未然，提高预测结果的可靠性。

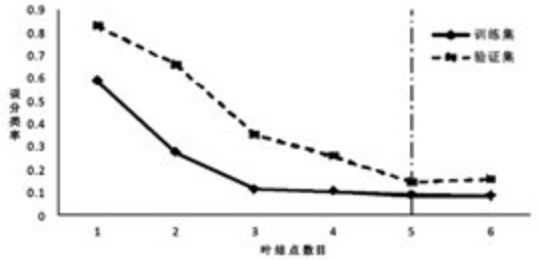

图 2 训练集与验证集误分类率趋势

图 3 最优决策树

2. BP 神经网络。根据神经网络算法及用于预测的指标 $X_{5(t-1)}$、$X_{3(t-1)}$、$X_{1(t-1)}$ 和 $X_{2(t-1)}$，对我国月度通货膨胀风险等级进行预测。将人民币兑美元实际汇率 $X_{1(t-1)}$、国房景气指数 $X_{2(t-1)}$、流动性过剩 $X_{3(t-1)}$ 和产出缺口 $X_{5(t-1)}$ 作为 BP 神经网络模型的输入变量，下个月的通货膨胀风险等级 R_t 作为 BP 神经网络模型的目标变量。

将指标数据输入神经网络模型，通过迭代（见图 4）不断调整其权值和阈值使网络的分类准确率达到最大，此时 BP 神经网络为最优。由图 4 可知，在迭代次数 $n=7$ 时，验证集的误分类率为最低，这时 BP 神经网络后向传播算法进行迭代得到的权重 ω_i（$i=1,2,3,5$）和偏倚 θ 为最优（见表 6），且最优 BP 神经网络模型（见图 5）只有一个隐藏单元。

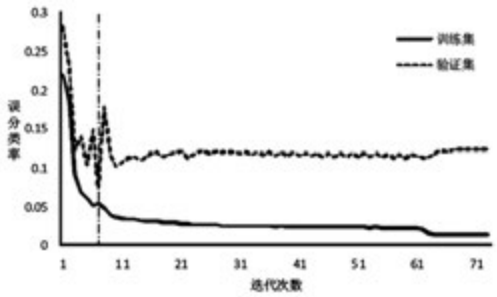

图 4 BP 神经网络迭代过程

表 6 BP 神经网络参数估计值

模型参数	ω_1	ω_2	ω_3	ω_4	ω_5	θ_H	θ_O
参数估计值	0.5724	0.3196	-0.0399	0.9974	0.7753	0.2478	0.3956

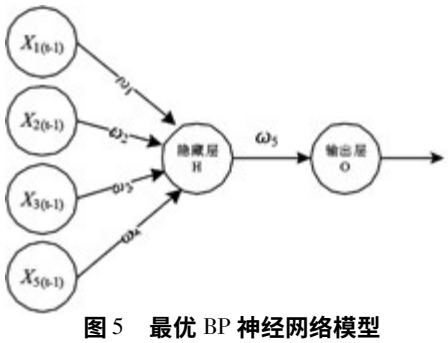

图 5 最优 BP 神经网络模型

3. 模型预测功能检验与比较分析。将测试集分别用决策树 - BP 神经网络模型和目前运用较多的 ARIMA 模型对我国通货膨胀风险等级进行预测，结果见表 7。

表 7 通货膨胀风险等级预测结果

	2011-06	2012-03	2012-10	2013-10	2014-05	2015-03	准确率	均方误差
决策树 - BP 神经网络	3（6.17%）	2（3.45%）	1（1.95%）	1（3.03%）	1（2.15%）	1（1.62%）	83.3%	0.29
ARIMA 模型	2（5.73%）	2（3.84%）	1（1.67%）	1（2.95%）	2（3.17%）	1（1.56%）	50%	0.49
实际风险等级	3（6.58%）	2（3.64%）	1（1.59%）	2（3.22%）	1（2.52%）	1（1.62%）		

注：括号中数字为对应的通货膨胀率。

由表 7 可知，本文所建议的决策树 – BP 神经网络模型准确率高且均方误差小，表明本文构建的决策树 – BP 神经网络模型对短期通货膨胀风险等级的预测功能较为理想。

4. 实证结果分析。(1) 各个影响指标动态乘数效应显著。通过运用分布滞后模型对通货膨胀的影响因素进行乘数效应分析可知，月度通货膨胀受各指标滞后一期的动态乘数效应影响较大，而即期影响不显著。这是因为本文选取的是我国 2010 年 1 月到 2015 年 3 月的月度数据，属于较高频数据，各种传导机制与效应由于不同的原因不能瞬时作用于物价。具体表现为：①货币因素。货币供给增加需要经过央行、商业银行等金融机构以及企业与个人才能最后作用于实体经济，这样便产生了时滞。②过度需求因素。过度需求的形成是在需求长期得不到满足的情况下逐渐累积而成的，暂时的或者即期的过度需求不会对通货膨胀产生显著的影响。③资产价格因素。无论是房地产还是股票价格都较多地依赖于人们的心理预期，人们很难随时跟随市场形势而动，需要一个观望与思考的时间，因此，从资产形势变化到真正影响到物价仍然需要一定的时间。④外部冲击因素。汇率通过国际贸易影响进出口产品实际价格，国际油价等原材料价格变化影响国内生产成本，这些外部冲击对国内经济产生影响是需要一定的时间的。因此，各指标对通货膨胀风险的影响存在滞后效应，但时滞是有限的，实证分析表明运用滞后一期的指标值来预测通货膨胀的风险等级会得到较准确的预测结果。

(2) 人民币兑美元实际汇率、国房景气指数、流动性过剩以及产出缺口对通货膨胀风险的影响较大。基于大数据技术的机器学习方法——决策树算法对滞后一期的指标进行了属性选择，结果表明人民币兑美元实际汇率、国房景气指数、流动性过剩以及产出缺口对通货膨胀的影响较大，其中影响最大的是人民币兑美元实际汇率。具体表现为：①人民币兑美元实际汇率。一方面汇率的变动会影响进出口原材料等非最终消费品的本币价格，进而影响到国内厂商的生产成本，最终作用于总体物价；另一方面是汇率的变动会影响进出口贸易余额，进而影响到本国外汇储备，而外汇储备的变化会最终反映到国内本币的供给上，因此人民币兑美元实际汇率会对通货膨胀产生较大的影响。②房地产价格。在房地产市场中，最主要的参与者包括政府、城市居民和农民工，城市居民对住房的需求不断推动着房地产价格的上涨，而房地产价格的上涨促使拥有土地产权的政府通过推动基础建设来实现产权增值，这些基础建设又刺激了对农民工的需求，进而拉动了农民工工资的提高即提高了社会的消费能力，推动了物价的上涨。

(3) 本文所建议的决策树 – BP 神经网络模型预测结果可靠性有明显的提高。借助大数据技术的决策树 – BP 神经网络模型对通货膨胀风险进行预测，与传统的 ARIMA 模型预测结果比较而言，前者准确率高且均方误差小，结果表明决策树通过机器学习对数据进行训练来选取用于预测的指标结果是有效的，所建议的决策树 – BP 神经网络模型对短期通货膨胀风险等级的预测结果较为理想。这是由于决策树算法本身拥有较为完善的属性选择机制，对缺失值、异常值、异方差等问题也存在内置的相应解决方法，对数据的要求较低，尤其是时效性强的非结构化数据，决策树算法不仅具有很强的适应性，而且得到的结果也较为稳健，容易形成分类规则，以便于先期监测和风险预防控制；同时，神经网络模型可以线性和非线性形式以任意精度逼近任意复杂问题，因此，决策树 – BP 神经网络模型不仅提高模型的可解释性和对指标的监测控制能力，而且还可提高预测结果的精度。

五、结论与展望

本文选取我国 2010 年 1 月到 2015 年 3 月的月度数据，运用分布滞后模型对通货膨胀影响因素的乘数效应进行理论分析，再利用决策树模型对影响我国短期通货膨胀水平的滞后一期指标进行筛选和优化，最后借助 BP 神经网络对我国月度通货膨胀风险等级进行预测，结果表明：第一，各个影响指标动态乘数效应显著，且受流动性过剩、产出缺口、国房景气指数、上证综合指数、人民币兑美元实际汇率、国际油价、通货膨胀预期的滞后一期影响较大，其中影响最大的是人民币兑美元实际汇率；第二，通过对分类准确率和均方误差两个指标的比较分析表明：所构建的决策树 – BP 神经网络模型分类准确率高且均方误差小，对短期通货膨胀风险等级的预测效果较为理想。虽然机器学习方法预测得到的结果更为准确，但由于机器学习方法是通过对数据进行训练得到模型，尤其是在选择分析指标的过程中缺乏经济理论支撑，所以将基于经济理论的计量经济模型与大数据技术相结合不仅能够发挥大数据技术的优势，而且更能准确预测宏观经济问题。

未来在大数据技术应用于宏观经济分析领域，

有以下几个方面需要不断加强和完善: 第一,对于政府统计部门而言,应利用政府的影响力优势不断加强对基于信息技术的非结构化数据(包括来源于网络的与宏观经济相关的文本、图像等) 的收集与整合; 同时,无论是关于宏观经济的结构化统计数据还是非结构化整合数据,都需要提高数据的频率,如周数据、日数据等,高频数据才能发挥大数据实时性的优势,对宏观经济进行及时预警与风险管理。第二,在大数据分析技术方面,机器学习等数据挖掘方法在图像或语音识别、自然语言处理等方面的应用已发展成熟,同样机器学习对复杂知识的获取能力会使得它应用于宏观经济分析成为一种必然,这就需要学者们进一步探究宏观经济分析与大数据技术之间成功的契合方式,并在此基础上构建专门的宏观经济预测系统,实现对宏观经济实时有效的监控和风险管理。

参考文献:

[1] RAY P, CHATTERJEE J S. The role of asset prices in Indian inflation in recent years: some conjectures [M]. Bank for International Settlements, 2001: 131 – 150.

[2] 王晓芳,高继祖. 股市收益与通货膨胀率: 中国数据的 ARDL 边界检验分析 [J]. 统计与决策, 2007(4): 86 – 88.

[3] 王维安,贺聪. 房地产价格与通货膨胀预期 [J]. 财经研究, 2005, 31(12): 64 – 76.

[4] BERGER H,ÖSTERHOLM P. Does money matter for U. S. inflation evidence from bayesian VARs [J]. Social Science Electronic Publishing,2011,57(3) :531 – 550.

[5] 项后军, 许磊. 汇率传递与通货膨胀之间的关系存在中国的“本土特征”吗? [J]. 金融研究, 2011(11): 74 – 87.

[6] 肖争艳, 安德燕, 易娅莉. 国际大宗商品价格会影响我国 CPI 吗——基于 BVAR 模型的分析 [J]. 经济理论与经济管理, 2009(8) :17 – 23.

[7] ENGLE E R F. Auto – regressive conditional heteroskedasticity with estimates of the variance of uK inflation [J]. Econometrica, 1982(50): 987 – 1008.

[8] BOS C S, FRANSES P H, OOMS M. Inflation forecast intervals and long memory regression models [J]. International Journal of Forecasting, 2002,18(2): 243 – 264.

[9] 夏荣尧. 基于 ARIMA 模型的我国通货膨胀预测研究 [D]. 长沙: 湖南大学, 2009.

[10] 张嘉为, 索丽娜, 齐晓楠. 基于 TEI@ I 方法论的通货膨胀问题分析与预测 [J]. 系统工程理论与实践, 2010 (12): 2157 – 2164.

[11] 刘涛雄, 徐晓飞. 大数据与宏观经济分析研究综述 [J]. 学科前沿, 2015(1): 57 – 64.

[12] 殷波. 中国经济的最优通货膨胀 [J]. 经济学(季刊) , 2011(3): 821 – 840.

[13] 姚余栋, 谭海鸣. 通胀预期管理和货币政策——基于“新共识”宏观经济模型的分析 [J]. 经济研究, 2013 (6): 45 – 57.

[14] 孙毅, 吕本富, 陈航,等. 大数据视角的通货膨胀预期测度与应用研究 [J]. 管理世界, 2014(4) :171 – 172.

Research on Inflation Risk Forecast of China Based on the Big – data Technology

XUE Ye, LIN Qi – zhu, REN Yao

(College of Economics and Management, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: This paper firstly analyses these factors that influence the inflation in China based on Distributed – Lag Model theoretically, then the attributes that influence the inflation risk are selected and optimized using the Decision Tree C4. 5 algorithm, and then the model of Decision Tree – BP neural networks (DT – BPNN) model based on big – data technology is suggested. ; the monthly inflation risk level is forecasted with the BP neural networks. The monthly data from January 2010 to March 2015 is collected for empirical analysis. The result shows the dynamical multiplier effect is significant, monthly inflation is affected by distributed – lag multiplier effect of indexes such as the excess liquidity, the output gap, the national housing boom index, the real exchange rate of RMB against the U. S. dollar, and the real exchange rate of RMB against the U. S. dollar has the greatest impact on monthly inflation. The forecast results of the proposed model are compared with those results of the ARIMA model, and our model have higher accuracy rate and lower mean square error. This also provides a new perspective for the construction of the real – time macro – economy assessment system based on the big – data technology.

Key words: Distributed – Lag Model; decision tree; BP neural networks; inflation; risk forecast

(责任编辑: 张爱英)