

基于深度神经网络结构的互联网金融市场动态预测

赵洪科¹ 吴李康² 李 徵² 张 兮¹ 刘 淇² 陈恩红²

¹(天津大学管理与经济学部 天津 300072)
²(大数据分析与应用安徽省重点实验室(中国科学技术大学) 合肥 230027)
(hongke@tju.edu.cn)

Predicting the Dynamics in Internet Finance Based on Deep Neural Network Structure

Zhao Hongke¹, Wu Likang², Li Zhi², Zhang Xi¹, Liu Qi², and Chen Enhong²

¹(College of Management and Economics, Tianjin University, Tianjin 300072)
²(Anhui Province Key Laboratory of Big Data Analysis and Application (University of Science and Technology of China), Hefei 230027)

Abstract In recent years, the Internet financial market has achieved rapid development across the globe. In the meantime, Internet finance has become a hot topic in academia. Compared with traditional financial markets, the Internet financial market has higher liquidity and volatility. In this paper, the dynamics (daily trading amount and count) of the Internet financial market is studied and a prediction model is proposed based on deep neural network for fusion hierarchical time series learning. Firstly, the model can process the multiple sequence (macro dynamic sequence and multiple subsequences) feature as the input variables. And then, an attention mechanism is proposed to fuse the input variables from both the time and subsequence feature dimensions. Next, the model designs an optimization function based on the stability constraint of the sequence prediction, which makes the model have better robustness. Finally, a large number of experiments have been carried out on real large-scale data sets, and the results have fully proved the effectiveness and robustness of the proposed model in the dynamic prediction of Internet finance market.

Key words Internet finance; time series; dynamic prediction; deep neural network; sequential modeling

摘 要 近些年,互联网金融市场在国内外迅速发展;同时,针对互联网金融市场的研究也成为了学术界的热点.相比于传统金融市场,互联网金融市场具有更高的流动性和易变性.针对互联网金融市场的动态(日交易量和日交易次数)进行研究,提出了基于深度神经网络结构的融合层次时间序列学习的预测模型.首先,该模型可以实现对多序列(市场宏观动态序列和多种子序列)特征变量输入的处理,并且在时间和序列特征 2 个维度上利用注意力机制来融合输入变量.其次,模型设计了基于预测序列平稳性约束的优化函数,使得模型具有更好的稳健性.最后,在真实的大规模数据集上进行了大量的实验,结果充分证明了所提出的模型在互联网金融市场动态预测问题上的有效性与稳健性.

关键词 互联网金融;时间序列;动态预测;深度神经网络;序列建模

中图法分类号 TP18

互联网金融(Internet finance, ITFIN)^[1]是指传统金融机构与互联网企业利用互联网技术和信息通信技术等实现资金融通、支付、投资和信息中介等服务的新型金融业务模式.其中,网络小额借贷(peer-to-peer lending or P2P lending)^[2-5]和众筹(crowdfunding)^[6-7]是互联网金融中2种发展比较成熟的模式.例如网络借贷中的Lendingclub, Prosper、人人贷等,众筹中的Indiegogo, Kickstarter和国内的众筹网等.最近几年,国内互联网金融市场持续保持着指数级的增长速度,2017年中国已超越美国成为全球最大的网络借贷交易市场,网贷平台综合成交量达1 163.98亿元^[7].根据世界银行预测,2025年全球众筹市场规模将达到3 000亿美元;发展中国家将达到960亿美元,其中我国将占52%^[8].

互联网金融快速发展的背后却存在着政策滞后、监管缺失和相关研究不足的问题,导致了行业乱象、平台停业、清盘等频繁发生.追踪和预测市场动态(例如日交易量)可以及时推测和把握平台的系统风险,对用户和平台管理都有极重要意义.然而,在互联网环境中,金融市场的流动性增强、易变性更高,使得金融市场的动态跟踪和预测变得十分困难.并且,网络环境下影响变量增多,多种序列的耦合关系更加复杂.例如对具体互联网金融市场,除市场宏观动态序列以外,平台下各种类别金融子市场的动态对平台整体动态也有显著影响.

传统的金融时间序列预测算法大多以目标变量序列作为主要的研究对象.其中,具有代表性的技术包括自回归模型^[9]、向量自回归模型^[10-11]等.随着深度学习技术发展,循环神经网络(recurrent neural network, RNN)在序列问题上被广泛应用,并展现出比传统线性模型更好的学习能力^[12].尤其是在金融市场的动态预测问题中,深度学习模型具有良好的非线性映射能力和较强的拟合泛化能力,因此能够更好地建模金融市场的易变特征、非线性关联以及时间序列依赖关系.在各种深度神经网络模型中,长短期记忆神经网络(long short-term memory network, LSTM)通过引入门结构,保留序列特征中需要记忆的长时信息,有效地解决了长序列依赖问题^[13].另外,借助神经网络强大的异构数据处理能力,部分学者基于长短期神经网络结构设计了融合变量特征的预测模型^[14].例如研究者通过结合新闻文本信息对股市涨跌进行动态预测^[15],还有研究工作利用图片等异构信息辅助预测社交媒体的流行度变化^[16].最近研究中发现现在序列预测问题中引入

注意力机制(attention mechanism)^[17]能够更加有效地快速筛选出对当前任务更为关键的信息,从而进一步增强模型的预测能力.然而,传统的注意力机制主要从时间维度进行设计,无法区分多变量时间序列间的不同影响.

本文针对互联网金融市场的动态序列(具体包括市场日交易量和日交易次数),设计了基于长短期记忆神经网络结构的预测模型.模型融合市场整体的宏观动态序列以及各种子类市场的动态序列,且该模型从时间和序列特征2个维度设计2种注意力机制,建模特征的时间依赖关系和多序列输入的共同影响.最后,模型设计了基于预测序列平稳性约束的优化函数,使得模型具有更好的稳健性.通过在真实互联网金融平台大规模数据上实验验证的结果,充分说明了本文所设计方法在预测互联网金融市场动态问题上的有效性.

1 相关工作

在金融场景中,市场信息随着时间而动态变化.因此基于时间序列的分析与预测是该领域研究所关注的重点.本节首先介绍传统的时间序列预测方法,然后介绍深度学习模型和注意力机制在时序问题中的相关研究,最后讨论金融场景中序列预测的最新相关研究工作.

1.1 传统时间序列预测方法

根据模型的输出结果对传统经典模型进行分类,预测模型可以分为随机时序模型和确定性时序模型.传统时间序列模型一般需要严格的数学原理作为支撑,并需要较严格的条件加以约束,从而借助外推原则来推测未来变化.

1) 确定性模型

在很多实际应用问题中,时间序列的变化由多种因素叠加或耦合得来,如季节变化、趋势因素、周期性变化和不规则性等因素.为了使时间序列预测消除不规则性因素带来的影响,学者进行了相关研究.在该类预测模型中,常采用分解方法、移动平均法、指数平滑法、季节系数法来构建时间序列分析模型.

① 分解方法^[18].分解方法的基本思想是将预测数据进行分解,得到季节因子、循环因子、误差因子、趋势因子和随机因子等.该类方法中一些新设计的优化调整可看做是在其原始思想上的延伸.虽然这种方法在独自使用时效果比较差,但是其作为一种基本获取数据特征的方法,仍然有着不错的应用场景.

② 移动平均法^[19].该方法的基本思路是先计算一组历史序列数据的平均值,然后以该平均值作为对未来时期的预测.因此,当时间序列受到周期性波动和随机波动的影响时,序列波动很大且很难显示事件的变化规律,这时使用移动平均法可以消除这些因素的影响并得到相对稳定的趋势线.移动平均法使用方便,但由于预测精度不够故实用性不高.

③ 指数平滑法^[20].该方法是一种特殊的加权移动平均法,会将较近时间距离内的数据进行加权.指数平滑法和指数平滑规则与全期平均值和移动平均长度兼容,该方法预测过程不会丢弃过往数据,并且结果会基于过往数据逐渐变化.

④ 季节系数法^[19].一些应用场景中,在不同的季节,时间序列会受到不同因素的影响.例如,观察金融时间序列数据,可以发现在较长的时间周期内,数据的变化显示出明显的季节性周期.为了适应季节性因素,季节系数法是一种基于季节周期性变化来进行预测的方法,其系数用来辅助模型给出更准确的预测结果.

2) 随机性模型

统计学领域学者通过利用随机理论研究时间序列,发现时间序列中由许多随机因素共同作用引起的不规则变化并不是完全混乱而是具有一定规律性.在这种现象启发下,设计基于随机理论的预测模型引起了众多学者的关注.这类模型一般基于以下的流程:首先,观察时间序列数据分布确定一些较为合理的规约条件,然后采用演绎推理得到刻画该时间序列的理论模型.如果理论模型满足实际数据特点,则建立实际模型,进而用来进行时间序列的分析和预测.该类模型中比较有代表性的包括 ARMA^[20], ARIMA^[21], ARCH^[22], GARCH^[23]等.

1.2 基于深度学习的时序预测方法

深度神经网络能够通过高维特征的非线性变化获得相比于传统方法更好的表示能力.在时间序列预测问题中,应用最为广泛的深度学习模型是循环神经网络(recurrent neural network, RNN)^[12]及长短期记忆神经网络(long short-term memory network, LSTM)^[13].循环神经网络通过引入循环机制来处理时间序列数据中的前后依赖关系.LSTM 在 RNN 的基础上,通过引入“门”结构,对前序信息进行筛选,将不重要信息有选择地遗忘,从而解决了 RNN 在长序列问题中梯度消失的困难,进一步强化学习时间序列中长距离信息的依赖关系.

1.3 基于注意力机制的时序预测方法

注意力机制是一种基于深度学习框架的优化方法^[24].目前,注意力机制在自然语言处理、计算机视觉和语音识别等领域都得到了很好的应用.由于注意力机制对准确率等指标提升明显,因此很多研究人员已经在探索如何将其应用到更多需要优化的场景中.具体地,注意力模型借鉴了人类视觉注意力机制,即人类在获得信息时,会优先关注信息中重要的或者自己目前需要的部分.事实上,基于视觉的注意力机制是人类在观察时所使用的信号过滤机制.在时间序列研究中,也有相关工作探索性地应用注意力机制,并取得了一定的成功.然而,目前针对时间序列问题的注意力设计大部分只从时间维度描述序列的作用关系,而不能建模多序列特征之间的关系,这是本文设计模型时的重要出发点.

1.4 金融市场时间序列分析与预测

金融时序预测是以金融市场中时间序列变量作为数据和研究对象,构建数学模型以分析建模市场的变动规律,并对未来市场的动态变化进行预测的研究.传统方法对金融市场的时序预测大部分是利用有效市场假说和随机理论来对价值的波动率进行预测,而并非直接预测价值的本身.近些年来,越来越多研究者从不同领域开始对金融的时序预测问题展开了研究,例如基于混沌理论的交易序列预测^[25]等.特别是随着计算机技术尤其是机器学习领域的技术发展,越来越多研究者开始关注机器学习方法在时序分析与预测问题上的应用,提出了众多的学习模型,例如深度网络模型^[26]、支持向量机模型^[27]以及基于聚类算法的模型^[28]等.这些模型可以根据历史数据来分析市场的变化模式,并进一步对市场未来走势进行预测.

针对互联网金融场景中的时间序列问题,有专家学者进行过相关研究并提出了一些预测模型,例如马尔科夫模型^[29]、转化回顾模型^[30]、生存分析模型^[31]和集成学习模型^[32]等.

2 长短期记忆神经网络与注意力机制

本节首先介绍长短期记忆神经网络模型的结构与原理,然后针对时间序列中重要信息的提取,分析注意力机制在序列预测问题中的设计与构建.

2.1 长短期记忆神经网络

长短期记忆网络(long short-term memory network, LSTM)^[13]在经典循环神经网络 RNN 的

基础上,通过对网络神经元内部添加非线性“门”(gate)结构,实现有选择地遗忘非重要信息,并对前序重要信息进行加强.因此,LSTM 避免了 RNN 对较长序列进行训练时产生无法避免的梯度消失问题.具体地,LSTM 的门结构定义为

$$\begin{aligned} i_t &= \delta(W_i x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + \hat{b}_i), \\ f_t &= \delta(W_f x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + \hat{b}_f), \\ c_t &= f_t c_{t-1} + i_t \tanh(W_c x_t + W_{hc} h_{t-1} + \hat{b}_c), \\ o_t &= \delta(W_o x_t + W_{ho} h_{t-1} + W_{co} c_t + \hat{b}_o), \\ h_t &= o_t \tanh(c_t), \end{aligned} \quad (1)$$

其中, i_t , f_t , c_t , o_t 以及 h_t 分别表示输入门、遗忘门、记忆模块、输出门以及神经元的隐状态.LSTM 的遗忘门和输入门结构通过有选择地保留历史信息,从而有效地传播梯度,避免了时序信息较长时导致的梯度消失问题.LSTM 解决了 RNN 无法有效学习时间序列中的长距离信息依赖关系的问题,因此在很多时序建模问题中得到了很好的应用.

2.2 注意力机制

时序问题中的注意力机制大部分基于深度循环网络的编解码(encoder-decoder)过程^[33].该过程将学习模型分为两大模块,首先是由一个单层或多层 RNN 构成的编码器,将输入序列按照时间关系输入编码器中,用于学习已知序列的前后依赖关系以及当前状态表示;得到最后一个时刻的隐状态将其保留,称之为向量 C ,此向量保留了输入序列的动态信息以及当前的序列状态.然后,同样由类似结构的神经网络单元组成一个解码器,将编码向量 E 转换为预测长度为 T' 的时序信息,其每一个时刻 j 的输入是由向量 E 与目标值序列 $(y_1, y_2, \dots, y_{j-1})$ 共同映射而得到的向量;时刻 j 的输出值即为对应时刻的

预测值,即:

$$\begin{aligned} E &= F(x_1, x_2, \dots, x_T), \\ y_j &= G(E, y_1, y_2, \dots, y_{j-1}). \end{aligned}$$

在传统的编解码模型中,解码时每个时刻所用的上下文向量 E 都固定不变,这种构造没有将不同时刻所关注的信息不同的原理融入模型中^[24].研究者对该问题进行了更深入的探索,将图像识别中的注意力机制引入到序列问题中^[24],通过设计注意力机制与编码解码器结构相结合,提出了序列注意力机制的方法,具体为

$$\begin{aligned} E_j &= F(x_1, x_2, \dots, x_T, h, h'_{j-1}), \\ y_j &= G(E_j, y_1, y_2, \dots, y_{j-1}), \end{aligned} \quad (2)$$

其中, F 表示注意力机制与编码器部分结合的过程, h'_{j-1} 是解码器时刻 j 上一步的隐状态, h 为编码器的隐状态集合.与传统编解码模型不同的是,对于每一个预测时刻 j ,该编码器都得到关注信息不同的动态上下文向量 E_j ,从而让解码过程能够更加注重对于当前时间预测内容更加重要的历史信息.

3 预测模型

在上述介绍长短期记忆神经网络与注意力机制的基础上,本节详细展示本文提出的融合多时间序列市场动态预测模型的具体内容.首先介绍模型的框架,随后从多序列输入、注意力机制的设计以及优化函数 3 个方面进行详细介绍.

3.1 模型框架

本节通过对互联网金融市场动态的特性分析,设计并实现了一种基于多时间序列的注意力网络模型(multiple time series based attention LSTM, MALSTM).图 1 中箭头表示各个模块中输入和输出

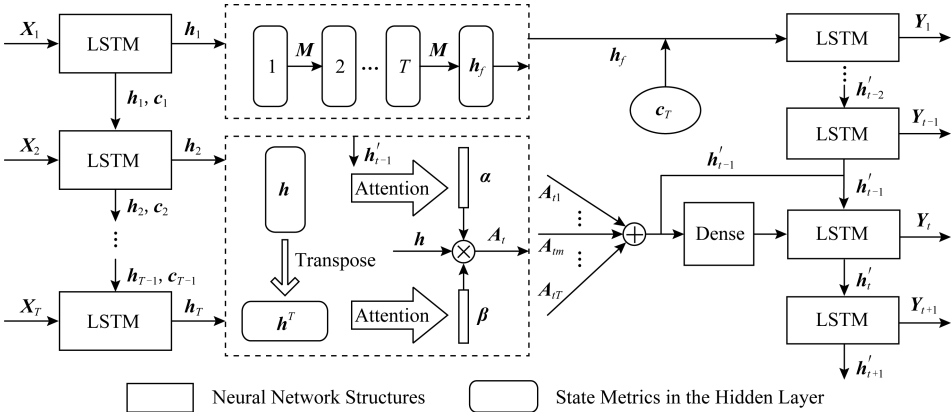


Fig. 1 Illustration of deep neural network prediction model based on multiple sequence attention mechanism
图 1 基于多序列注意力机制的深度神经网络预测模型示例

的对应的向量,LSTM 和全连接层等结构用带有标记的矩形框表示,隐含层状态矩阵用圆角矩形表示.具体地,模型首先提出一种基于多时间序列的输入模块,然后从时间和特征 2 个维度设计注意力机制;最后,模型设计了基于预测序列平稳性约束的优化函数,使得预测结果具有更好的稳健性.

3.2 多序列输入

时间序列研究中的传统方法对于时序的预测往往仅考虑目标变量时间序列自身的变化情况,没有考虑同时刻其他特征序列的相互影响.但是在互联网金融市场的动态分析中,其市场动态是会受到多个不同细分子类市场动态的共同影响;因此,本文对市场整体宏观动态以及多个类别市场的子序列同时进行建模,建立多类时间序列输入:

$$\mathbf{x}_t = [x_t^1, x_t^2, \dots, x_t^n], \tag{3}$$

其中, \mathbf{x}_t 表示时刻 t 模型的整体输入, x_t^c 表示当前时刻第 c 个子类序列的输入数据,例如 x_t^1 表示平台在时刻 t 交易量观察值;而 x_t^i 则表示某个类别(例如科技类或者稳健类项目)市场的交易量.接下来,如图 1 所示,模型通过一个单层 LSTM,建模多输入时间序列的动态演变过程.具体地,对于每一个时间步都有:

$$\mathbf{h}_t, \mathbf{c}_t = \text{LSTM}(\mathbf{h}_{t-1}, \mathbf{c}_{t-1}), \tag{4}$$

其中 $\mathbf{h}_t, \mathbf{c}_t$ 分别表示当前时刻 t 的市场隐状态以及记忆状态.因此,通过 LSTM 可以得到当前多序列的动态表征 \mathbf{h}_T .

3.3 注意力机制设计

通过多序列的 LSTM 输入建模后,可以得到每时刻市场动态状态 \mathbf{h}_t , 以及多序列输入的记状态 \mathbf{c}_t .考虑到金融市场的宏观动态建模需要前序时间段序列中的重要信息,因此模型需要学习不同时刻的信息对于长期变化预测所起的关键性作用;同时,考虑到宏观市场的组成,模型应能够自动挖掘对整体宏观序列预测影响最大的子类序列,并进一步增强对市场动态预测的能力.因此,本文分别从时间以及多序列特征维度设计 2 种注意力机制对市场动态的重要影响因素进行了建模.

首先,通过借鉴传统编解码模型设计思想,即不同时刻的隐状态被关注程度是不同的,在时间维度中构建基于市场历史状态的注意力机制.

依据 3.1 节中 LSTM 输入模块处理流程,可得每个时刻隐层状态所组成的输出矩阵 $\mathbf{h} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_T]$.模型进一步将 \mathbf{h} 作为编码器的输出,矩阵 \mathbf{h} 中的每个列向量 \mathbf{h}_t 表示时刻 t 的市场状态,并将其

作为注意力机制的输入.进而通过式(5)计算出当前时刻状态对预测状态 y_j 的重要程度:

$$\begin{aligned} e_{ij} &= \mathbf{V}_a^T \tanh(\mathbf{W}_a \mathbf{h}'_{j-1} + \mathbf{U}_a \mathbf{h}_i), \\ a_{ij} &= \text{softmax}(e_{ij}), \end{aligned} \tag{5}$$

其中, \mathbf{h}'_{j-1} 是解码器上一阶段的隐状态, \mathbf{h}_i 是编码器中时刻 t 的市场状态, $\mathbf{W}_a, \mathbf{U}_a$ 和 \mathbf{V}_a^T 分别是解码器隐状态、编码器状态以及注意力机制的参数矩阵, e_{ij} 表示编码器中时刻 t 的市场状态对于当前预测时刻 j 状态输出的影响程度.最后通过 softmax 函数,对 e_{ij} 进行归一化操作,从而获得每一个历史时刻市场状态对当前预测的权重因子 a_{ij} ,即在时间维度上的注意力值.

另外,考虑到市场宏观动态受到多个子类时序变化的影响,本文提出另一种基于多序列特征维度的注意力机制.不同于时间维度,多序列维度注意力机制需要考虑到每一个序列的历史编码状态对于市场完整宏观动态的影响.通过根据每一个序列在编码器序列的所有状态计算当前序列状态对宏观市场动态的影响程度,具体过程:

$$\begin{aligned} e'_k &= \mathbf{V}_\beta^T \tanh(\mathbf{U}_\beta \mathbf{h}_k^T), \\ \beta_k &= \text{softmax}(e'_k), \end{aligned} \tag{6}$$

其中, \mathbf{h}_k^T 表示第 k 个子序列的历史隐状态, \mathbf{U}_β 和 \mathbf{V}_β^T 是注意力机制的参数矩阵, β_k 即为当前序列对宏观市场的影响权重因子.考虑到解码器需要同时融合时间和特征维度的注意力机制,本文设计注意力权重线性联合方法,即可得到历史序列对时刻 j 市场预测的总权重因子 \mathbf{E}_j :

$$\mathbf{E}_j = \sum_{i=1}^T a_{ij} \sum_{k=1}^n \beta_k \mathbf{h}_i. \tag{7}$$

最后,通过利用 LSTM 结合注意力因子 \mathbf{E}_j 即可组成模型的解码器部分,从而可以逐步获取在时刻 j 的预测值 y'_j :

$$y'_j = \text{LSTM}(\mathbf{E}_j, \mathbf{h}'_{j-1}, \mathbf{c}'_{j-1}). \tag{8}$$

当 $j = 1$ 时, $y'_1 = \text{LSTM}(\mathbf{E}_1, \mathbf{h}_f, \mathbf{c}_T)$, 其中, $(\mathbf{h}_f, \mathbf{c}_T)$ 为历史序列的编码输出, \mathbf{h}_f 为宏观市场的隐状态, \mathbf{c}_T 表示历史序列最后的记忆状态.

3.4 优化函数设计

从宏观层面分析,相比于具体金融产品而言,金融市场的动态变化由于总需求与总供给的相对平衡而趋于平缓^[28].因此对于宏观市场的表征也应是趋于稳定平缓变化.本研究对于模型宏观市场的输出表征,提出一种线性演变约束过程,即采用式(9)的条件分布使模型对宏观动态编码过程中满足线性平稳约束:

$$\mathbf{h}_t | \mathbf{h}_{t-1} \sim N(\mathbf{M}\mathbf{h}_{t-1}, \boldsymbol{\Sigma}), \quad (9)$$

其中 \mathbf{M} 是状态转移矩阵, 其值在模型训练过程中进行优化, $\boldsymbol{\Sigma}$ 为协方差矩阵. 由此可以看出, 宏观市场的隐状态不再是直接由 LSTM 生成, 而是由市场的最终状态演化而来, 演化方式具体为

$$\mathbf{h}_f = \mathbf{M}\mathbf{h}_T, \quad (10)$$

其中, \mathbf{h}_f 表示的是宏观市场的隐状态, \mathbf{h}_T 表示历史信息的最末时间步状态. 为了进一步满足市场宏观动态的平稳特性, 我们设计一种基于历史隐状态线性演变过程的优化目标, 用于约束模型学习过程中对宏观市场的表征满足线性平稳特性. 具体优化目标为最小化 L_p :

$$L_p = \sum_{t=1}^{T-1} \|\mathbf{h}_{t+1} - \mathbf{M}\mathbf{h}_t\|_2^2 + \|\mathbf{h}_p - \mathbf{M}\mathbf{h}_T\|_2^2. \quad (11)$$

模型最终目的是通过对历史序列信息的学习预测未来 T' 时间的市场状态; 因此, 准确度是最重要的优化目标. 模型选择目前时序预测中最常见的均方根误差 (root mean squared error, RMSE)^[29,31] 作为模型准确性的优化函数, 具体如式 (12) 所示:

$$L_n = \sqrt{\frac{1}{T'} \sum_{j=1}^{T'} (y'_j - y_j)^2}. \quad (12)$$

进一步, 为了防止模型训练造成过拟合, 引入可调参数 λ 平衡模型在平稳性以及准确程度上的优化目标. 最终的优化函数:

$$L = \min \sum_{i=1}^N (L_n + \lambda L_p). \quad (13)$$

具体地, 模型使用初始学习率为 0.000 1 的 RMSProp 算法^[34] 优化模型参数直至模型收敛.

本节提出的模型提出解码器需要同时考虑时间和特征维度不同特征的重要程度, 从而设计了一种新的注意力结构. 并且模型考虑到从宏观层面分析, 金融市场的动态变化由于总需求与总供给的相对平衡会相对趋于平缓, 因此设计了线性平稳约束函数对预测过程进行约束. 这 2 点改进为本文模型最为重要的创新点.

4 实验评估

本节通过在真实数据集上构建实验验证所设计模型在互联网金融市场宏观动态预测问题上的有效性.

4.1 实验数据

本文实验数据全部来源于 Prosper 平台, 该平台是全球第二大互联网 P2P 借贷平台. 实验收集了平台自 2006-04-01 至 2011-05-25 平台有记录的总

计 1 622 d 的贷款人投资记录. 实验中按照 4:1 的比例将所有数据样本划分为训练集和测试集, 即 1 297 d 交易数据用于模型训练, 325 d 交易数据用于测试. 具体地, 实验中主要以日交易额和交易次数 2 种变量作为研究对象. 参照平台的用户信用评级, 将 7 种 (风险由高到低依次为: “HR”, “E”, “D”, “C”, “B”, “A”, “AA”) 不同风险评级的项目的动态序列提取出来, 构成宏观市场的子类序列作为模型的输入, 并按照滑动窗口的方法构建每个样本的历史序列和目标序列 (例如采用平台上 10 d 的历史数据预测未来 10 d 每天的目标变量).

考虑到平台采集的真实数据可能存在不完整、不一致的特性, 本文对训练集以及测试集进行了预处理: 1) 首先使用线性插值法处理数据集中的缺失值. 使用线性插值是因为在时间序列问题中对于缺失值不能做简单的删除操作, 它会导致时间序列索引的断裂. 2) 对原始时间序列进行数据标准化, 不同信用等级时间序列的变量数值大小差异较大, 数值较大的序列对模型训练和预测结果的影响会有较大干扰, 直接输入到模型中可能会导致模型的权重偏移. 实验中采用数据归一化的方法将所有的时间序列变量映射到 $[0, 1]$ 区间内:

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \quad (14)$$

其中, X 为序列变量的原始数据, X_{\max} , X_{\min} 分别是原始序列变量中的最大值和最小值, X^* 为经过数据归一化的值, 其取值范围为 $[0, 1]$.

4.2 实验设置

本节分别介绍实验平台、对比方法和评测指标.

4.2.1 实验平台

实验中本文所提出的 MALSTM 模型以及全部对比方法均基于 python 3.5 以及 tensorflow1.2.1 实现, 并且涉及到 LSTM 结构的所有模型中神经元个数设置为 200, Dropout 为 0.5, $\lambda = 0.001$, Batchsize 大小设置为 8. 本文所有方法程序均在配置 2 块 2.20 GHz Intel Xeon E5-2650 v4 CPU 以及 4 块 Tesla K80 GPU 的 Linux 环境下运行.

4.2.2 对比方法

为了对比验证本文所提出模型的效果, 共选取和设计了 4 种方法作为对比实验.

1) 自回归差分移动平均模型 (ARIMA). 经典的传统时间序列预测模型, 输入为单时间序列, 只包括平台总金额或总贷款笔数.

2) 单序列输入的 LSTM 模型^[12] (记为 LSTM).

在循环网络模型中使用 LSTM 构造编码器和解码器,每个时间步的输入只包括每天的平台总金额或总交易笔数.

3) 多序列输入的 LSTM 模型(记为 LSTM-M),使用 LSTM 构造编码器和解码器,参考本文的多序列输入机制,每个时间步的输入除了当前每天的平台总金额数或总交易笔数之外,还包含不同风险等级项目每天的总金额数或总交易笔数.此模型也看作是传统模型的一个改进.

4) 多序列注意力机制模型(记为 MALSTM-t).在 LSTM-M 模型的基础上增加了时间纬度上的注意力机制.

5) 多序列注意力机制模型(记为 MALSTM),即本文所提出的基于多序列输入的注意力模型. MALSTM 在训练中只使用 RMSE 作为优化目标.

6) 加入平稳性优化目标的模型(MALSTM-L).在 MALSTM 模型中加入了本文提出的线性平稳约束对模型进行优化,并使用本文设计的完整优化函数.

4.2.3 评测指标

本文使用在时间序列预测领域中最为广泛采用

的均方根误差(root mean squard error, $RMSE$)^[27,29]以及平均绝对误差(mean absolute error, MAE)^[29]作为预测精准度评价标准,两者基本原理是计算预测序列的目标值与预测值之间的差异性大小, $RMSE$ 和 MAE 值越小代表预测精准度越高, $RMSE$ 和 MAE 计算方法为

$$RMSE = \sqrt{\frac{1}{T'} \sum_{j=1}^{T'} (y'_j - y_j)^2},$$
$$MAE = \frac{1}{T'} \sum_{j=1}^{T'} |y'_j - y_j|.$$

(15)

4.3 实验结果与分析

本文分别对平台的总贷款金额和总贷款笔数进行预测对比实验,固定历史天数为 10 的情况下,将预测天数调整为 1~10,观察不同模型的表现情况.

表 1 和表 2 分别给出了 4 种算法在预测天数 1~10 d 中平台总交易金额预测结果的 $RMSE$ 和 MAE 值.从表中可以很清晰地观察到本文提出的模型(MALSTM 与 MALSTM-L)显著优于其他对比模型.其中,在 $RMSE$ 指标上,本文提出的 MALSTM 与 MALSTM-L 相较于经典注意力架构 MALSTM-t 分别取得了最大 2.04%和 4.18%的效果提升,相对

Table 1 The RMSE Results of the Market Amount Prediction in the Next Ten Days

表 1 对未来 10 d 市场的交易总额预测结果的均方根误差

Method	Days									
	1	2	3	4	5	6	7	8	9	10
ARIMA	0.064 0	0.063 3	0.064 0	0.064 6	0.065 9	0.065 6	0.066 4	0.067 5	0.068 8	0.070 7
LSTM	0.062 7	0.063 3	0.063 7	0.064 1	0.065 1	0.065 5	0.067 2	0.068 0	0.069 5	0.071 8
MLSTM	0.062 4	0.062 8	0.063 1	0.062 9	0.063 6	0.064 2	0.064 6	0.065 3	0.066 0	0.066 9
MALSTM-t	0.058 9	0.059 3	0.060 2	0.060 7	0.061 1	0.061 8	0.062 2	0.062 6	0.063 0	0.063 7
MALSTM	0.059 2	0.059 4	0.059 7	0.060 1	0.060 5	0.060 9	0.061 6	0.061 4	0.061 8	0.062 4
MALSTM-L	0.058 3	0.058 4	0.058 6	0.059 1	0.059 4	0.059 6	0.059 6	0.060 0	0.060 6	0.061 1

Note: The best results are in bold.

Table 2 The MAE Results of the Market Amount Prediction in the Next Ten Days

表 2 对未来 10 d 市场的交易总额预测结果的平均绝对误差

Method	Days									
	1	2	3	4	5	6	7	8	9	10
ARIMA	0.044 8	0.045 5	0.04 6	0.046 9	0.048 8	0.048 9	0.049 6	0.050 7	0.052 4	0.053 6
LSTM	0.044 1	0.045 3	0.045 8	0.046 5	0.048 1	0.048 7	0.050 2	0.051 3	0.053 2	0.055 4
MLSTM	0.043 7	0.044 5	0.044 9	0.044 6	0.045 2	0.046 2	0.046 9	0.048 0	0.049 1	0.049 7
MALSTM-t	0.041 6	0.04 2	0.042 8	0.043 0	0.043 9	0.044 4	0.044 7	0.045 0	0.045 3	0.046 2
MALSTM	0.041 3	0.041 7	0.042 2	0.042 7	0.043 2	0.043 8	0.044 9	0.044 2	0.044 7	0.045 9
MALSTM-L	0.040 6	0.040 9	0.041 2	0.042 0	0.042 6	0.043 1	0.042 9	0.043 2	0.043 8	0.044 2

Note: The best results are in bold.

于未使用注意力机制的其他模型取得了 4.45%~6.72%以及 6.04%~8.67%的提升.同时,在 MAE 指标上,MALSTM 与 MALSTM-L 相较于经典注意力架构 MALSTM-t 分别取得了最大 1.78%和 4.33%的效果提升,相对于未使用注意力机制的其他模型取得了 4.26%~8.96%以及 5.82%~11.06%的提升.这些结果说明了本文提出的多输入注意力深度网络模型的有效性与先进性.同时,通过比较本文提出的 2 个模型,可以发现 MALSTM-L 相比于 MALSTM 在 RMSE 指标和 MAE 指标上,分别平均提升了 2.02%以及 2.52%,这说明了模型基于的平稳性约束条件的合理性和有效性,并证明了市场宏观的动态变化是趋于平稳的.

进一步,通过对预测结果按照时间维度的横向分析,可以发现随着预测时间的推移,传统的 LSTM 与 MLSTM 模型预测效果显著减低,例如,2 个模型在第 10 天预测结果相比较于第一天准确度在 RMSE 指标下分别降低了 14.51%与 7.21%.由此可以看

出,对于长期预测问题,多序列输入模型显著优于直接输入目标单序列的模型,这证明了多序列之间的相互关系能够有效地辅助市场的宏观动态建模,并提高模型对于远期预测结果的准确性.同时,对于 MALSTM 和 MALSTM-L 模型,远期的预测结果在 RMSE 指标下仅降低了 5.41%以及 4.80%,这表明本文提出的模型能够很好的建模金融市场宏观动态的长期波动情况,同时也说明在市场宏观预测问题中,平稳性约束的有效性,进一步证明了本文提出的模型在金融宏观市场预测问题中的稳健性.

图 2 展示的是 4 种算法在预测 1~10 d 中平台总交易笔数预测结果的 RMSE 和 MAE 值.从图 2 可以清晰看出,本文提出的带有平稳性约束的 MALSTM-L 模型效果最佳,其次是多序列注意力模型 MALSTM,然后是多序列输入的 MALSTM-t 和 MLSTM 模型,效果最差的是 ARIMA 模型和原始的 LSTM 模型.结果进一步表明本文所提出的多序列对市场宏观动态预测的重要性,同时也证明了基于多序列注意力

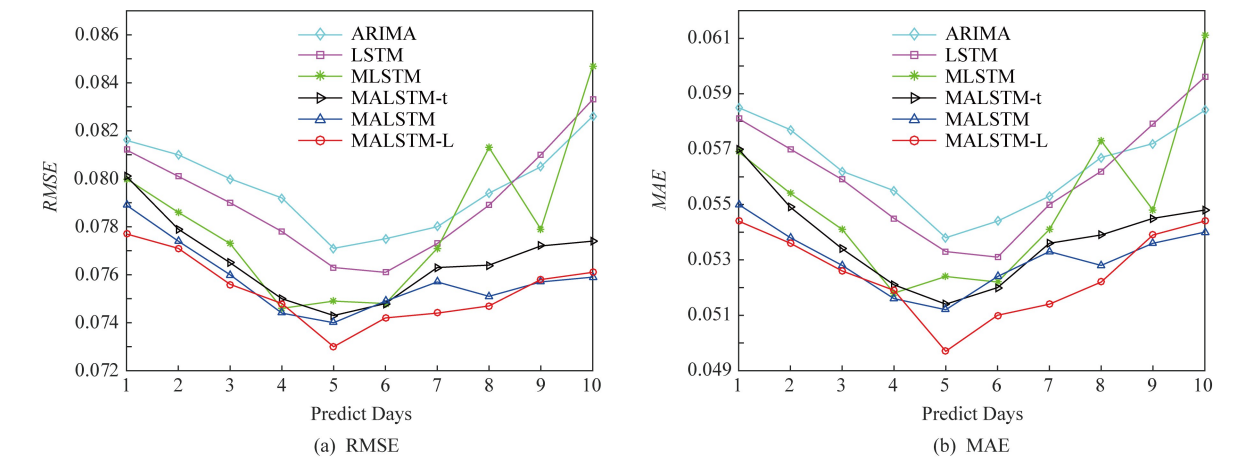


Fig. 2 The results of the market transaction quantity prediction in the next ten days

图 2 模型对未来 10 d 金融市场的交易单数预测结果

Table 3 The RMSE Results of the Market Counts Prediction in the Next Ten Days

表 3 对未来 10 d 市场的交易笔数预测结果的均方根误差

Method	Days									
	1	2	3	4	5	6	7	8	9	10
ARIMA	0.0816	0.0810	0.0800	0.0792	0.0771	0.0775	0.0780	0.0794	0.0805	0.0826
LSTM	0.0812	0.0801	0.0790	0.0778	0.0763	0.0761	0.0773	0.0789	0.0810	0.0833
MLSTM	0.0800	0.0786	0.0773	0.0746	0.0749	0.0748	0.0771	0.0813	0.0779	0.0847
MALSTM-t	0.0801	0.0779	0.0765	0.0750	0.0743	0.0748	0.0763	0.0764	0.0772	0.0774
MALSTM	0.0789	0.0774	0.0760	0.0744	0.0740	0.0749	0.0757	0.0751	0.0757	0.0759
MALSTM-L	0.0777	0.0771	0.0756	0.0748	0.0730	0.0742	0.0744	0.0747	0.0758	0.0761

Note: The best results are in bold.

Table 4 The MAE Results of the Market Counts Prediction in the Next Ten Days
表 4 对未来 10 d 市场的交易笔数预测结果的平均绝对误差

Method	Days									
	1	2	3	4	5	6	7	8	9	10
ARIMA	0.058 5	0.057 7	0.056 2	0.055 5	0.053 8	0.054 4	0.055 3	0.056 7	0.057 2	0.058 4
LSTM	0.058 1	0.057 0	0.055 9	0.054 5	0.053 3	0.053 1	0.055 0	0.056 2	0.057 9	0.059 6
MLSTM	0.056 9	0.055 4	0.054 1	0.051 8	0.052 4	0.052 2	0.054 1	0.057 3	0.054 8	0.061 1
MALSTM-t	0.057 0	0.054 9	0.053 4	0.052 1	0.051 4	0.052 0	0.053 6	0.053 9	0.054 5	0.054 8
MALSTM	0.055 0	0.053 8	0.052 8	0.051 6	0.051 2	0.052 4	0.053 3	0.052 8	0.053 6	0.054 0
MALSTM-L	0.054 4	0.053 6	0.052 6	0.051 9	0.049 7	0.051 0	0.051 4	0.052 2	0.053 9	0.054 4

Note: The best results are in bold.

机制的有效性以及本文所设计模型的先进性.同时,图 2 中结果显示序列预测模型的准确度随着预测时间的推移表现出先上升后下降.同时,对比 4 种方法的预测误差的变化趋势可以发现,MALSTM-L 模型的误差随着时间的推移变化幅度最小,这进一步验证本文提出的模型在宏观市场预测问题中的稳健性,也说明了引入的市场平稳性约束的有效性.另外,相比较而言,由于交易笔数序列比交易金额序列平稳性差,所以,在表 3 和表 4 的结果中,MALSTM-L 有个别情况时表现弱于 MALSTM.

5 总 结

本文针对互联网金融市场的宏观动态预测进行研究,首先介绍了互联网金融市场时间序列动态的特点与预测问题的研究背景.然后回顾了传统时间序列预测方法,基于深度学习的方法和注意力机制,并对当前金融市场时序预测研究进行了简要介绍.针对互联网金融市场的流动性强、易变性高的特点,本文提出了一种基于深度神经网络融合层次多时间序列学习的预测模型.首先,该模型可以实现对多序列(宏观动态序列和多种子序列)特征变量输入的处理,并且在时间和序列特征 2 个维度上利用注意力机制融合输入变量;其次,模型设计了基于预测序列平稳性约束的优化函数,使得模型预测具有更好的稳健性;最后,在真实大规模互联网金融数据集上进行了大量实验,实验结果表明本文所提出的基于深度神经网络融合层次多时间序列学习的预测模型取得了最佳的预测性能,充分证明了模型在互联网金融市场宏观动态预测问题上的有效性与稳健性.

本文探索了互联网金融市场的宏观动态预测问题,并建模了多序列输入对于市场动态的影响,同时

利用了市场宏观动态的平稳性特点,提升了时序预测准确度,为时间序列预测提供了一种新的研究思路.未来的研究可以从 2 个方面展开:1)进一步探索子序列对于宏观动态的影响以及各个子序列之间的相互影响关系;2)探索外部信息,例如新闻媒体文本等对市场宏观动态预测的影响.

参 考 文 献

[1] Zhao Hongke. Data mining and its applications for micro entities in Internet finance [D]. Hefei: University of Science and Technology of China, 2018
(赵洪科. 面向互联网金融微观对象的数据挖掘方法及应用研究[D]. 合肥: 中国科学技术大学, 2018)

[2] Belleflamme P, Lambert T, Schwienbacher A. Crowdfunding: Tapping the right crowd [J]. Journal of Business Venturing, 2014, 29(5): 585-609

[3] Zhao Hongke, Ge Yong, Liu Qi, et al. P2P lending survey: Platforms, recent advances and prospects [J]. ACM Transactions on Intelligent Systems and Technology, 2017 (6): 72

[4] Zhang Yu, Lü Bing, Chen Chunyan, et al. Product multi-dimensional evaluation and global recommendation in P2P lending market [J]. Journal of Chinese Computer Systems, 2016, 37(2): 308-311 (in Chinese)
(张钰, 吕冰, 陈春燕, 等. 网络小额贷款平台产品多维度分析与全局推荐[J]. 小型微型计算机系统, 2016, 37(2): 308-311)

[5] Haewon Y, Byungtae L, Myungsin C. From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms [J]. Electronic Commerce Research and Applications, 2012, 1(6): 469-483

[6] Akaike H. Fitting autoregressive models for prediction [J]. Annals of the Institute of Statistical Mathematics, 1969, 21 (1): 243-247

- [7] Feng Bo, Ye Qiwen, Chen Dongyu. Review on P2P online lending and new research opportunities for China's case [J]. *Journal of Management Sciences in China*, 2017, 20(4): 113-126 (in Chinese)
(冯博, 叶绮文, 陈冬宇. P2P 网络借贷研究进展及中国问题研究展望[J]. *管理科学学报*, 2017, 20(4): 113-126)
- [8] Luo Min, Zhou Mengya. Bottleneck analysis of independent development of crowdsourcing Websites [J]. *China Newspaper Industry*, 2015 (20): 26-27 (in Chinese)
(罗敏, 周梦雅. 众筹网站独立发展瓶颈分析[J]. *中国报业*, 2015 (20): 26-27)
- [9] Sims C A. Macroeconomics and reality [J]. *Econometrica: Journal of the Econometric Society*, 1980, 48(1): 1-48
- [10] Johansen S. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models [J]. *Econometrica: Journal of the Econometric Society*, 1991, 59(6): 1551-1580
- [11] Hopfield J J. Neural networks and physical systems with emergent collective computational abilities [J]. *Proceedings of the National Academy of Sciences*, 1982, 79(8): 2554-2558
- [12] Hochreiter S, Schmidhuber J. Long short-term memory [J]. *Neural Computation*, 1997, 9(8): 1735-1780
- [13] Zhang Yi, Guo Quan, Wang Jianyong. Big data analysis using neural networks [J]. *Journal of Sichuan University (Engineering Science Edition)*, 2017, 49(1): 9-18 (in Chinese)
(章毅, 郭泉, 王建勇. 大数据分析的神经网络方法[J]. *工程科学与技术*, 2017, 49(1): 9-18)
- [14] Hu Ziniu, Liu Weiqing, Bian Jiang, et al. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction [C] //Proc of the 11th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2018: 261-269
- [15] Wu Bo, Cheng Wenhua, Zhang Yongdong, et al. Sequential prediction of social media popularity with deep temporal context networks [C] //Proc of the 26th Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2017: 3062-3068
- [16] Cinar Y G, Mirisae H, Goswami P, et al. Position-based content attention for time series forecasting with sequence-to-sequence RNNs [C] //Proc of the 31st Int Conf on Neural Information Processing. Berlin: Springer, 2017: 533-544
- [17] Tsay R S. *Analysis of Financial Time Series* [M]. New York: John Wiley & Sons, 2005
- [18] Cleveland R B, Cleveland W S, McRae J E, et al. STL: A seasonal-trend decomposition [J]. *Journal of Official Statistics*, 1990, 6.1: 3-73
- [19] Das, Samarjit. *Time Series Analysis* [M]. Princeton, NJ: Princeton University Press, 1994
- [20] Montgomery D C, Johnson L A, Gardiner J S. *Forecasting and Time Series Analysis* [M]. New York: McGraw-Hill, 1990
- [21] Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation [J]. *Econometrica: Journal of the Econometric Society*, 1982, 50(4): 987-1007
- [22] Bollerslev T. Generalized autoregressive conditional heteroskedasticity [J]. *Journal of Econometrics*, 1986, 31(3): 307-327
- [23] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. *arXiv preprint arXiv:1409.0473*, 2014
- [24] Kazem A, Sharifi E, Hussain F K, et al. Support vector regression with chaos-based firefly algorithm for stock market price forecasting [J]. *Applied Soft Computing*, 2013, 13(2): 947-958
- [25] Azoff E M. *Neural Network Time Series Forecasting of Financial Markets* [M]. New York: John Wiley & Sons, Inc, 1994
- [26] Sapankevych N I, Sankar R. Time series prediction using support vector machines: A survey [J]. *IEEE Computational Intelligence Magazine*, 2009, 4(2): 24-38
- [27] Wang Weina, Pedrycz W, Liu Xiaodong. Time series long-term forecasting model based on information granules and fuzzy clustering [J]. *Engineering Applications of Artificial Intelligence*, 2015, 41(C): 17-24
- [28] Tobin J. *Essays in Economics: Macroeconomics* [M]. Cambridge, MA: MIT Press, 1987
- [29] Zhao Hongke, Liu Qi, Zhu Hengshu, et al. A sequential approach to market state modeling and analysis in online P2P lending [J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48, 2018, (1): 21-33
- [30] Zhao Hongke, Zhang Hefu, Ge Yong, et al. Tracking the dynamics in crowdfunding [C] //Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 625-634
- [31] Jin Binbin, Zhao Hongke, Chen Enhong, et al. Estimating the days to success of campaigns in crowdfunding: A deep survival perspective [C] //Proc of the 28th Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2019: 1-8
- [32] Ren Xiaoying, Xu Linli, Zhao Tianxiang, et al. Tracking and forecasting dynamics in crowdfunding: A basis-synthesis approach [C] //Proc of the 18th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2018: 1212-1217
- [33] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. *arXiv preprint arXiv:1406.1078*, 2014
- [34] Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude [J]. *COURSERA: Neural Networks for Machine Learning*, 2012, 4(2): 26-31



Zhao Hongke, born in 1988. PhD, lecturer. His main research interests include data mining and Internet finance.



Zhang Xi, born in 1982. PhD, professor, PhD supervisor. His main research interests include knowledge management and social network.



Wu Likang, born in 1996. Master candidate. His main research interests include deep learning, data mining.



Liu Qi, born in 1986. PhD, associate professor. Member of CCF. His main research interests include data mining, optimization, recommender systems, and social network analysis.



Li Zhi, born in 1993. PhD candidate. His main research interests include deep learning, data mining and recommender systems.



Chen Enhong, born in 1968. PhD, professor, PhD supervisor. CCF Fellow, Senior member of IEEE. His main research interests include data mining, artificial intelligence, recommender systems and social network analysis.

《计算机研究与发展》征订启事

《计算机研究与发展》(Journal of Computer Research and Development)是中国科学院计算技术研究所和中国计算机学会联合主办、科学出版社出版的学术性刊物,中国计算机学会会刊.主要刊登计算机科学技术领域高水平的学术论文、最新科研成果和重大应用成果.读者对象为从事计算机研究与开发的研究人员、工程技术人员、各大专院校计算机相关专业的师生以及高新企业研发人员等.

《计算机研究与发展》于 1958 年创刊,是我国第一个计算机刊物,现已成为我国计算机领域权威性的学术期刊之一.并历次被评为我国计算机类核心期刊,多次被评为“中国百种杰出学术期刊”.此外,还被《中国学术期刊文摘》、《中国科学引文索引》、“中国科学引文数据库”、“中国科技论文统计源数据库”、美国工程索引(Ei)检索系统、日本《科学技术文献速报》、俄罗斯《文摘杂志》、英国《科学文摘》(SA)等国内外重要检索机构收录.

国内邮发代号:2-654;国外发行代号:M603

国内统一连续出版物号:CN11-1777/TP

国际标准连续出版物号:ISSN1000-1239

联系方式:

100190 北京中关村科学院南路 6 号《计算机研究与发展》编辑部

电话: +86(10)62620696(兼传真); +86(10)62600350

Email: crad@ict.ac.cn

http://crad.ict.ac.cn