

DOI: 10.20098/j.cnki.1001-5132.2022.1211

## 基于图嵌入技术的产业集群分析方法研究

牛士银, 余军合\*, 战洪飞, 王 瑞, 刘东洋

(宁波大学 机械工程与力学学院, 浙江 宁波 315211)

**摘要:** 以往对产业集群的相关实证研究存在数据获取困难、数据维度片面、传统复杂网络理论分析方法可拓展性差等问题. 针对以上问题, 本文以互联网上的大量非结构化数据为基础, 采用图嵌入模型提取集群网络特征的向量空间分析方法, 利用互联网公开数据构建产业集群关联网络, 结合企业行业分类标准与分析目的设计部分节点标签, 使用关系型图卷积神经网络模型(R-GCNs), 从产品关联层面进行产业集群特征学习. 根据产业集群内企业的嵌入表示和地理位置信息, 提出了集群网络嵌入应用分析方法. 通过对宁波地区制造业集群相关数据进行实验分析和论证, 验证了图嵌入分析方法在量化分析产业集群关联网络特征上的有效性.

**关键词:** 产业集群; 关联网络; 关系型图卷积网络; 区域特性

**中图分类号:** F276; T-9

**文献标志码:** A

**文章编号:** 1001-5132 (2023) 05-0029-08

产业集群是由多种异构主体(企业、研究机构、中介机构、行业协会、政府部门等)通过竞争、协作组成的复杂系统, 其结构特性、演化发展规律及核心竞争力受到诸多学者关注<sup>[1-3]</sup>. 从已有相关文献来看, 学者有侧重于对产业集群识别方法进行研究<sup>[4-6]</sup>, 也有对产业集群技术演化展开研究<sup>[7]</sup>. 同时, 对产业集群相关问题研究的范式也受到学者学科背景的影响, 难以统一产业集群的研究主题和方法, 目前较为主流方法是采用复杂网络理论, 通过统计性网络结构指标来分析产业集群.

产业集群的重要表现形式是由其各主体所形成的关联网络. 从研究方法特点来说, 对产业集群关联网络的研究总体上可以分为两种: 一种是对产业集群关联网络的仿真研究<sup>[8]</sup>; 另一种是基于现实数据对产业集群关联网络的实证研究<sup>[9-10]</sup>. 其中, 产业集群关联网络实证分析的研究多数以调研数据为主, 而获取产业集群企业间的关联关系来构建网络模型是关键, 也是难点. 随着产业集群体系的不断扩张和企业对自身信息保护意识的增强,

当下调研数据更加难以大量获取, 因此基于小样本的产业集群关联网络分析局限性显得愈发显著. 互联网技术的发展有效地解决了数据量问题, 有部分学者使用大批量的企业投入产出数据构建了大规模的产业集群网络<sup>[11-12]</sup>. 而企业地理位置作为产业集群的一个重要属性, 是比较容易批量获取的企业数据. 例如, 李经成等<sup>[13]</sup>基于创新型企业地理位置信息, 采用复杂网络理论、无监督聚类等多种空间研究方法, 对南京市创新型企业展开了分析; 沈体雁等<sup>[14]</sup>联合了投入产出和地理邻近双重信息, 识别出多种中国国家标准产业集群, 并分析它们的特征差异.

综上所述, 产业集群现实中所包含关系是复杂的. 但在国内外学者相关研究中, 存在调研数据样本数量小, 或者所构建的产业集群网络维度单一问题; 分析方法也都偏向于采用复杂网络理论, 依靠某些网络指标来分析产业集群网络结构特征, 难以有效挖掘产业集群的微观信息. 本文针对以上问题, 提出从嵌入视角学习产业集群关联网络

收稿日期: 2022-12-12.

宁波大学学报(理工版)网址: <http://journalg.nbu.edu.cn/>

基金项目: 国家重点研发计划项目(2019YFB1707101, 2019YFB1707103); 国家自然科学基金(71671097); 浙江省省属高校基本科研项目(SJLZ2023001); 浙江省公益技术应用研究计划项目(LGG20E050010, LGG18E050002).

第一作者: 牛士银(1996-), 男, 河南周口人, 在读硕士研究生, 主要研究方向: 知识管理、企业信息化. E-mail: niushiying\_nb@163.com

\*通信作者: 余军合(1971-), 男, 湖北天门人, 博士/副教授, 主要研究方向: 知识管理、智能制造. E-mail: yujunhe@nbu.edu.cn

特征的角度,从互联网获取各种半结构化和非结构化数据来构建产业集群关联网络,并通过图嵌入模型学习产业内丰富的关联特征,以及结合企业地理位置信息展开产业集群分布特征和区域特性分析,为产业集群相关研究提供新的视角。

## 1 网络模式与图嵌入模型选择

### 1.1 网络模式

大数据时代背景下,各种企业信息管理平台的存在使得从互联网获取产业集群的公开数据成为可能。本文以行政区域作为产业集群研究数据边界,获取互联网上的企业数据,主要包含企业的经营状态、规模、地理位置、成立时间、法人等相关信息。产业集群内复杂的关联形式表现为企业的产品相似性、技术相关性、地理邻近性、供应链关系等,这些关系将集群内企业紧密联系成为一个具有竞合关系的区域经济体。本文在考虑产业集群关联网络时,认为企业的业务围绕并服务于所经营的产品,即集群关系网络依托于集群内的产品。因此,通过企业与产品间关系构建的关联网络是一种符合实际情况的集群网络模式。文中以企业为核心、以产品为关联节点的产业集群关联网络模式如图1所示。

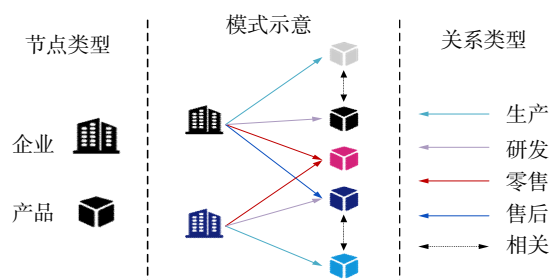


图1 产业集群关联网络模式示意图

图1为有向的异质图,生产关系体现企业的加工设备资源和加工能力;研发关系体现企业的新技术开发能力;零售关系体现企业该类型产品的业务终端是面向企业,还是面向大众;售后服务关系体现企业是否具有该产品的技术支持能力;产品间相关关系是符合实际特征的关系,如产品名称语义相同或相近,但表述不同的情形。

### 1.2 图嵌入模型选择

图嵌入模型是将网络(图结构)数据转换至低维向量空间,同时保留原始数据的关键信息。产业

集群的关联网络数据隐藏于文本中,对产业集群的实证分析而言,图嵌入是挖掘产业集群关联网络特征的高效技术手段。基于不同理念,经典的图嵌入模型可分为三类:(1)Trans系列的线性翻译模型<sup>[15-16]</sup>,通过 $h+r=t$ 形式,将网络中存在具体关系 $r$ 的两节点 $h$ 和 $t$ 转至向量空间;(2)基于图的随机游走模型<sup>[17]</sup>,沿着图(网络)上的随机路径采样,以形成“句子”,将其中“单词”(网络中节点)转化为向量;(3)以GCN为代表的图神经网络模型<sup>[18]</sup>,通过卷积核聚集节点周围信息来获取节点的嵌入向量表示。

依据构建的集群网络模式,企业节点可能与多个产品节点存在某种关系,而线性翻译模型不适用于处理该类型网络数据,同样,随机游走模型也可能遗失企业间重要的产品关联特征,但GCN卷积核思想可用产品节点来表示企业。针对其无法处理多关系网络数据的问题,Schlichtkrull等<sup>[19]</sup>对其进行改进,提出了关系型图卷积网络(R-GCNs)模型,该模型又分为节点分类和链接预测。节点分类模型侧重于挖掘网络中隐含于标签信息下的重要特征,而链接预测模型则是在获取特征基础上,使用特定的评分函数进行节点间关系的预测。

综上所述,选择半监督(部分节点具有标签)的R-GCNs节点分类模型,旨在通过聚合不同关系下的邻居产品节点信息,以获取产业集群关联网络中各企业节点的向量表示,从而利用嵌入所学习到的企业特征展开产业集群的实证研究。

## 2 基于图嵌入技术的产业集群分析方法

基于图嵌入技术的产业集群分析研究主要分为三部分。首先将从互联网获取的产业集群公开数据进行预处理,构建产业集群关联网络图结构数据;其次结合行业类别标准设计部分企业标签,使用R-GCNs节点分类模型进行半监督的节点嵌入特征学习,并评估集群关联网络嵌入质量;最后结合地理信息,在产业集群向量空间展开分析。

### 2.1 产业集群关联网络构建方法

企业经营文本信息中包含相应的产品及关系,通过文本句子关键词筛选,将经营信息分割成四类:生产信息、研发信息、零售信息和售后服务信息。分割后的文本信息包含很多专用名词,且不是

常见的主-谓-宾句子类型,而目前的实体抽取<sup>[20]</sup>方法很难应用到此类文本中。因此考虑采用 jieba 分词库来处理此类信息,并通过构建专有名词名称词典及去除停用词,将分割后的经营文本转化为多个独立的产品专有名词进行产品实体抽取。

对产品间相关关系建模,可使用预训练 BERT 模型<sup>[21]</sup>,并结合余弦相似度阈值计算产品名称间的相关关系,但不同阈值所建立的相关关系对于图的质量以及集群网络特征的学习有较大影响。对图的结构分析而言,度分布是其最基本的特征,因此可根据网络度分布分析方法来选择阈值。

## 2.2 产业集群关联网络嵌入及效果评价方法

本文使用的 R-GCNs 节点分类模型是以交叉熵为损失函数的半监督机器学习模型,需要对构造图设计部分标签,以指导模型学习特征。以挖掘产业集群企业嵌入特征的有效表示为目的,将行业类别作为网络中企业节点的标签,结合企业经营信息和制造业行业分类标准对部分特征明显的企业节点进行类别标注。

R-GCNs 模型将网络中每种业务关系看作是有向(入边和出边)关系。企业与产品间的单向关系对企业是出边关系,对产品是入边关系,产品与产品间关系是双向关系,模型输入为不同方向关系所构建的邻接矩阵集合。

模型输出为节点的嵌入向量和节点类别,评估模型训练嵌入质量是后续应用分析的必要前提。结合传统分类模型评价指标及产业集群特点,在分类准确率和混淆矩阵评估指标的基础上,增加方差指标和主成分分析法来评估模型的嵌入效果,其中方差与主成分分析均根据模型训练后输出的企业嵌入表示其具体计算值。

## 2.3 产业集群关联网络嵌入应用分析方法

产业集群的分析离不开企业地理位置。作为企业的唯一地理属性标识,一般为短文本数据,通常可使用百度地图地名查询 API(应用程序编程接口)以获取企业的经纬度编码信息。在计算企业间距离时,可通过经纬度编码直接计算企业间地理上的绝对距离,也可通过百度地图直接获取企业间的行驶距离。相比于计算地理上的绝对距离,行驶距离可以很好地捕获由地貌环境特殊性产生的影响,从而更合理地在地理空间上分析产业集群

内各企业主体与其邻近企业的关联性。

为应用嵌入所学习到的企业的业务特性来研究产业集群,本文首先作了下列定义。

**定义 1** 产业集群  $C_n = \{c_i\}$ ,  $i = 1, 2, \dots, n$ ,  $c_i$  为产业集群内的某个企业节点,其向量表示为  $\mathbf{v}_i$ 。  $C_n$  所包含的企业全部行业分类  $K_m = \{k_a\}$ ,  $a = 1, 2, \dots, m$ , 其中,  $k_a$  泛指某行业类别下的企业,其行业质心向量表示为  $\bar{\mathbf{v}}_a^k$ 。

**定义 2** 产业集群企业地理位置区域范围  $R_t^d$  指以企业  $c_i$  为区域中心,以道路距离等于  $d$  为虚拟半径的某个区域,在该区域包含的企业集合定义为  $C_t^d$ ,  $C_t^d \subseteq C_n$ , 其中,  $t$  为该集合中的企业个数。

**定义 3** 在定义 1、2 基础上,企业  $c_i$  所在局部区域  $R_t^d$  内,该类型企业综合关联程度评估指标定义为  $S^d(c_i, C_t^d)$ :

$$S^d(c_i, C_t^d) = \frac{1}{t} \sum_{j=1}^t \cos(\mathbf{v}_i, \mathbf{v}_j), \quad (1)$$

式中:  $\mathbf{v}_j$  为  $C_t^d$  中的某个企业向量;  $\cos(\mathbf{v}_i, \mathbf{v}_j)$  为 2 个向量的余弦值。

在上述定义基础上,首先以热力地图形式展示集群的行业地理分布。以行业质心向量  $\bar{\mathbf{v}}_a^k$  与企业向量  $\mathbf{v}_i$  的相似度为热点权重,并设定显示阈值;以企业经纬度为坐标绘制热力地图,对比不同行业属性下的企业地理分布特征,挖掘集群内不同行业属性下企业的分布共同点和集群特性。这种方式是从整个产业集群企业嵌入,而不是仅从行业分类视角来分析产业集群内的企业,且还需考虑  $k_a$  行业之外企业也可能具有隐性的关联性特征。

已有文献表明,地理上邻近的企业知识和技术会迅速传播,从而导致产业集群内部区域企业间的同质性增强<sup>[22]</sup>。使用定义 3 中的式(1),从评估区域特性角度应用企业的嵌入,以集群内某个特征的企业作为  $c_i$ ,从产业集群地理邻近视角来看,不同企业所在的不同区域位置在集群中应当表现出各自的区域特性,而企业综合关联程度评估指标  $S^d$  能够将每种情形都定量化表示。

## 3 实验与结果分析

本文数据集来源于“企查查”企业信息管理平台公开数据,截至 2022 年 5 月份,共获取 63 000 多条宁波地区内制造业产业集群相关企业数据,经

表1 文本处理结果示例

| 企业(ID) | 经营文本(处理前)                                             | 经营文本(处理后)             |            |    |            |
|--------|-------------------------------------------------------|-----------------------|------------|----|------------|
|        |                                                       | 生产                    | 研发         | 零售 | 售后         |
| 1      | 一般经营项目: 液压机械设备、电动工具、气动工具、五金工具的制造、液压机械设备及其配件的研发、售后维修服务 | 液压机械设备、电动工具、气动工具、五金工具 | 液压机械设备及其配件 | -  | 液压机械设备及其配件 |

过删除经营目录信息缺失或异常的数据后, 59 261 条半结构化企业数据被用来构建集群关联网络。

### 3.1 产业集群关联网络构建结果

通过构建的 5 000 多个产品名称专有名词库, 本文将企业的经营目录文本信息转化为生产、研发、零售、售后服务 4 种关系及对应产品, 共包含 11 563 个产品, 表 1 即为某企业经营文本数据处理结果展示。

在处理产品间的相似关系时, 本文以拼接预训练 BERT 模型后两层的平均输出作为产品的向量表示。在建立产品间相关关系阈值选择时, 图 2 揭示了不同阈值下产品关联图(仅由产品构成的同构图)的度分布, 以及本文最终构建的产业集群关联网络图结构数据度分布。

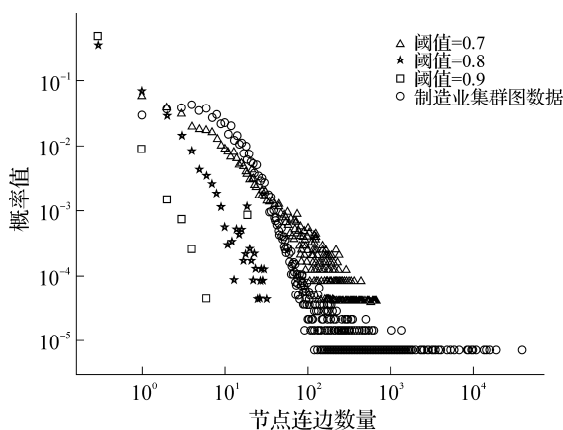


图2 不同阈值下产品相关性度分布

由图 2 可见, 当相似度阈值为 0.7 时, 其平均节点度约为 19.503, 表明该阈值建立的产品间关联关系过于稠密, 显然不符合制造业集群企业的产品潜在关联特征; 当相似度阈值为 0.9 时, 其平均节点度约为 0.064, 产品相似关系过于稀疏; 而当相似度阈值为 0.8 时, 其平均节点度约为 0.783, 能够真实地表示该制造业集群产品间的相关性联系特点, 故最终选定相似度阈值为 0.8。在本文制造业集群网络数据的度分布中, 加入产品之间的相

关关系后, 仅有 3% 的节点是稀疏的(仅与一个节点存在联系), 这种数量的稀疏关系完全是模型所能承受的。虽然存在有超过 10 000 条的连边情况, 但这也是制造业产业集群的特点——某些通用产品会被较多的企业经营, 且每个企业节点可能与一个产品节点建立 4 条连边(4 种关系)。

### 3.2 产业集群关联网络嵌入及效果评价

根据制造业企业行业分类标准, 本文对部分企业进行了 15 类单标签标注, 并对其进行 One-Hot 编码, 作为模型标签输入。模型共输入 59 261 个企业节点和 11 563 个产品节点, 对任一单向关系构建  $70824 \times 70824$  维的关系邻接矩阵, 最后将所有关系邻接矩阵(包括 8 个企业-产品关系矩阵, 2 个相同产品相关矩阵, 1 个节点自连接对角矩阵)水平堆叠, 得到  $70824 \times 779046$  维矩阵作为模型的输入。

将标注数据的 60% 作为训练集, 余下作为测试集, 使用 Pytorch 框架搭建模型, 嵌入维度设置为 50 维(经多次试验证明, 当嵌入维度为 50 维时,

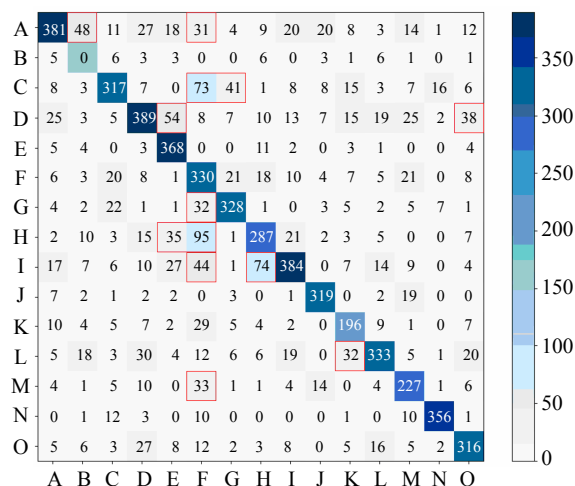


图3 R-GCNs 模型分类结果混淆矩阵

A:金属制品制造业; B:金属冶炼加工业; C:文体用品制造业; D:其他专用设备制造业; E:塑机及机床制造业; F:橡胶和塑料制品业; G:造纸印刷业; H:模具制造业; I:汽车及其配件制造业; J:非金属建材制造业; K:电器制造业; L:通信电子元件制造业; M:化学及纤维制品制造业; N:服装纺织皮革制造业; O:仪器仪表制造业。

模型分类准确率表现较为稳定), 测试准确率为  $72.5\% \pm 0.5\%$ . 使用测试集混淆矩阵作为模型的分

类效果评价指标(图 3), 图中空心方框区域表示预测错误个数大于 30. 尽管该模型分类准确率不够理想, 但通过混淆矩阵及图 3 的空心方框区域标记可以看出, 其误分类情况多发生在集群中关联性更高的行业中. 由于其本身产业集群下的很多企业产品业务相对繁多, 且有些行业在定义上就比较宽泛, 在行业标签标注时, 很容易产生误差, 导致标签误判. 但该模型对不同关系下的邻居节点信息泛化表达能力依然可以通过训练集群网络各标签下的企业关系传播权重得到有意义的企业嵌入表示.

为评估嵌入的可靠性, 本文列出了嵌入分类后各行业下企业数量, 使用式(2)计算各行业下企业嵌入向量的方差, 其具体结果见表 2(按方差降序排列).

$$\sigma_a^2 = \frac{1}{q} \sum_{i=1}^q \left[ (\mathbf{v}_i - \bar{\mathbf{v}}_a^k) \cdot (\mathbf{v}_i - \bar{\mathbf{v}}_a^k) \right], \quad (2)$$

式中:  $q$  为某行业  $k_a$  下的企业数量.

从表 2 数据可见, 各行业企业数量基本符合制造业集群现状. 橡胶和塑料制品业行业方差最大, 说明其在制造业集群中的业务范围最广; 而非金属建材制造业的行业方差最小, 这与其在集群中

表 2 各行业企业数量及嵌入方差

| 行业类别       | 企业数量 $q$ | 方差 $\sigma_a^2$ |
|------------|----------|-----------------|
| 橡胶和塑料制品业   | 4 887    | 0.082 4         |
| 电器制造业      | 7 856    | 0.067 4         |
| 模具制造业      | 4 519    | 0.065 1         |
| 仪器仪表制造业    | 2 374    | 0.053 7         |
| 金属冶炼加工业    | 1 047    | 0.053 7         |
| 汽车及其配件制造业  | 5 808    | 0.047 8         |
| 通信电子元件制造业  | 4 036    | 0.043 2         |
| 造纸印刷业      | 2 834    | 0.042 2         |
| 塑机及机床制造业   | 1 708    | 0.040 8         |
| 金属制品制造业    | 5 936    | 0.039 2         |
| 服装纺织皮革制造业  | 4 397    | 0.038 3         |
| 化学及纤维制品制造业 | 2 052    | 0.037 8         |
| 文体用品制造业    | 3 698    | 0.037 4         |
| 其他专用设备制造业  | 6 746    | 0.036 7         |
| 非金属建材制造业   | 1 363    | 0.021 7         |

的行业特殊性相吻合. 总体来看, 方差排在前几位的行业事实上在制造业集群中也都有所关联; 方差排在后几位的行业, 在制造业集群的生产业务中更偏向于业内独立运营.

图 4 展示了企业嵌入经 TSNE 降维后的分布情形. 图中 A 体现塑机及机床制造业行业的聚集性和交叉性, B 体现模具行业与其上下游企业所在行业

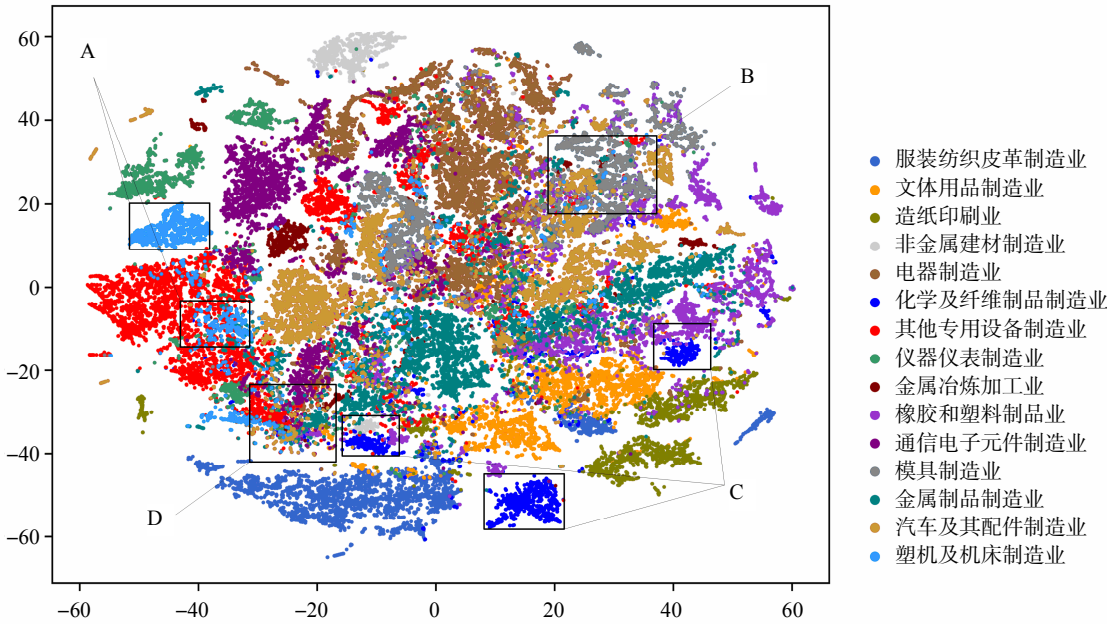


图 4 产业集群企业嵌入行业分类可视化

(塑机及机床制造业、汽车及其配件制造业、橡胶和塑料制品业等)的交叉性, C 体现化纤制品制造行业的聚集性和与某些行业的近似性, D 区域的向量聚类却拥有多种分类, 体现了宁波地区产业集群内少量企业的多行业复合性。

组合聚集、交叉、近似和复合 4 种基本现象后, 形成产业集群的普遍状况: 在某区域下的产业集群中, 以产品为中介关系所形成的企业网络, 其部分企业会表现得比较离群, 它们的产品经营范围独特且单一, 在经过模型训练分类后, 此类企业会表现为离群聚类, 如图 4 中的非金属建材制造业、服装纺织皮革制造业、造纸印刷业和文体用品制造业。而在集群中, 大部分企业的产品经营范围相对宽泛, 一家上规模的模具企业不只经营模具产品, 同时还可能经营塑料制品、汽车内饰配件等其他产品。同理, 一些产品如五金件、模具、各类零件等产品也会被不同行业下的企业所兼营, 导致图 4 中表现出各行业企业向量跨度大和聚类不分明特点。

### 3.3 产业集群关联网络嵌入应用分析

通过模型分类效果评估和企业嵌入可视化分析, 进一步从嵌入角度量化分析宁波地区产业集群企业地理分布特征及集群局部区域特性。

图 5 和图 6 显示了塑机及机床制造业和服装纺织皮革制造业两类行业属性下, 相似度大于 0.9 的企业分布热力地图。受宁波地区特殊地理环境的影响, 各类企业主要集中在各县区靠中心位置, 其余则沿道路零散分布。图 5 和图 6 在热点数量上存在明显差距, 即使考虑了行业外有些复合型企业可能的相关性, 不同行业属性下的相关企业数量

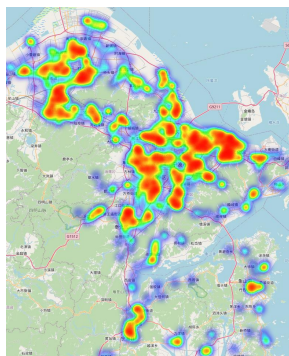


图 5 塑机及机床制造业相关企业分布情况

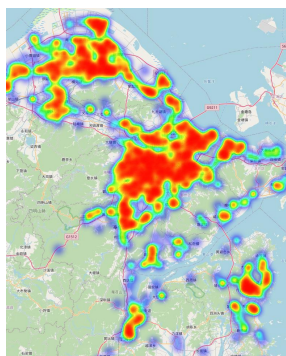


图 6 服装纺织皮革制造业相关企业分布情况

仍然没有发生质的改变。两图的企业分布整体上都由不同大小的斑块状图案构成, 且局部区域上的分布存在较为明显差异。

为进一步分析集群这些行业分布特征, 选取 8 个宁波地区不同行业且具有一定规模的企业作为分析对象, 以各企业所在位置  $d = \{1, 2, \dots, 15\}$  范围内的  $S^d$  为指标, 平滑处理后, 绘制图 7(a) 中的曲线图, 曲线标签结构为行业分类简写-所选企业经营产品举例-该企业所在区域, 两端数字分别显示每个企业在其方圆 1~15 km 范围内的企业数量。

以图 7 为依据, 本文从 4 点对宁波地区产业集群区域特性展开分析。

(1) 不同类型企业所在局部区域 1 km 和 15 km 范围内的企业数量明显不同, 不同区域内的企业密集程度存在差异, 且企业密集程度没有对  $S^d$  指标值产生线性干扰, 导致企业密集程度差异或许

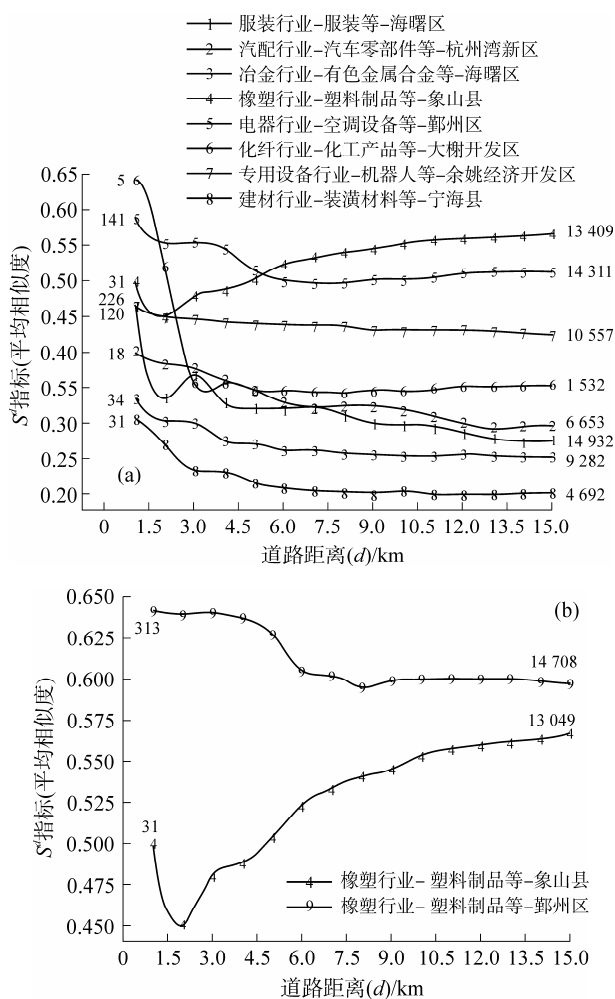


图 7 产业集群不同区域特性指标曲线

是由区域政策、地理环境等因素造成。

(2)图 7(a)中的各条曲线在局部的 2 km 范围内均呈下降趋势,大多数曲线在 5 km 范围内呈现出差异性的变化趋势,5 km 后的变化趋势相对缓和,这一现象与上文热力地图的企业斑块状分布特征吻合。由此推测可能是具有一定规模的企业对周围企业的影响导致业务相似性提升, $S^d$  指标同样也反映这种情形的影响范围,也间接地验证了文献[22]关于集群企业同质性增强的观点。

(3)不同类型的企业  $S^d$  指标有较大差异。如建材行业下的企业,其  $S^d$  指标在周围 15 km 范围内都远低于电器行业下的企业,此差异体现的是企业在集群中的关联程度,所选企业越具有跨行业性质,其  $S^d$  指标相对越高,反之则越低。

(4)图 7(a)中的曲线 4,在其 2 km 外,随着局部区域范围的扩大而逐渐上升。针对该现象,本文在图 7(b)中选择另一家来自鄞州区的同类型企业进行对比(曲线 9)。可以看出,曲线 9 与其他多数曲线变化趋势相似。经过对源数据的筛查,确定曲线 4 的企业来自象山县定塘镇边缘,后续曲线的上升是由于镇中心存在较多类似企业。两个企业处于不同的区域位置,其  $S^d$  指标大小完全不同,这也符合宁波地区该类型企业的分布特点,并且鄞州区企业数量和种类更繁多,相对业务关联程度也比宁波边缘地区更高。

综上分析,本文在通过嵌入学习捕获产业集群关联网络结构特征的基础上,利用企业嵌入和企业地理信息,可以有效反映出集群各行业分布特征,并能分析集群内不同的区域特性。

#### 4 总结与展望

本文以一种新颖的分析方法对产业集群关联网络进行了相关研究,首先阐述了研究思路和研究方法的可行性,通过对互联网企业公开数据的预处理和关联网络的构建及评估,使用 R-GCNs 模型,将以产品为显性中介关系的产业集群关联网络转入至向量空间,在对嵌入结果的质量进一步评估后,通过调用百度地图 API 获取企业地理位置编码,从集群视角分析行业的分布特征,结合企业间距离信息构建区域企业综合关联程度分析评估指标,探索了产业集群关联网络的嵌入在分析产

业集群区域特性相关问题中的初步应用。

同时,本文研究也存在着一定不足。如在人工进行企业行业分类标签制作时,所制作标签存在一定误差,导致模型分类结果准确率还有较大提升空间。在本文相关研究中,尝试通过模型训练挖掘企业的临近特征,但结果也并不理想,因此后续将对产业集群的地理邻近特征做出进一步分析研究。总体来说,本文的研究方法将对产业集群组织网络相关研究带来更多启发,同时未来的研究也将聚焦于更全面、更细粒度的企业业务知识、企业属性、企业技术及高校研究院等产业集群组织网络相关信息,进行更多维度的嵌入研究和关联性挖掘分析。

#### 参考文献:

- [1] 赵炳新,杜培林,肖雯雯,等. 产业集群的核结构与指标体系[J]. 系统工程理论与实践, 2016, 36(1):55-62.
- [2] 吴宇,余军合,战洪飞,等. 数据驱动的产业集群辨识及空间关联性分析[J]. 科技与经济, 2020, 33(4):51-55.
- [3] Sozinova A A, Okhrimenko O I, Goloshchapova L V, et al. Industrial and innovation clusters: Development in Russia[J]. International Journal of Applied Business and Economic Research, 2017, 15(11):111-118.
- [4] Brachert M, Titze M, Kubis A. Identifying industrial clusters from a multidimensional perspective: Methodical aspects with an application to Germany[J]. Papers in Regional Science, 2011, 90(2):419-439.
- [5] Papagiannidis S, See-To E W K, Assimakopoulos D G, et al. Identifying industrial clusters with a novel big-data methodology: Are SIC codes (not) fit for purpose in the internet age?[J]. Computers & Operations Research, 2018, 98(10):355-366.
- [6] Kozonogova E V, Kurushin D S, Dubrovskaya J V. Computer visualization of the identify industrial clusters task using GVMap[J]. Scientific Visualization, 2019, 11(5):126-141.
- [7] 郑荣,魏明珠,高志豪,等. 基于 SCAN-CPM 的产业新兴技术识别与演化路径分析:以新能源汽车产业为例[J]. 图书情报工作, 2022, 66(11):100-109.
- [8] 张岩,朱艳硕,吴曼,等. 中小企业智能制造产业集群创新扩散动力学复杂网络模型仿真[J]. 青岛科技大学学报(社会科学版), 2019, 35(1):35-40.
- [9] Lei H S, Huang C H. Geographic clustering, network relationships and competitive advantage two industrial clusters in Taiwan[J]. Management Decision, 2014, 52(5):

- 852-871.
- [10] 高菲, 俞竹超, 江山. 多核式中卫型产业集群的网络结构分析——以沈阳装备制造业集群为例[J]. 产经评论, 2014, 5(5):63-77.
- [11] 关峻, 徐泽磊, 邢李志. 全球产业集群发展关联网模型研究——以汽车产业集群为例[J]. 科技进步与对策, 2017, 34(17):72-79.
- [12] Titze M, Brachert M, Kubis A. The identification of regional industrial clusters using qualitative input-output analysis (QIOA)[J]. *Regional Studies*, 2011, 45(1):89-102.
- [13] 李经成, 黄春晓. 创新型企业城市内部空间分布及组织逻辑——以南京市为例[J]. 经济地理, 2022, 42(7):93-105.
- [14] 沈体雁, 李志斌, 凌英凯, 等. 中国国家标准产业集群的识别与特征分析[J]. 经济地理, 2021, 41(9):103-114.
- [15] Bordes A, Usunier N, Garcia-Durán A, et al. Translating embeddings for modeling multi-relational data[C]// *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada. New York: ACM, 2013:2787-2795.
- [16] Song D D, Zhang F, Lu M Y, et al. DTransE: Distributed translating embedding for knowledge graph[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2021, 32(10):2509-2523.
- [17] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations[C]// *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA. New York: ACM, 2014:701-710.
- [18] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2022-10-16]. <https://arxiv.org/abs/1609.02907>.
- [19] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]// *European Semantic Web Conference*. Cham: Springer, 2018:593-607.
- [20] Nayak T, Majumder N, Goyal P, et al. Deep neural approaches to relation triplets extraction: A comprehensive survey[J]. *Cognitive Computation*, 2021, 13(5):1215-1232.
- [21] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]// *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019:4171-4186.
- [22] 童旭红. 从企业的进退机制看集群发展——以高新技术企业为例[D]. 成都: 西南财经大学, 2005.

## Industrial cluster analysis method based on graph embedding technology

NIU Shiyin, YU Junhe\*, ZHAN Hongfei, WANG Rui, LIU Dongyang

( Faculty of Mechanical Engineering & Mechanics, Ningbo University, Ningbo 315211, China )

**Abstract:** Previous research on industrial clusters has problems such as the difficulty in data acquisition or one-sidedness of data dimensions, and poor scalability of analysis method in traditional complex network theory. To address these problems, a vector space analysis method is proposed, which is based on a large amount of unstructured data on the Internet, and employs a graph embedding model to extract the features of cluster associated network. The public data from Internet are used to construct industrial cluster associated network, and part of node labels are designed based on standards of industry classification and purposes of analysis, with which the model of relational graph convolutional networks (R-GCNs) is used to learn the characteristics of industrial clusters from the level of product association. According to the embedding and geographic location information of enterprises in the industrial cluster, the application method of cluster network embedding is also proposed. Through the experimental analysis and demonstration of the relevant data of the manufacturing industry cluster in Ningbo, the effectiveness of the graph embedding analysis method in the quantitative analysis of the characteristics of the industrial cluster associated network is verified.

**Key words:** industrial cluster; associated network; relational graph convolutional networks; regional characteristics

( 责任编辑 章践立 )