

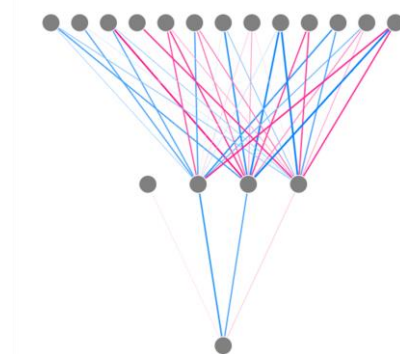
基于机器学习的社会融资规模的预测

✍️：申屠业勤 执业证书编号：S1230519120001
☎️：021-80106016
✉️：shentuyeqin@stocke.com.cn

报告导读

将人工神经网络应用与社会融资规模的预测，采用对 Adaptive-Lasso 方法筛选出宏观审慎管理体系下最有代表性的变量，将社会融资规模与金融机构：住户贷款的同比增速、私营非金融部门的偿债比率、个人住房贷款加权平均利率、私营非金融部门的偿债比率、加权平均利率：个人住房贷款、居民部门债务率、中央政府债务率、地方政府债务率、非金融部门债务率 M2 增速，GDP 增速以及地产去库存周期代入人工神经网络进行迭代，预测未来社会融资规模及增速。

计算流程 AI 化示意图



数据支持人：陈昊

风险提示：市场环境变动风险，模型基于历史数据的表达，数据集固有局限，未来预测存在误差

正文目录

1. 研究背景.....	3
2. 研究目标.....	3
3. 分析方法.....	3
3.1. Adaptive-Lasso 正则方法.....	3
3.2. BP 神经网络预测模型.....	4
4. 建模过程.....	5
4.1. 主要步骤.....	5
4.2. 数据探索.....	5
4.2.1. 变量解释.....	5
4.2.2. 数据预处理.....	5
4.3. 模型构建.....	6
4.3.1. Adaptive-Lasso 变量选择.....	6
4.3.2. 社融同比增速预测.....	6
4.4. 模型评估及优化.....	8
5. 模型结论与研究展望.....	9
5.1. 模型结论与解释.....	9
5.2. 研究与应用展望.....	9

图表目录

图 1: 单一隐藏节点人工神经网络预测模型网络拓扑图.....	7
图 2: 社融同比增速真实值与预测值对比图.....	8
图 3: 模型综合误差与神经网络隐藏层数量的关系.....	9
图 4: 三维隐藏节点人工神经网络预测模型网络拓扑图.....	9
表 1: 基于 Adaptive-Lasso 算法的变量选择结果.....	6
表 2: 社融同比增速及其相关因素历史数据和预测表.....	7

1. 研究背景

根据已有文献,总的来看,我国很多研究者已经对宏观经济数据的影响因素及宏观经济数据的预测进行了很多研究,部分研究者先建立被解释变量与各待定的影响因素之间的多元线性回归模型,运用最小二乘估计方法来估计回归模型的系数,通过系数能否通过检验来变量之间的关系,这样的结果对数据的依赖程度很大,并且普通最小二乘估计求得的解往往是局部最优解,同时因为预测因子间存在相关性,常规线性模型会存在如下缺陷,比如过度拟合,预测变量共线性带来不准确的参数估计,后续的检验可能就会失去应有的意义。而与此同时,更多的研究者是基于经验进行的预测,不可避免地带来一些主观性。

近几十年来,随着现代统计技术的不断完善和发展,由此可见,机器学习方法对于无法通过理论分析得出规律的复杂问题有着极其优越的拟合效果,近年来,机器学习算法凭借其高效的拟合预测效果,逐渐在各个研究领域被广泛应用,部分学者通过引入人工神经网络、支持向量机等机器学习算法来尝试解决相关分类与预测的问题。基于人工神经网络的非线性回归预测模型的有效性与基于神经网络的非线性回归对于解决无法得出具体函数表达的或者本身就没有具体的函数表达的问题的有效性和优越性被相继证明。由此可见,对新的数据运用新的方法来考察宏观经济数据的影响因素及其预测是可行并且有价值的。本文在已有研究的基础上运用 Adaptive-lasso 变量选择方法来研究影响宏观经济数据变化的因素,在其基础上根据反向传播算法建立人工神经网络(Artificial Neural Network, ANN)来预测未来短期内的宏观经济数据。其优点为:新的现代统计方法的运用,能更好的度量各因素的影响程度,从而全面地反映各指标与被解释变量的关系。

2. 研究目标

本次建模目标在于利用 2011 第三季度到 2019 年第二季度的宏观经济数据,采用 Adaptive-Lasso 方法,梳理影响社融同比增速关联指标的有关数据,分析并识别影响社融同比增速的关键影响因素,构建社融同比增速预测模型。进而利用 BP 神经网络对未来(2019 年第三季度)的社融同比增速进行预测,并提出相应的投资建议。

3. 分析方法

3.1. Adaptive-Lasso 正则方法

在以往的文献中,大多使用普通最小二乘法来对回归模型的系数进行估计,预测变量的选取则采用的是逐步回归。然而,无论是最小二乘法还是逐步回归,都有其不足之处。它们一般都局限于局部最优解而不是全局最优解。如果预测变量过多,子集选择的计算过程具有不可实行性,且子集选择具有内在的不连续性,从而导致子集选择极度多变。Lasso 是近年来被广泛应用于参数估计和变量选择的方法之一,并且 Lasso 进行变量选择在确定的条件下已经被证明是一致的。本文选用了 Adaptive-Lasso 方法来探究社融同比增速与各因素之间的关系。

Lasso 是由 Tibshirani (1996)提出的将参数估计与变量选择同时进行的一种正则化方法。Lasso 参数估计被定义如下:

$$\hat{\beta}(\text{lasso}) = \arg \min^2 \left\| y - \sum_{j=1}^p x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (3-1)$$

其中, λ 为非负正则参数, $\lambda \sum_{j=1}^p |\beta_j|$ 称为惩罚项。

Lasso 方法虽然可以解决最小二乘法和逐步回归局部最优估计的不足,但是其自身需要满足一定的苛刻条件。OU 提出了一种改进的 Lasso 方法,其改进之处在于给不同的系数加上了不同的权重,称为 Adaptive-Lasso 方法,定义如下:称为 Adaptive-Lasso 方法,定义如下:

$$\hat{\beta}^{*(n)} = \arg \min^2 \left\| y - \sum_{j=1}^p x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^p \hat{\omega}_j |\beta_j| \quad (3-2)$$

式中, 权重 $\hat{\omega}_j = \frac{1}{|\beta_j|^\gamma}$ ($\gamma > 0$), $j = 1, 2, \dots, p$, $\hat{\beta}_j$ 为由普通最小二乘法得出的系数。

3.2. BP 神经网络预测模型

人工神经网络算法是当今机器学习领域中比较热门的一类算法, 已有研究证明三层人工神经网络^①可以拟合任何连续有界函数, 因此人工神经网络已经得到了广泛应用。人工神经网络模型是由大量处理单元单向加权相连而成的网络, 该处理单元称为人工神经元。人工神经网络主要由输入层、隐藏层以及输出层组成。其中输入层神经元将输入值传递到隐藏层。隐藏层神经元将输入值加权求和, 并通过激励函数产生一个输出值发送至输出层。隐藏层神经元的计算公式如公式 (3-3) 所示:

$$H = f(\vec{l} \times \vec{w}_{ih}) \quad (1-3)$$

其中 $\vec{l}$ 为输入向量, $\vec{w}_{ih}$ 为隐藏层权重参数向量, H 为隐藏层某个神经元的输出。函数 f 称为激励函数, 一般选取可微非线性函数作为激励函数。

输出层的各神经元将隐藏层的计算结果加权求和并产生输出结果, 所有输出神经元的输出结果即为神经网络模型的输出。输出层神经元计算公式如下:

$$o = \vec{h} \times \vec{w}_{ho} \quad (3-4)$$

其中 $\vec{h}$ 为所有隐藏神经元的输出, $\vec{w}_{ho}$ 为输出层的权值参数。

模型输出与目标输出的差值称为误差, 若误差不为 0, 则模型会根据误差调整其权值, 使得误差向 0 逼近。根据误差调整权值的算法称为学习规则, 人工神经网络可以根据学习规则的不同进行分类, 其中 BP 神经网络是一类应用比较广泛的算法。在 BP 神经网络算法中, 当误差非 0 时, 算法会根据误差, 依次修改输出层、隐藏层的连接权值, 从而使得误差达到允许的范围之内, 从而实现对数据的拟合。

假设现有 K 个训练数据组成的训练数据集, 即 $\{(x_i, y_i)\}$, $i = 1, \dots, K$ 。其中 x_i 为输入值, y_i 为目标输出。从节点 i 到节点 j 的输入表示为 l_{ij} , 节点 i 到节点 j 的权值表示为 w_{ij} 。BP 神经网络学习过程具体如下:

第一步, 初始化模型。人为设定隐藏层的节点数 l ; 根据样本的输入输出维度确定输入层节点数 n 和输出层节点数 m ; 初始化神经网络的参数, 即对 w_{ij} 赋予接近于 0 的随机值

第二步, 用训练数据 (x_i, y_i) 对模型进行训练, 根据 BP 网络的学习规则进行参数调整, 参数调整的具体步骤如下:

将训练数据输入神经网络, 通过公式 (3) 计算出隐藏层各个节点的输出 H ; 然后通过公式 (5) 计算出模型输出 o 。

计算输出层节点的实际输出值 o 与目标输出值 y 之间的偏差, 此偏差是当前模型的误差:

$$e_k = y_k - o_k, \text{ 其中 } k = 1, 2, \dots, m \quad (3-5)$$

BP 神经网络认为模型误差是各个节点产生的误差综合作用的结果, 因此依次反向计算各个节点贡献的误差项。

对于网络的输出单元 k , 计算它的误差项

$$\delta_k = o_k(1 - o_k)e_k \quad (3-6)$$

^① 符合该结论的神经网络需要满足以下条件: 隐藏层采用 sigmoid 函数作为激励函数; 输出层采用线性单元

对于网络的隐藏单元 h ，计算它的误差项

$$\delta_h = o_h(1 - o_h) * \sum_{k=1}^m w_{kh} \delta_k \quad (3-7)$$

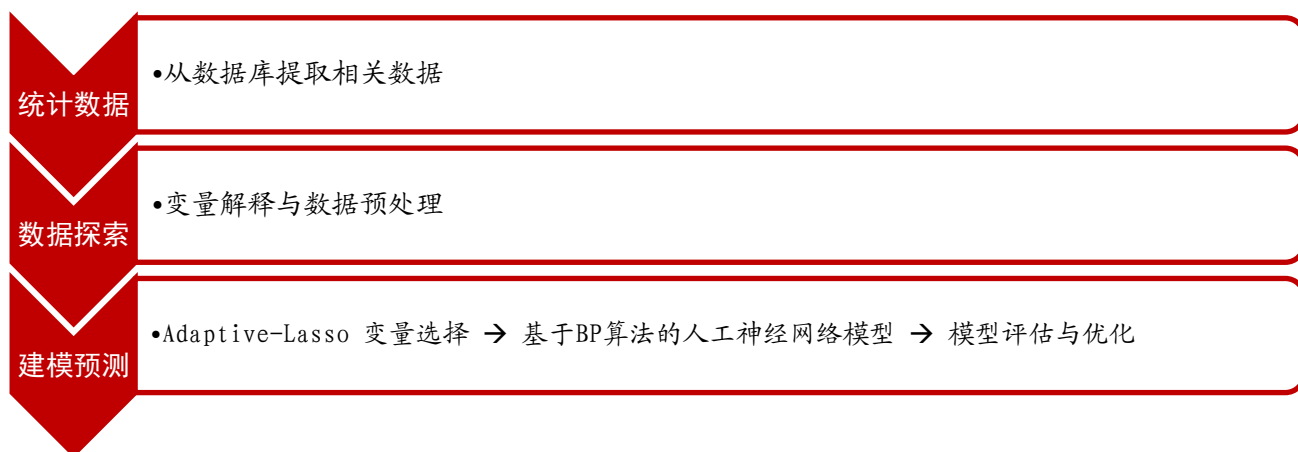
根据误差项，更新节点权值，其中 η 为学习速率，一般设置为接近于 0 的值

$$w_{ij} = w_{ij} + \Delta w_{ij}, \text{ 其中 } \Delta w_{ij} = \eta \delta_i l_{ij} \quad (3-8)$$

第三步，将训练数据依次放入模型，重复第二步训练方法，直到模型误差低于目标要求或者迭代次数达到一定设定值停止训练。至此神经网络训练完毕，可输入相应的 x 值利用该模型进行预测。

4. 建模过程

4.1. 主要步骤



4.2. 数据探索

4.2.1. 变量解释

基于宏观审慎评估，我们选择了与社会融资规模相关的指标进行梳理，初始选取的数据包括:金融机构:住户贷款的同比增速，城镇居民人均可支配收入同比增速，私营非金融部门的偿债比率，个人住房贷款加权平均利率，利息债务比，十年期国债与三年期国债差值，居民部门债务率，中央政府债务率，地方政府债务率，广义政府部门债务率，非金融企业债务率，非金融部门债务率，社会融资规模与 GDP 之比，M2，GDP，房地产投资，房地产新开工以及地产去库存周期。主要思路是通过宏观审慎框架建立与社融的相关性；其中指标涵盖收入水平，杠杆比率，宏观运行数据，银行流动性以及定价数据，分部门融资数据以及房地产相关数据。

4.2.2. 数据预处理

本次研究中所用的宏观经济数据集共包含 19 个变量，包括 18 个特征（自变量）和 1 个被解释变量（因变量），包含的数值范围下至 0，上至 100 以上。而最适合神经网络的运行的数据集应该是位于 0 附近的狭窄波动范围内，并且分布服从一钟形曲线。为此，我们通过标准化处理将输入数据转化为一个 0~1 的标准化范数，即：

$$Z = \frac{x - \text{Min}(x)}{\text{Max}(x) - \text{Min}(x)} \quad (4-1)$$

4.3. 模型构建

4.3.1. Adaptive-Lasso 变量选择

运用 LARS 算法来模拟式(3-2)的 Adaptive-Lasso 估计, 对于每一个给定的 γ , 该算法会寻找一个最优的 λ_n 。此处取 $\gamma = 1$, 用 R 语言编制相应的程序后运行得到如下结果, 如表 1 所示。

表 1: 基于 Adaptive-Lasso 算法的变量选择结果

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
17.77853	0.00000	-0.83560	1.98778	0.00000	-3.00479	-1.12611	1.18748	-0.23196	0.00000
X11	X12	X13	X14	X15	X16	X17	X18		
0.00000	0.32065	3.94506	0.37771	-1.05992	0.00000	0.00000	-0.25634		

由表 1 可看出, 城镇居民人均可支配收入同比增速(X2)、十年期国债与三年期国债差值(X5)、广义政府部门债务率(X10)、非金融企业债务率(X11)、房地产投资(X16)以及房地产新开工(X17)的系数为 0, 因而, 在模型建立的过程中这几个变量被剔除了。主要的排除原因包括①多重共线性; ②与被解释变量的相关性不显著; ③追踪数据的有效性, 同时需要说明的是, 影响因变量的因素不局限于文中所用的数据, 模型只是对已有数据的表达。

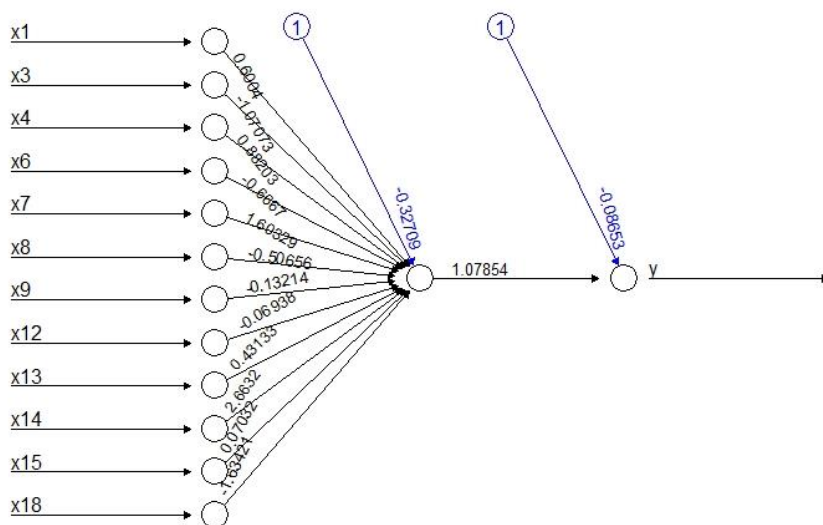
综上所述, 利用 Adaptive-Lasso 方法识别影响社融同比增速的关键影响因素是金融机构: 住户贷款的同比增速、私营非金融部门的偿债比率、个人住房贷款加权平均利率、私营非金融部门的偿债比率、加权平均利率: 个人住房贷款、居民部门债务率、中央政府债务率、地方政府债务率、非金融部门债务率 M2 增速, GDP 增速以及地产去库存周期。

4.3.2. 社融同比增速预测

对 Adaptive-Lasso 变量选择方法识别的影响宏观经济数据的因素建立 BP 神经网络预测模型 (基于 neuralnet 包, 可自动实现神经网络迭代次数、学习速度的设置。), 在该模型中, 12 个特征分别对应一个输入节点, 每一个输入节点后跟随相应的隐藏节点, 最终得到单一的预测社会融资同比增速的单一节点。城镇居民人均可支配收入同比增速(X2)、私营非金融部门的偿债比率(X3)、个人住房贷款加权平均利率(X4)、十年期国债与三年期国债差值(X6)、居民部门债务率(X7)、中央政府债务率(X8)、地方政府债务率(X9)、非金融部门债务率(X12)以及非标融资同比增速(X18)指标的未来两个报告期数值通过得出。代入所建立的 BP 神经网络预测模型, 按照随机原则划分训练数据集与测试数据集, 比例为 25:8。

初始状态下, 这是一个单一隐藏节点的多层前馈网络。得益于网络结构拓扑图, 我们可以窥见人工神经网络黑箱的冰山一角, 每个输出节点至隐藏层的权值亦如图所示。

图 1：单一隐藏节点人工神经网络预测模型网络拓扑图



资料来源：浙商证券研究所

实际上，具有一个隐藏节点的神经网络与目前计量经济学常用的线性回归模型类似，每个输入节点与隐藏节点间的权重类似于回归系数，也就是说，如果构建一个与上述神经网络原理相同的线性回归模型，则模型整体的相关系数即为神经网络预测值与真实值之间的相关系数。

通过进一步评估与优化（详见 4.4 节），我们得到了预测性能更佳的模型，通过这一模型预测得到的社融同比增速预测值为 10.62224465%，相关数据见表 2，其中红色字体的数据为预测数据。图 1 为 BP 神经网络社融同比增速真实值与预测值对比图，可见预测误差位于可接受范围内。

表 2：社融同比增速及其相关因素历史数据和预测表

期数	X1	X3	X4	X6	X7	X8	X9	X12	X13	X14	X15	X18	Y
2019Q3	0.173	19.283	5.68	0.3285	54.976	16.341	21.324	249.206	0.278	8.49	6.33	11.872	10.622
2019Q2	0.17	19.284	5.7	0.266	55.304	16.413	21.534	249.591	0.293	8.5	6.3	11.848	10.962*
2019Q1	0.176	19.296	5.7	0.37	54.278	16.250	21.420	248.385	0.383	8.6	6.4	11.996	11.169*
2018Q4	0.182	19.2	5.75	0.374	53.199	16.528	20.422	243.652	0.214	8.1	6.6	10.743	10.183*

预览已结束，完整报告链接和二维码如下：

https://www.yunbaogao.cn/report/index/report?reportId=1_8462

