基于专利数据应用LDA和N-BEATS组合方法的技术主题预测研究*

吴雷 杜文研 林超然
（哈尔滨工程大学经济管理学院，哈尔滨 150001）

摘要：预测技术主题未来热点，有助于企业在技术层面判别现状、识别未来技术方向并提前规划战略布局。提出LDA和N-BEATS组合方法，运用LDA模型提取专利文献数据的技术主题，引入N-BEATS网络模型分析各技术主题专利数量的时间序列，发挥其分析可解释性时间序列的优势，在预测模型中加入技术研发活动周期性模块，并以芯片技术为例，运用该组合方法预测产业的技术主题和未来趋势。对比实验中LDA和N-BEATS组合方法的预测精度高于LDA-LSTM、IPC-N-BEATS和IPC-LSTM三种基准方法。案例结果表明，未来芯片产业研发热点是电子级树脂、蚀刻机、芯片封装、芯片键合、抛光液。

关键词：LDA；N-BEATS网络模型；深度学习；芯片产业；技术预测
中图分类号：G255 DOI：10.3772/j.issn.1673-2286.2023.11.008
引文格式：吴雷，杜文研，林超然. 基于专利数据应用LDA和N-BEATS组合方法的技术主题预测研究[J]. 数字图书馆论坛，2023（11）：62-73.

当今世界正面临新一轮科技革命和产业变革，新兴技术不断迭代。技术主题是技术文献的主旨和核心。通过掌握并预测技术主题未来热点，企业能在技术层面判别现状、洞察趋势、识别未来技术方向以及提前规划战略布局[1]。在技术主题的研究中，专利文献涵盖了超过90%的行业技术信息[2]，具有长期性和超前性[3]。如今，专利成果数量与日俱增、技术迭代速度不断提高，如何从海量文本中挖掘技术主题、提高预测技术主题未来趋势的精准度，是当前亟需解决的问题之一。

本文采用专利文献作为技术主题的数据来源，运用隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）算法提取专利文献数据的技术主题；引入神经网络底层扩展分析（Neural Basis Expansion Analysis，N-BEATS）模型，实现对周期性、趋势性等时序变化的解释，分析各技术主题的专利数量时间序列；以芯片技术为例展开实证分析，验证方法的有效性。相比其他专利挖掘预测算法，LDA和N-BEATS网络模型组合算法在抗干扰的前提下，考虑了技术活动的周期性，提高了技术主题预测精度，为国家提前规划战略布局和企业洞察未来技术趋势提供新思路。

1 文献综述

为挖掘专利文献中的技术主题，早期研究通过分析代表性专利获得技术发展现状[4-5]。后为减少对专家经验的依赖，以及弥补专利代表性难以衡量等缺陷，学者采用专利的自身属性划分技术主题，这些属性包括专利的IPC分类号[6]、德温特手工代码[7]等。随着算法的发展，为了提高技术主题提取的精度，专利聚类技术逐渐应用于技术主题研究，学者利用专利申请人[8]、专利共现网络[9]、引用关系[10]、SAO语义分析[11]、TF-IDF模型[12]、LDA模型[13]等方法实现聚类。其中，LDA模型采用贝叶

收稿日期：2023-08-14
*本研究得到国家社会科学基金一般项目“‘双链融合’视角下IC产业关键核心技术甄别及突破研究”（编号：23BGL076）资助。

斯模型、多项式分布等多种方法，实现了完全通过所给的文本信息分析的效果，避免了人工分析方法的不可重复性，该方法目前已广泛应用于对专利文本的提取[13]。例如，罗恺等[14]运用LDA算法提取关节机器人专利主题，分析关节机器人技术融合趋势。马建红等[15]基于产品生命周期，利用LDA模型提取不同阶段的产品专利技术主题。Choi等[16]运用LDA模型识别专利背后的物流相关技术主题并进一步调查领域层面和公司层面的趋势。

在专利技术主题预测分析中，学者运用专利文献的时间信息，采用时间序列分析[17]、灰色预测[18]、技术路线图[19]等方法获取相关技术的发展趋势。其中，随着数据分析方法的快速发展，时间序列分析法逐渐成为主流的预测方法[20]。早期的时间序列模型主要应用于相关病症传染率预测[21]、旅游城市的旅游需求预测[22]、原油市场的价格预测[23]以及国内粮食产量的预测 [24]等。21世纪初期，学者结合时间序列模型预测分析从专利文本中提取出的技术主题，使用有限元方法模拟[25]、曲线拟合[26]、差分整合移动平均自回归模型[27-28]、支持向量机算法[29]等方法挖掘专利文献的数据规律并预测后续的时序变化。后来学者发现建模假设、突发事件等因素会显著干扰这类模型的预测结果[30]，为了解决这一问题，学者们引入机器学习的方法。Lu等[31]、Zhang等[32]运用HMM模型预测相关技术主题的未来趋势，随后学者们进一步引入深度学习模型，如通过LSTM模型[33-34]、CNN模型[35]、RNN模型[36]、GRU模型[37]等实现了基于自然语言处理的技术预测。

综上所述，针对技术主题预测的研究较多，主要应用时间序列模型和深度学习模型。时间序列模型能够灵活预测时间序列趋势变化点，但由于其表达能力较低，可能无法准确预测复杂模型。相比时间序列模型，深度学习模型的抗干扰性能较高，但常见的深度学习模型作为黑盒模型，存在无法拆解时序变化因素、难以描述模型的内部结构两个缺陷[38-39]。同时，企业技术研发活动受宏观经济周期[40-41]及技术迭代的影响，显现周期性，例如英特尔公司交替更新制程和微架构，使得技术研发产生明显的“Tick-tock”周期[42]。但现有的时间序列模型和深度学习模型都尚未考虑到技术活动的周期性，会高估周期低谷的研发活动，这影响了预测结果的准确性。

为解决上述问题，本文提出了结合LDA算法和N-BEATS模型的组合方法。N-BEATS网络模型将深度学习同经典的时间序列预测方法结合，是一种基于反向和正向剩余链接以及深度全连接层堆栈的神经结构[43]，目前已广泛应用于风力发电[44]、空间科学[45]、心电图信号[46]等众多领域的数据预测，尤其在大型时间序列数据分析中展现出了较高的预测能力。因此，本文运用LDA算法提取专利文献数据的技术主题，并引入N-BEATS网络模型，通过构造不同的堆栈模型实现对周期性、趋势性等时序变化的解释，分析各技术主题的专利数量时间序列。同时，以芯片技术为例展开实证分析，对方法的有效性进行验证。

2 研究方法

将LDA算法和N-BEATS模型组合。在智慧芽全球专利数据库中检索专利数据，形成芯片产业专利数据库，经过数据清洗和专利文本向量化后，运用LDA算法将文本分类并构建相应的技术主题，并采用困惑度算法确定分类的主题数，再将各技术主题划分为训练集和测试集。通过构建N-BEATS网络模型的不同堆栈，实现对技术活动周期性、趋势性等时序变化的解释，训练N-BEATS网络模型并测试模型训练结果，经过多次迭代后得到最佳的N-BEATS网络模型，完成对各技术主题的时序预测。研究思路见图1。

2.1 LDA算法提取技术主题

LDA模型在传统的概率潜在语义分析（Probabilistic Latent Semantic Analysis，PLSA）模型的基础上引入了参数的先验分布概念。在LDA模型中，每个文档以及每个话题在所有单词上的概率分布都被赋予先验分布，因此LDA模型能够比PLSA更好地刻画文档、主题、单词这三者的关系。假定专利文献的技术主题服从超参数狄利克雷先验分布，见式（1）。

$$\mathrm{Dir}\left(\theta_d \mid \alpha\right)=\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{K}\theta_{dk}^{\alpha_k-1} \tag{1}$$

式中：θ_{dk}表示专利d在技术主题k中的分布；K表示技术主题总数；α表示每篇专利文献中主题的狄利克雷分布的超参数。对每一个技术主题k生成主题词项分布ϕ_k~Dir（β），β表示每个主题下词汇的狄利克雷分布的超参数。对专利d生成主题词分布θ_d~Dir（α），并基于文本中第n个词项生成主题项z_{dn}~Multinomial（θ_d）和词项w_{dn}~Multinomial（$\phi_{z_{dn}}$）。LDA模型见式（2）。

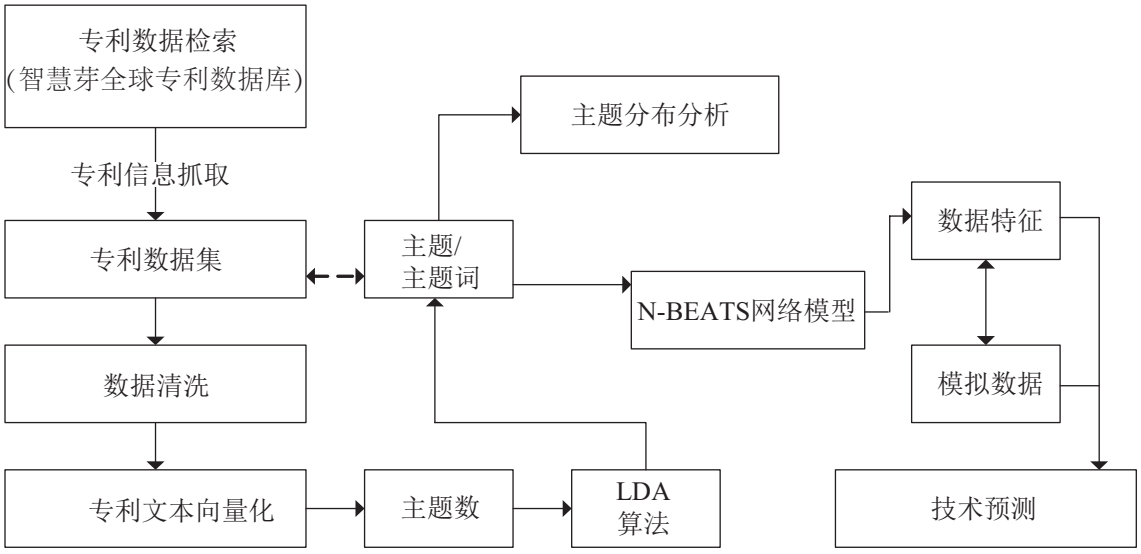

图1 研究思路

$$p(W \mid \alpha, \beta) = \prod_{d=1}^{D} \int p(\theta_d \mid \alpha) \prod_{n=1}^{N_d} \sum_{z_d} p(z_{dn} \mid \theta_d)\ p(w_{dn} \mid \phi_{z_{dn}})\ d\theta_d \quad (2)$$

此外，要使LDA主题模型发挥降维作用，关键在于对异构文本潜在主题数量的准确设定，但LDA方法自身并不能生成最佳的主题数量。因此，选用Blei等[13]提出的困惑度（Perplexity）作为确定主题数量的标准。

2.2 N-BEATS网络模型预测技术主题

N-BEATS网络模型的框架结构见图2。首先根据窗口大小输入原始时间序列数据，其次通过若干堆栈提取原始数据的多个特征，其中每个堆栈含有多个由残差模块联结的区块，每个区块内存在若干计算神经元。

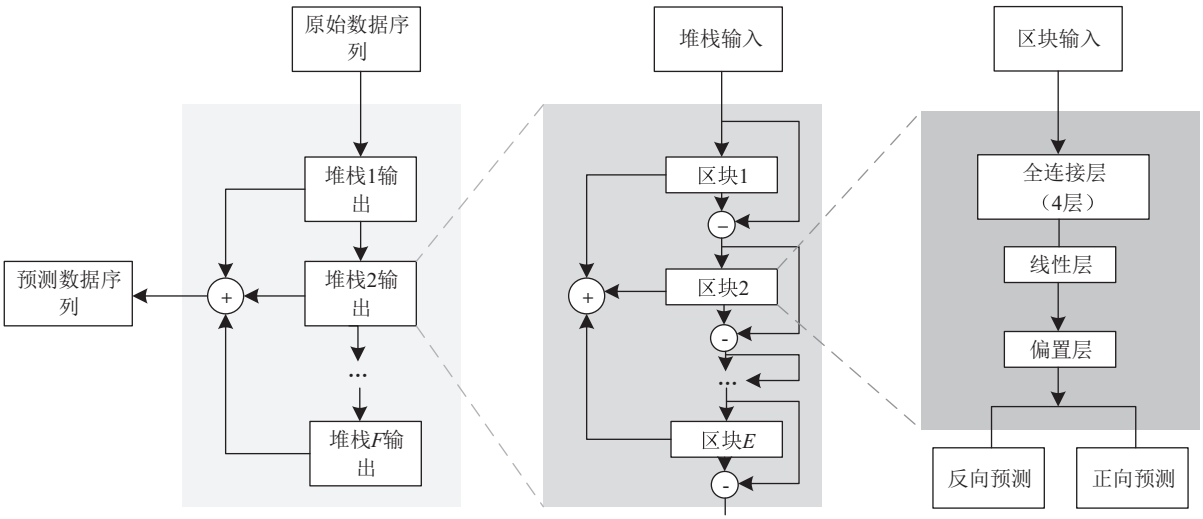

图2 N-BEATS网络模型架构

（1）区块层分解正向反向预测结果。时间序列数据输入后经过4层全连接层后进入线性层，在线性层经过线性投影后生成正向预测系数和反向预测系数，接下来在偏置层对两个系数采取归纳偏差以适当地限制输出结构，最终生成正向预测结果和反向预测结果。其中正向预测结果是保留下来对预测任务起正向作用的信息，反向预测结果是从该时间序列数据剔除出的偏置误差。

（2）堆栈层提取时间序列数据特征。针对区块层区分的正向反向预测结果，该模块采取了一种新的分层

双残差拓扑结构，将区块层正向预测结果保留，反向预测结果转化为残差进入下一个区块层，最终堆栈层的输出结果为每个区块层的正向预测结果的整合。由于每个区块层之间由残差连接，这种结构不但提高下游区块的预测效率，还促进了梯度的反向传播。

（3）整合时间序列特征预测分量。将每个堆栈层输出的预测结果按照各个堆栈设定的参数维度叠加整合在一起，得到最终的预测数据序列。

模型设置了3个堆栈，分别用来处理周期的变化。每个堆栈内包含2个区块，区块内包含4层全连接层、256个隐藏层单元。区块层中的线性层和偏置层在处理周期变化、趋势变化和其他形变因素变化的堆栈模型中参数维度设置分别为3、2以及1。模型训练过程选用16个样本点的小批量梯度下降法。为了提高模型的预测精度，运用优化器减小模型预测值与实际值的差异。N-BEATS网络预测模型运用均方根误差作为误差函数，即预测数据和实际数据对应点误差的平方和的算术平方根，其均值优化器采用了学习率为0.001、一阶矩阵和二阶矩阵估计指数衰减率分别为0.900和0.999、Epsilon参数为10^{-8}的Adam优化算法。Adam优化算法能基于训练数据迭代地更新神经网络权重。同其他优化算法相比，Adam算法更适合解决数据和参数方面的问题和处理非平稳的时间序列。训练每个单个样本的迭代次数为300次。通过多次试验对N-BEATS网络模型的推荐参数调整后获得该模型使用的参数。N-BEATS网络模型的参数设置见表1，模型的结构见图3。

表1　N-BEATS网络模型参数设置

参　数	设　定
堆栈数	3
每个堆栈的区块数	2
每个区块的全连接层数	4
隐藏层单元数	256
周期堆栈参数维度设置	3
趋势堆栈参数维度设置	2
其他因素堆栈参数维度设置	1
损失函数	均方根误差
优化器	Adam算法，学习率为0.001
迭代次数	300
小批量梯度下降的样本点数	16

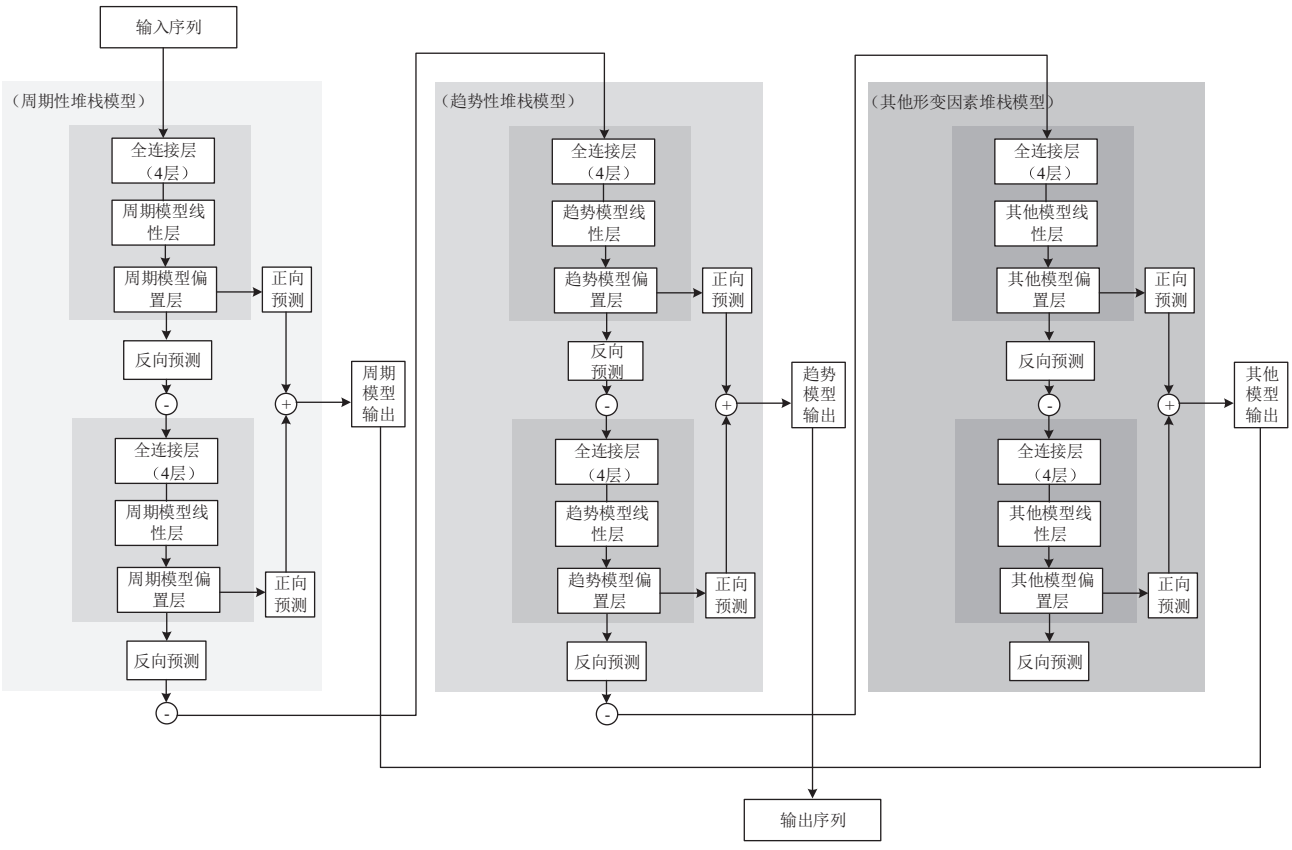

图3　N-BEATS预测网络模型

3 案例分析

本节通过分析芯片产业的技术主题预测情况验证LDA和N-BEATS网络模型组合预测方法的可行性和有效性。目前芯片产业是支撑国民经济发展的战略性和先导性产业，是应对结构转型升级的根本产业[47]。芯片产业既是保障我国战略安全的重要屏障，也是当前国家经济、科技发展建设的关键[48]。近年来，我国芯片产业的技术水平提高，市场份额逐渐增加，但我国对芯片技术的规划布局不足，面对以技术保护为导向的国际环境，我国应当将政策转向关键技术领域。而分析芯片产业的技术主题，预测其未来趋势不但可以帮助企业在技术层面判别现状、洞察趋势，也可以帮助国家发现技术演变方向，提前谋划芯片产业的战略布局。本节以芯片产业技术为例，探讨芯片产业技术主题的现状及其趋势。

3.1 数据构成与数据来源

数据来源于智慧芽全球专利数据库，检索时间为2022年6月30日，考虑到研究需求为整年数据，检索范围为1978—2021年美国、欧洲各国以及经由世界知识产权组织登记的专利。选择了集成电路、电子设备、模拟电路等多个关键词，在B24、B81、C09等9个IPC分类号中检索，最终在数据库中获取所需要的专利数据，并对提取的数据筛查，得到所需的芯片产业专利文本库，共20 952条专利数据。

3.2 LDA主题识别

计算了5~150个主题数时的Perplexity值，当划分41个主题时，Perplexity值最小，如图4所示，因此设定主题数为41个。运用LDA模型将查询到的专利聚类，划分为41个主题组并得到各组的关键词。根据主题词情况过滤掉与芯片技术无关的2个主题聚类以及1个由虚词组成的聚类，剩余38个主题。主题命名依据来自《英汉电子信息技术缩略语词典》《英汉电子技术与电路词典》《新编英汉计算机与电子技术词典》等。由于芯片产业多用缩略词指代一些专有名词，后续也使用缩略词指代相关专有名词。通过查验发现各主题之间区别明显，边界清晰，符合主题分类的标准。为便于后文写作，对各主题进行编号与赋名，见表2。选用由2011—2021年的数据得到的技术主题绘制累计数量趋势图，见图5。

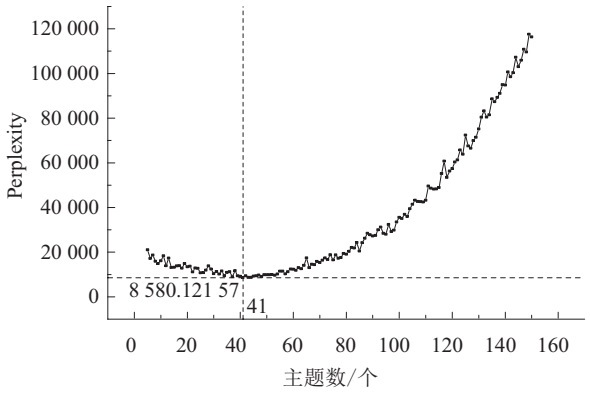

图4 不同主题数下的Perplexity值

静态随机存取存储器、桥式电路、放大器、芯片天线、传感器、碳化硅单晶等在2011—2021年的占比都保持在3%在以下。其中：静态随机存取存储器是二级高速缓存，并使用电晶体技术来保存数据，该技术在近十年间已逐渐被读写能力更强且本身的集成度也较高、可以多次读写的不挥发性磁性随机存储器取代；桥式电路的主要功能则是把交流变压电路输出的交流电转换成单向脉动线系数性直流电，该电路的工作原理较简单，因而近十年的主题占比较少；放大器放大高频的已调波信号的功率，达到适应高传输功率的目的；芯片天线将天线的设计集成到芯片上，具有质量轻、体积小、成本低等优点；传感器的作用主要是将那些非电信号的能用于测量或感受到的信息转换为电信号等可以输出的信息。由于放大器、芯片天线以及传感器技术早已成熟，近十年间的主题占比都较少。碳化硅由于具有耐高温与高导热率的特点，在芯片市场上性能优势十分显著，但碳化硅单晶对制造条件的要求较高，因而被边缘化。

少数的几个主题抛光液、电路布局、电子级树脂等占比较大。其中：抛光液的作用是使半导体材料表面平整，为后续的芯片加工打下基础；电路布局主要用于确定集成电路单元在芯片中的具体位置，其技术难点在于如何对越来越大规模的电路布局与优化；电子级树脂主要用于覆铜板制作、半导体封装等。抛光液、电路布局以及电子级树脂同芯片元器件密切相关，近十年间的占比较大。大多数主题的占比依然较少，随着时间的推移占比增长的现象不明显。

表2 主题名称

编 号	主题英文名	主题中文名
topic01	SRAM	静态随机存取存储器
topic02	CMOS	互补金属氧化物半导体
topic03	Bridge Circuit	桥式电路
topic04	Semiconductor Chip	半导体芯片
topic05	MRAM	不挥发性磁性随机存储器
topic06	Electronic Grade Resin	电子级树脂
topic07	Etch	蚀刻机
topic08	Chip Package	芯片封装
topic09	Peripheral	外围设备
topic10	Lithium Ion Battery	锂离子电池
topic11	Chip Bonding	芯片键合
topic12	Timing Model	时序模型
topic13	FPGA	现场可编程逻辑门阵列
topic14	Datum Memory	数据存储器
topic15	Chip Probe	芯片探针
topic16	Virtual Circuit	虚拟电路
topic17	Wafer Dicing	晶圆切割
topic18	SOC	系统级芯片
topic19	Wafer Level Camera	晶圆级相机
topic20	Chip Emulation	芯片仿真
topic21	Ceramic Module	陶瓷模块
topic22	Functional Test	功能测试
topic23	Circuit Layout	电路布局
topic24	Analog Signal	模拟信号
topic25	Amplifier	放大器
topic26	Film Adhesive	薄膜黏合剂
topic27	Electronic Apparatus	电子设备
topic28	Chip Antenna	芯片天线
topic29	Conductive Heat Adhesive	导热黏合剂
topic30	Sic Battery	碳化硅电池
topic31	Optical Chip	光学芯片
topic32	Silicon Wafer	硅晶片
topic33	Transistor	晶体管
topic34	Sensor	传感器
topic35	Polishing Liquid	抛光液
topic36	Logic Circuit	逻辑电路
topic37	Light Emitting Diode Chip	发光二极管芯片
topic38	Silicon Carbide Single Crystal	碳化硅单晶

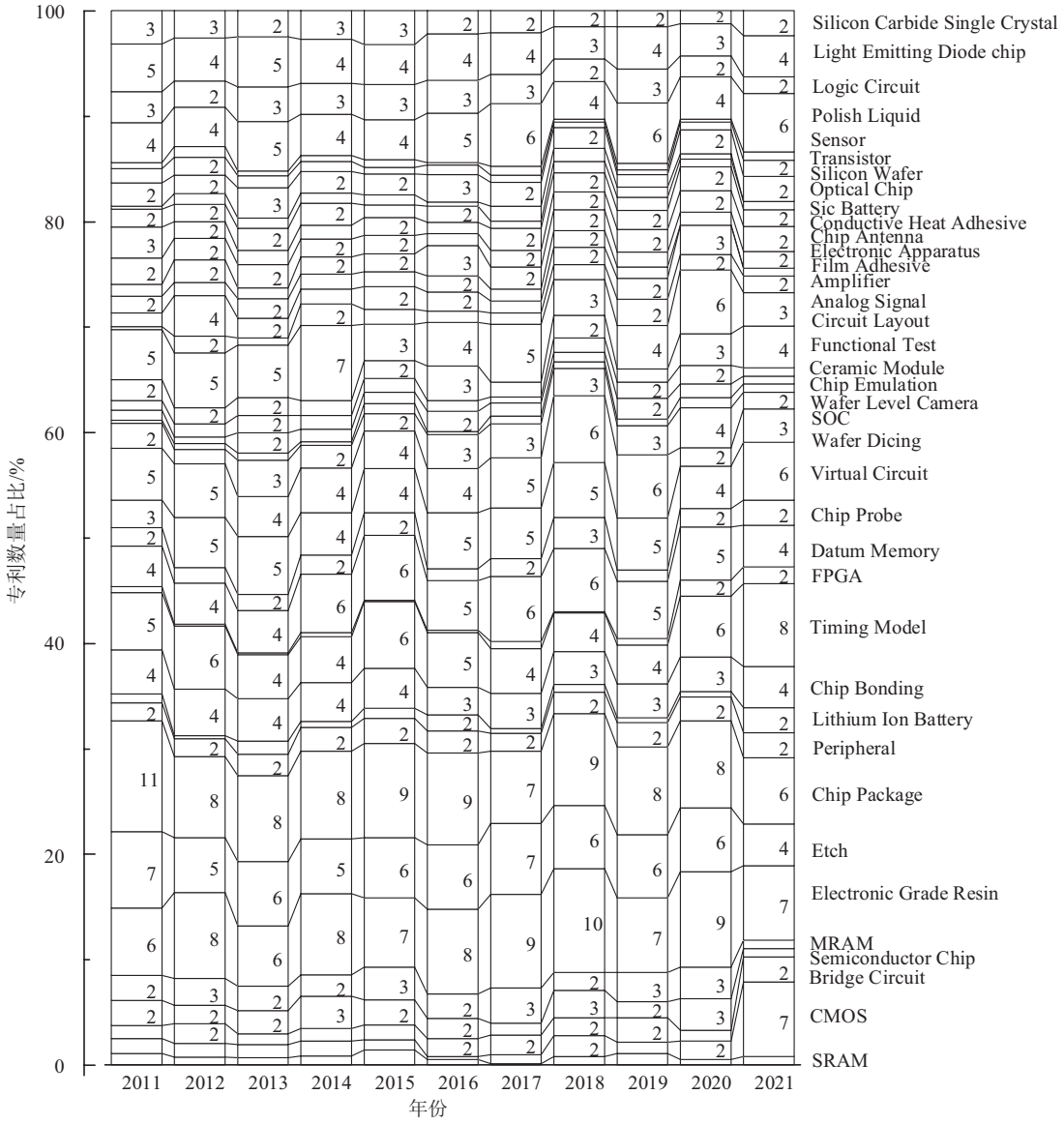

图5 主题累计占比

3.3 技术主题预测结果

通过预测各技术主题研发强度的变化趋势来体现芯片产业技术的未来演变。首先运用LDA对芯片产业专利主题分类，其次采用N-BEATS方法对技术主题预测。38个技术主题中每个样本的长度为176个专利，模型正向预测长度为25个专利、反向预测长度为16个专利，最终得到2022—2025年芯片产业技术主题的专利数量演化趋势预测结果，见图6。

从图6可以看出，2022—2025年芯片产业技术主题演化趋势之间的差别较大。这些技术主题中，未来的专利数量稳步上升的主题包括：主题6（电子级树脂）、主题7（蚀刻机）、主题8（芯片封装）、主题11（芯片键合）、主题35（抛光液），这5个技术主题是未来几年芯片产业技术热点。《2022年美国半导体行业现状》强调，未来与芯片-生物接口相关的设备通常有独特的封装需求，需要大量的材料和封装方面的研究来设计制造高性能医疗设备，以减少炎症反应和设备污染，这同所提到的芯片封装、芯片键合以及电子级树脂3个技术主题相关。此外《2022年美国半导体行业现状》强调要重视未来对晶圆制造企业的投资，这同蚀刻机和抛光液两个技术主题相关，说明了所提出的识别方法具有一定的可行性与合理性。

电子级树脂主要包括环氧树脂、酚醛树脂、硅树脂

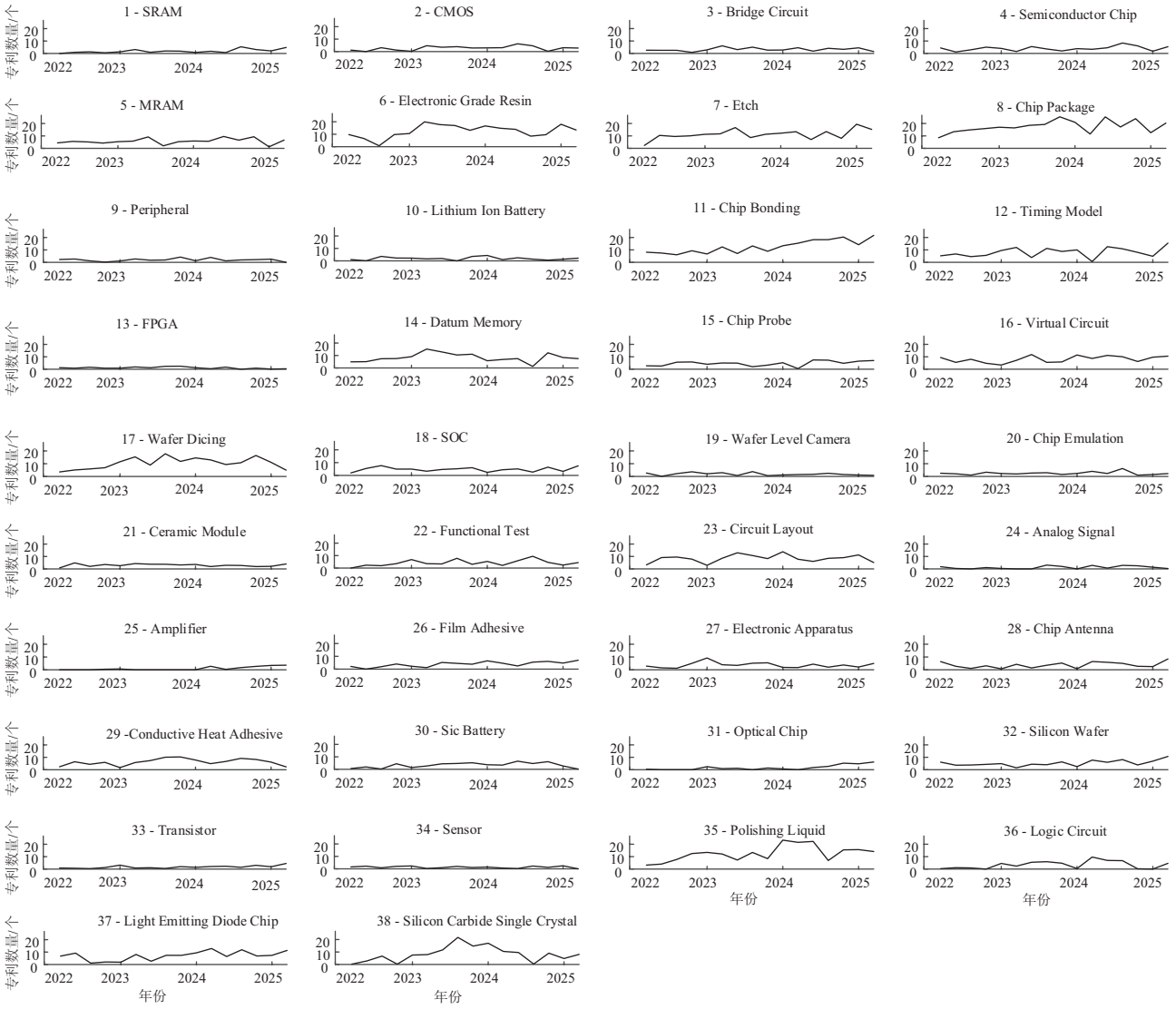

图6　2022—2025年技术主题演化趋势预测

等。环氧树脂作为集成电路的封装材料，可进一步降低成本、提高阻燃耐热性能、提高模量等，对未来技术提升有着很大帮助。此外对酚醛树脂，应进一步提高性能以及开发新子类，以适应未来芯片的轻薄化、高频高速化以及多功能化发展。硅树脂由于具有电绝缘性更高、更为耐臭氧以及热氧化稳定性更高等优点，未来可朝着用于芯片、电路板的定制化产品以及结构可控等方向发展。

蚀刻机的工作原理是按光刻机刻出的电路结构，在硅片上进行微观雕刻，刻出沟槽或接触孔。在3D NAND制造工艺中，增加集成度的方法主要是增加堆叠的层数。3D化集成电路对蚀刻机提出了更高的要求。

随着未来对芯片功能以及功率的要求提升，解决多芯片ESOP（Exposed-Pad Small Outline Package）集成电路的分层控制问题、降低超大尺寸芯片封装内应力的影响将成为芯片产业热点。芯片键合工艺有两类：一类是引线键合，通过金属引线实现芯片与基板间的电气互连和芯片间的信息互通，其优点包括可实现性高、成本较低等；另一类是倒装芯片键合，通过倒置芯片同封装基板直接连接，不需要引线的参与，其中金属粒子烧结键合具有更优异的导电、耐高温、导热等特性，受到了芯片键合领域的广泛关注。

抛光液在晶圆抛光的过程中起着重要作用。随着芯片特征尺寸的缩小，对晶圆抛光效果的要求提高，但当前抛光效果不好、成本过高以及环保压力过高等问题亟需解决，因此新型阻挡层材料和环保材料的抛光液是未来抛光液的热点开发方向。

未来的专利数量先增长后回落的主题有主题14（数

据存储器）、主题17（晶圆切割）、主题27（电子设备）以及主题38（碳化硅单晶）。这几种技术都已经趋于成熟，例如早期的数据存储器采用由分离式小规模芯片所搭建的控制模块，后期则针对大规模器件采用遥感系统数据存储器控制模块。晶圆切割是指将整片晶圆按芯片大小分割成单一的芯片，早期采用钻石锯片砂轮进行切割，后期则采用激光进行无切割式加工。碳化硅单晶由于其热稳定性高、键合能高、耐氧化等，成为第三代半导体材料，在军工等领域被大规模应用。这些主题专利数量的回落可能是因为主题已经趋于成熟。

未来的专利数量增速较为平缓的主题包括：主题2（互补金属氧化物半导体）、主题3（桥式电路）、主题4（半导体芯片）、主题5（不挥发性磁性随机存储器）、主题19（晶圆级相机）、主题21（陶瓷模块）、主题22（功能测试）、主题23（电路布局）、主题28（芯片天线）、主题30（碳化硅电池）等。互补金属氧化物半导体由于几乎没有静态功耗，并且有高噪声容限等优点，如今在计算机内存、手机中广泛应用。半导体芯片虽然正在达到物理的极限，但由于未来智能技术的爆发式增长必然是半导体芯片所驱动的，当前研究集中于增强半导体芯片的性能，包括智能传感、能源效率、安全性等方面。不挥发性磁性随机存储器比起传统的存储器所消耗的功率更低，更为持久，应用极为广泛。晶圆级相机则因为其更小的光学直径以及更高的分辨率受到了医学界的关注。芯片天线是一种带有导体的天线，当前芯片天线最大的优势就是体积极小，例如广泛应用于纳米收发器等。碳化硅电池由于其更低的损耗、更小的体积以及更方便集成的特性，已经在头部新能源车厂获得了青睐与应用。此外桥式电路等技术领域已经较为成熟，未来专利数量增速比较平缓。

未来的专利数量先下降再小幅增加的主题是主题26（薄膜黏合剂）、主题36（逻辑电路）、主题37（发光二极管芯片）。当前芯片市场上对更小、更薄和更可靠器件的黏结薄膜的需求不断增长，如何满足不同芯片对薄膜黏合剂的要求是芯片产业当前的热点问题之一。随着集成电路设计的复杂化，进一步采用近似逻辑综合优化电路成为了当前的热点，但多级逻辑近似优化算法如何同时优化大型电路效果和算法速度的问题依然有待解决。由于氮化镓基垂直结构发光二极管相比普通半导体发光二极管更为节省能源、体积更小、寿命更长、光效更强，受到了发光二极管芯片领域的广泛关注。但当前如何在不导致器件短路的前提下得到单个氮化镓基垂直结构发光二极管芯片的问题依然难以解决，此外如何在分离器件时减少反向漏电也是一大问题。这些主题的专利数量先降后升可能是因为早期技术要求无法达到，后续技术水平提高的时候主题的研究热度逐步上升。

未来的专利数量减少的主题包括：主题1（静态随机存取存储器）、主题9（外围设备）、主题24（模拟信号）、主题33（晶体管）、主题34（传感器）。这可能是由于上述主题目前已经形成成熟完备的技术，这些技术主题未来将不再是相关领域最为关注的主题。比如静态随机存取存储器早期由于数据不随电路更新而改变的优点受到了广泛关注，但不挥发性磁性随机存储器的读写时间更短，目前已经取代静态随机存取存储器成为高速缓冲存储器的主要存储媒介。

3.4 效果评价

为验证LDA和N-BEATS网络预测模型组合方法的优异性，运用LSTM模型和IPC分类号分别作为N-BEATS网络模型预测方法和LDA的替代模型。在Pytorch 1.12.1以及Python 3.8.3软件中设置LSTM模型迭代的次数为300，输入时间序列长度为16，隐藏层数为100，损失函数和优化函数为MSE函数和Adam优化函数。按各专利的IPC主分类号划分9个类别，为B24、B81、C09、G01、G05、G06、H01、H03以及H05。采用所收集的芯片产业专利文本库，共20 952条芯片产业专利数据，分别计算4个组合方法（LDA-N-BEATS、LDA-LSTM、IPC-N-BEATS、IPC-LSTM）下2018—2021年每个季度各个技术主题专利数量的预测值与实际值的均方根误差（见表3）。

表3 各方案均方根误差统计

方案名	均方根误差
LDA-N-BEATS	4.604 078
LDA-LSTM	4.753 839
IPC-N-BEATS	15.854 868
IPC-LSTM	19.426 080

可以看出，LDA-N-BEATS的均方根误差在4个方案中最小，表示LDA-N-BEATS的预测效果最好。此外IPC分类号仅能将芯片产业专利数据分成9个技术主题分类，而LDA可以划分38个技术主题，技术主题划分精

度高于IPC分类方法，并且IPC-N-BEATS的均方根误差小于IPC-LSTM的均方根误差，这说明N-BEATS的预测精度高于LSTM模型。

4 结论和建议

本文结合LDA算法和N-BEATS网络预测模型，运用LDA模型提取专利文献数据的技术主题，引入N-BEATS网络模型分析各技术主题专利数量的时间序列，将技术活动周期性纳入技术主题预测模型，提高了模型的可解释性，并以芯片技术为例，运用该组合方法预测了产业的技术主题和未来趋势。对比实验发现，所提出的预测模型的分类精度和预测精度均高于LDA-LSTM、IPC-N-BEATS以及IPC- LSTM模型。

在芯片产业技术主题和未来趋势案例中，该方法将芯片产业专利数据分为38个技术主题，其中电子级树脂、蚀刻机、芯片封装、芯片键合、抛光液是未来几年芯片产业技术热点。

我国芯片产业的相关企业技术研发机构以及政府相关部门应做好芯片产业未来的技术布局，此外提高芯片制作的精准度与适应未来芯片的轻薄化、高频高速化以及多功能化发展，解决多芯片ESOP集成电路的分层控制问题，降低超大尺寸芯片封装内应力的影响，开发新型阻挡层材料和环保材料的抛光液是当前芯片领域面对的挑战，应引起芯片企业技术研发机构及政府有关部门的重视。

未来研究应扩充数据库，纳入多元数据源，如科研文章、项目文本等，降低计算复杂度，提升迭代效率，以实现针对更大规模数据集的分析工作。

参考文献

[1] 陈伟，林超然，孔令凯，等. 基于专利文献挖掘的关键共性技术识别研究[J]. 情报理论与实践，2020，43（2）：92-99.

[2] 中国科学院综合计划局，中国科学院成都文献情报中心. 中国科学院专利分析报告[M]. 成都：中国科学院成都文献情报中心，2015.

[3] 林岩. 基于专利数据的知识计量研究评述[J]. 科技管理研究，2008，28（9）：91-93.

[4] TRAN T A，DAIM T. A taxonomic review of methods and tools applied in technology assessment[J]. Technological Forecasting and Social Change，2008，75（9）：1396-1405.

[5] POPPER R. How are foresight methods selected[J]. Foresight，2008，10：62-89.

[6] 杨祖国，李文兰. 中国专利被专利文献引用的主题分析[J]. 情报科学，2005，23（12）：1845-1851.

[7] 车尧，李雪梦. 基于德温特手工代码的专利技术分析：以风能为例[J]. 情报科学，2015，33（4）：132-138.

[8] 方伟，曹学伟，高晓巍. 技术预测与技术预见：内涵、方法及实践[J]. 全球科技经济瞭望，2017，32（3）：46-53.

[9] 毛云莹，陆伟. 基于IPC关联的专利技术和产业双向分析框架研究[J]. 情报科学，2022，40（4）：33-39.

[10] 冉从敬，徐晓飞. 基于NodeJS+ECharts的专利权人引证关系可视化方法研究[J]. 情报科学，2018，36（8）：77-83.

[11] 马铭，王超，周勇，等. 基于语义信息的核心技术主题识别与演化趋势分析方法研究[J]. 情报理论与实践，2021，44（9）：106-113.

[12] LIU J，WEI J，LIU Y. Technology forecasting based on topic analysis and social network analysis：a case study focusing on gene editing patents[J]. Journal of Scientific & Industrial Research，2021，80（5）：428-437.

[13] BLEI D，NG A，JORDAN M. Latent Dirichlet allocation[J]. Journal of Machine Learning Research，2003，3（1）：993-1022.

[14] 罗恺，袁晓东. 基于LDA主题模型与社会网络的专利技术融合趋势研究：以关节机器人为例[J]. 情报杂志，2021，40（3）：89-97.

[15] 马建红，王晨曦，闫林，等. 基于产品生命周期的专利技术主题演化分析[J]. 情报学报，2022，41（7）：684-691.

[16] CHOI D，SONG B M. Exploring technological trends in logistics：topic modeling-based patent analysis[J]. Sustainability，2018，10（8）：2810.

[17] 陈荣，王小庆，李建霞，等. 基于多参数动态时序的技术预测方法与实证研究[J]. 科技管理研究，2022，42（14）：173-183.

[18] DUAN H M，PANG X Y. A multivariate grey prediction model based on energy logistic equation and its application in energy prediction in China[J]. Energy，2021，229：120716.

[19] YUSKEVICH I，SMIRNOVA K，VINGERHOEDS R，et al. Model-based approaches for technology planning and roadmapping：technology forecasting and game-theoretic modeling[J]. Technological Forecasting and Social Change，2021，168：120761.

[20] INOUE H，SOUMA W，TAMADA S. Spatial characteristics of

joint application networks in Japanese patents[J]. Physica A: Statistical Mechanics and Its Applications, 2007, 383(1): 152-157.

[21] 王璐, 杨书, 张强, 等. 时间序列组合式模型在艾滋病感染率变化趋势中的应用[J]. 中国卫生统计, 2013, 30(2): 196-198.

[22] 陈萍萍. 基于时间序列的旅游需求预测模型[J]. 统计与决策, 2013, 29(18): 11-13.

[23] 刘金培, 林盛, 郭涛, 等. 一种非线性时间序列预测模型及对原油价格的预测[J]. 管理科学, 2011, 24(6): 104-112.

[24] 王惠婷. 基于混合时间序列模型的粮食产量预测[J]. 统计与决策, 2013, 29(12): 23-25.

[25] GHAREHDAGHI M, FAKHER A, CHESHOMI A. The combined use of long-term multi-sensor insar analysis and finite element simulation to predict land subsidence[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019, XLII-4/W18: 421-427.

[26] 白敬毅, 颜端武, 陈琼. 基于主题模型和曲线拟合的新兴主题趋势预测研究[J]. 情报理论与实践, 2020, 43(7): 130-136, 193.

[27] 刘自强, 王效岳, 白如江. 基于时间序列模型的研究热点分析预测方法研究[J]. 情报理论与实践, 2016, 39(5): 27-33.

[28] ABUHAY T M, NIGATIE Y G, KOVALCHUK S V. Towards predicting trend of scientific research topics using topic modeling[J]. Procedia Computer Science, 2018, 136: 304-310.

[29] 徐路路, 王芳. 基于支持向量机和改进粒子群算法的科学前沿预测模型研究[J]. 情报科学, 2019, 37(8): 22-28.

[30] 郭澳庆, 胡俊, 郑万基, 等. 时序InSAR滑坡形变监测与预测的N-BEATS深度学习法: 以新铺滑坡为例[J]. 测绘学报, 2022, 51(10): 2171-2182.

[31] LU X H, ZHANG Y W, LIN C R, et al. Evolutionary overview and prediction of themes in the field of land degradation[J]. Land, 2021, 10(3): 241.

[32] ZHANG Y W, LU X H, LIN C R, et al. A new method for identifying key and common themes based on text mining: an example in the field of urban expansion[J]. Discrete Dynamics in Nature and Society, 2021, 2021: 1-14.

[33] 许学国, 桂美增. 基于深度学习的技术预测方法: 以机器人技术为例[J]. 情报杂志, 2020, 39(8): 53-62.

[34] 唐晓灵, 刘嘉敏. 基于PSO-LSTM网络模型的建筑碳排放峰值预测[J]. 科技管理研究, 2023, 43(1): 191-198.

[35] KUMAR B P, HARIHARAN K, SHANMUGAM R, et al. Enabling internet of things in road traffic forecasting with deep learning models[J]. Journal of Intelligent & Fuzzy Systems, 2022, 43(5): 6265-6276.

[36] SUN D Z, HUANG X, ZHAO Z R, et al. Deep learning-based solar flare forecasting model. III. extracting precursors from EUV images[J]. The Astrophysical Journal Letters Supplement Series, 2023, 266(1): 8.

[37] ZAMEER A, JAFFAR F, SHAHID F, et al. Short-term solar energy forecasting: integrated computational intelligence of LSTMs and GRU[J]. PLoS ONE, 2023, 18(10): e0285410.

[38] MHASKAR H N, POGGIO T. Deep vs. shallow networks: an approximation theory perspective[J]. Analysis and Applications, 2016, 14(6): 829-848.

[39] 陈晋音, 陈奕芃, 陈一鸣, 等. 面向深度学习的公平性研究综述[J]. 计算机研究与发展, 2021, 58(2): 264-280.

[40] KABUKCUOGLU Z. The cyclical behavior of R&D investment during the Great Recession[J]. Empirical Economics, 2019, 56(1): 301-323.

[41] 成力为, 朱孟磊, 李翘楚. 政府补贴对企业R&D投资周期性的影响研究: 基于融资约束视角[J]. 科学学研究, 2017, 35(8): 1221-1231.

[42] SHAPHIR E, PINTER R Y, WIMER S. Efficient cell-based migration of VLSI layout[J]. Optimization and Engineering, 2015, 16(1): 203-223.

[43] ORESHKIN B N, CARPOV D, CHAPADOS N, et al. N-BEATS: neural basis expansion analysis for interpretable time series forecasting[J/OL]. arXiv Preprint, arXiv: 1905.10437 [2023-05-24]. https://arxiv.org/abs/1905.10437.pdf.

[44] PUTZ D, GUMHALTER M, AUER H. A novel approach to multi-horizon wind power forecasting based on deep neural architecture[J]. Renewable Energy, 2021, 178: 494-505.

[45] 苏行, 杨韬, 孙保琪, 等. 基于N-BEATS的单站对流层天顶总延迟预报[J]. 中国空间科学技术, 2022, 42(2): 56-63.

[46] PUSZKARSKI B, HRYNIÓW K, SARWAS G. Comparison of neural basis expansion analysis for interpretable time series (N-BEATS) and recurrent neural networks for heart dysfunction classification[J]. Physiological Measurement, 2022, 43(6): 064006.

[47] 方莹莹, 刘戒骄. 从开放式协同创新看中国芯片产业生态圈营造[J]. 产经评论, 2018, 9(6): 104-115.

[48] 范旭, 刘伟. 中美贸易冲突下的半导体创新政策工具选择[J]. 科学学研究, 2020, 38(7): 1176-1184.

作者简介

吴雷，男，博士，副研究员，研究方向：技术创新、知识产权管理。
杜文研，女，硕士研究生，研究方向：技术创新。
林超然，男，博士，讲师，通信作者，研究方向：技术创新、知识产权管理，E-mail：linchaoran@hrbeu.edu.cn。

Technology Themes Prediction Based on Combination of LDA and N-BEATS Methods Applied to Patent Data

WU Lei DU WenYan LIN ChaoRan
(School of Economics and Management, Harbin Engineering University, Harbin 150001, P. R. China)

Abstract: Predicting the future hot topics of technology themes can enable enterprises to distinguish the current situation, identify future technological directions, and plan their strategic layout in advance at the technical level. This study proposes a combination method of LDA and N-BEATS, which uses the LDA model to extract technical topics from patent literature data. The N-BEATS network model is introduced to analyze the time series of the number of patents for each technical topic, leveraging its advantage in analyzing interpretable time series. The periodic module of technological research and development activities is added to the prediction model, and the chip technology is used as an example to predict the technical themes and future trends of the industry using this combination method. The prediction accuracy of the LDA and N-BEATS combination method in the comparative experiment is higher than that of the three benchmark methods: LDA-LSTM, IPC-N-BEATS, and IPC-LSTM. The case results indicate that the future research and development hotspots in the chip industry are electronic grade resins, etching machines, chip packaging, chip bonding, and polishing fluids.

Keywords: LDA; N-BEATS Network Model; Deep Learning; Chip Industry; Technological Forecast

（责任编辑：王玮）