文章编号:1005－2542(2020)04-0629-10

基于自然语言处理与深度学习的信用贷款评估模型

赵雪峰[1]， 吴伟伟[1]， 时辉凝[2]
(1.哈尔滨工业大学 管理学院，哈尔滨 150001；2.广东外语外贸大学 会计学院，广州 510006)

【摘要】针对信用贷款评估模型存在特征预处理复杂、受主观因素干扰、准确率较低等现象，提出一种新模型。该模型首先组建连续性信贷特征文本数据，然后使用 Word2Vec 算法进行词向量化后通过词嵌入层衔接卷积神经网络(CNN)进行评估，通过 Keras 框架并依据 2008～2018 年的银行个人信贷数据进行实证分析。结果表明，新模型的总体评估准确率高达 91.7%，无需对缺失特征进行处理并可直接评估，且评估准确率更优异，达到 85.8%。新模型将离散型的信贷特征转变为连续性文本，降低特征预处理复杂度，结合 Word2Vec 与自然语言处理实现直接评估缺失信贷特征的目的，并基于 CNN 优异的特征分析能力最终提高信贷评估模型鲁棒性，进一步改善信用贷款评估模型中存在的部分问题，同时避免评估中主观因素的干扰。

关键词：自然语言处理；卷积神经网络；深度学习；信用贷款
中图分类号：C 932.1　　文献标志码：A　　DOI：10.3969/j.issn 1005－2542 2020 04 002

Credit Loan Evaluation Model Based on Natural Language Processing and Deep Learning

ZHAO Xuefeng[1]， *WU Weiwei*[1]， *SHI Huining*[2]
(1.School of Management, Harbin Institute of Technology, Harbin 150001, China;
2.School of Accounting, Guangdong University of Foreign Studies, Guangzhou 510006, China)

【Abstract】In view of the fact that current credit evaluation model has the characteristics of complex preprocessing, subjective factors interference, and low accuracy, a novel model is proposed, which first constructs text data of continuous credit characteristics, and uses the Word2Vec algorithm for word vectorization, and then evaluates by connecting convolutional neural network (CNN) with word embedding layer. Besides, an empirical analysis is conducted through the Keras framework and based on the personal credit data of the bank from 2008 to 2018. The results show that the overall evaluation accuracy of the novel model is as high as 91.7%. Missing features can be evaluated directly, i.e., missing features need not processed, with an accuracy rate of 85.8%. The novel model transforms discrete credit features into continuous text, which reduces the complexity of feature preprocessing. The combination of Word2Vec and natural language processing achieves direct assessment of missing credit features. The excellent feature analysis capabilities based on CNN improves the robustness of the credit evaluation model, improves some of the problems, and avoids subjective factors in the current credit evaluation model.

Key words: natural language processing; convolutional neural network (CNN); deep learning; credit loan

收稿日期：2019-05-13　修订日期：2019-08-02
基金项目：国家自然科学基金资助项目(71472055)；国家社会科学基金重点项目(16AZD0006)；中央高校基本科研业务费专项资金资助项目(HIT.NSRIF.2019033)
作者简介：赵雪峰(1993-)，男，硕士生。研究方向为技术预测与创新管理。
通信作者：吴伟伟(1978-)，男，教授。E-mail:46254086@qq.com

信用贷款是银行或其他金融机构向资信良好的企业或个人发放无需提供担保的人民币贷款方式。近年来，国家不断出台政策，鼓励金融机构加大贷款投放力度，使得信用贷款更好地解决了“贷款难”问题，有力地支持了各行各业的经济发展，其功能、效益也日趋显现[1-2]。据《2018—2025 年中国信用贷款行业现状研究分析及发展趋势预测报告》显示，随着市场经济的快速发展，短中长期贷款额进一步增加，且无论企业还是个人对信贷需求都有所增长，特别是随着国内居民逐渐接受如提前透支消费、贷款消费等消费观念，使得个人贷款业务获得更快发展，个人信贷业务也因较优惠的贷款条件获得商业银行的青睐。作为贷款金融机构，完善信用贷款制度不仅是认真贯彻落实国家政策的重要举措，也是进一步巩固金融机构在金融市场地位的必然要求[3-4]。但综合而言，目前信用贷款体系还有待进一步完善，特别是对于信用贷款的评估模型。目前对于信用贷款评估模型的研究多从信用贷款评估结果与信用贷款特征两方面着手。

对信用贷款评估结果的研究有：石宝峰等[5]建立了由年龄、非农收入/总收入等 13 个指标组成的农户小额贷款信用评级指标体系，在此基础上，利用熵权法求解评价指标权重，构建了基于 ELECTRE（消去与选择转换评价）的农户小额贷款信用评级模型；Prager 等[6]基于马尔可夫模型对贷款进行建模评估；葛兴浪等[7]利用偏相关系数和 wald 统计量遴选出一套最能代表企业信息的指标体系作为解释变量，以外部评级作为被解释变量，找出企业信息与其外部评级之间的动态权重映射关系；肖斌卿等[8]运用线性判别分析、二项逻辑回归和 10 种基于不同学习算法的 BP 神经网络模型构建内部信用评级模型，并在评级指标体系中加入宏观经济变量的方法；曹勇等[9]以对 m 个行业的贷款权重为优化变量，以银行整体信贷资金的差异系数最小化为目标函数，建立了基于违约状态联合概率的商业银行信贷资金行业间优化配置模型；石宝峰等[10]通过将偏好顺序结构评估法（PROMETHEE-II）引入商户小额贷款信用评级，构建了基于 PROMETHEE-II 的小额贷款信用评分模型；李丹[11]在渐进单因子模型基础上，引入违约概率与违约损失之间的正相关性，给出一个容许违约损失服从任意分布的信用风险经济资本测度方法；赵志冲等[12]以信用差异度和违约金字塔为标准，构建非线性规划模型划分信用等级；卞世博等[13]假设开放式信用债券型基金的资金净流入为一个随机过程，基金的投资目标为基于最终财富的期望效用最大化，研究了基金如何对信用债券以及银行存款进行最优投资，并利用鞅方法给出了此优化问题的解析解。

对信用贷款特征的研究有：王星等[14]利用商业银行实际经营数据，构建了客户特征对信用卡业务盈利水平影响的结构方程模型，研究了复杂客户特征共同作用于信用卡业务盈利水平的影响路径及程度；熊志斌[15]在分析 CFS 方法基础上引入 Gebelein 最大相关系数 GMC，提出一种非线性相关性特征选择方法——基于 Gebelein 最大相关性特征选择方法（GCFS），可有效地识别变量间的非线性相关关系，更真实估计特征间相关系数大小，从而筛选出最优特征子集。

综合而言，由于信贷评估关乎到经济稳定性、社会财力分配均匀等经济与社会问题，因而不管是对信用贷款特征的研究，还是对信用贷款评估结果的研究，都衍生出一批优异的模型和方法。如针对农户小额贷款的熵权法和消去与选择转换评价模型、在信贷评估中添加了宏观经济变量的 BP 神经网络模型等，不仅节省人力投入，也有效提高了信贷评估的贷款效率，对信贷评估的发展和完善做出了巨大贡献，具有重要的研究参考价值。

但客观而言，信贷评估依然受到多种因素干扰，干扰因素主要分为信贷特征、信贷模型和人为主观因素等 3 个方面。由信贷特征和信贷模型可以看出，由于信贷特征都是基于离散型的原因，使得多数信贷模型对于信贷特征处理较为繁琐，特别是信贷特征相似度、信贷特征定性到定量的计算等，进而影响模型的最终评估准确率；由信贷特征和人为主观因素可以看出，实际评估难免遇到缺失少量信贷特征的数据，放弃该类数据违背了最大化利用已有数据的原则，但多数缺失特征的补充方法标准不一，且对于补充方法的选择上也具有较多的人为主观因素，直至影响评估过程公正公平的原则。

针对上述现象，本文以自然语言处理为基础，使用 Word2Vec 算法并结合卷积神经网络的方法，将其作为信贷评估模型（Credit Loan based on Word2Vec and Convolutional Neural Network，WV-CNN）。WV-CNN 模型首先组建连续性特征文本数据并使用自然语言方法进行预处理；然后，基于 Word2Vec 对所述连续性特征文本数据进行词向量化；最后，通过卷积神经网络训练词向量数据并完成评估。实验结果表明：WV-CNN 一方面简化了特征处理流程，实现了特征处理与信贷评估端到端的形式，同时也提高了模型评估的准确率；另一方面，

WV-CNN 可直接对缺失特征的数据进行信贷评估，无须对缺失特征进行补充操作，减少时间投入成本和主观因素对客观评估的影响。

1　模型构建过程

按照信贷特征在 WV-CNN 模型的处理过程，可统一分为信贷特征文本预处理、信贷特征文本词向量化、模型训练及信贷评估 3 个阶段。本节先引入背景知识和参数，然后依次构建出 3 个阶段。

1.1　背景知识及参数定义

（1）Huffman 二叉树。树是数据元素（又称为结点）按照分支关系组织出的一种非线性数据结构，若干颗树的集合称为森林。二叉树是每个结点最多只有两个子树的有序树，两个子树分别为左子树和右子树。若存在一颗二叉树的路径长度最小，则称为 Huffman 二叉树[16]。

（2）Huffman 编码。结合 Huffman 二叉树，单词为叶子结点，各叶子结点的权值通过 Huffman 编码表现[16]，本文用 0、1 码的不同排列表示单词。

词向量。若干不重复的词组成的集合称为词典，用 ϕ 表示，对词典 ϕ 内的任意词 ω，指定一个固定长度的实体向量 $\boldsymbol{V}(\omega)\in R^m$，则 $\boldsymbol{V}(\omega)$ 称为 ω 的词向量，m 为词向量长度。词向量是为了将特征从定性转变为定量，为后续模型研究做好准备。

（3）统计语言模型。自然语言具有上下文相关的特性，而统计语言模型就是使用概率预估出衡量一个语句是否合理的语言相关特性模型[17]。假如句子 S 由 n 个词 ω 组成，即 $S=\{\omega_1,\omega_2,\cdots,\omega_n\}$，则统计语言模型的目的是求出 $P(S)$。

（4）参数定义。ϕ、ω、c、$\boldsymbol{V}(\omega)$ 和 m 分别为词典、词典内的任意某单词、任意某单词前或后的单词数、任意单词所对应的词向量以及词向量维度；$\mathrm{Context}(\omega)$ 为词 ω 与其周边词的集合；ζ 表示语料，语料是包含了文本所有内容的集合，存在重复的单词，上述词典 ϕ 是剔除语料 ζ 中重复的词而形成的集合；p^ω、l^ω 为根结点出发到达单词 ω 对应叶子结点的路径、路径 p^ω 中包括结点的数量；$p_j^\omega\{p_1^\omega,p_2^\omega,\cdots,p_{l^\omega}^\omega\}$，$p_j^\omega$ 表示在路径 p^ω 内，每个结点的名称，其中，p_1^ω 为根结点，$p_{l^\omega}^\omega$ 为词 ω 所在位置的对应结点；$d_{j\epsilon}^\omega\{d_2^\omega,d_3^\omega,\cdots,d_{l^\omega}^\omega\}$，$d_j^\omega$ 表示在路径 p^ω 内，第 j 个结点对应的 Huffman 编码，根结点无编码，$d_{l^\omega}^\omega$ 为词 ω 的编码；$\boldsymbol{\theta}_{j\epsilon}^\omega\{\theta_1^\omega,\theta_2^\omega,\cdots,\theta_{l^\omega-1}^\omega\}$，$\boldsymbol{\theta}_j^\omega$ 为路径 p^ω 内，第 j 个非叶子结点对应的向量，因为词 ω 是叶子结点，故无对应向量；$p(\omega\mid\mathrm{Context}(\omega))$，根据 Huffman 二叉树理论，路径 p^ω 存在 $l^\omega-1$ 个分支，每个分支的每次分类都涉及到概率判断，$p(\omega\mid\mathrm{Context}(\omega))$ 即表示所有分支概率的乘积。

1.2　信贷特征文本预处理

传统的文本预处理包括清理和标准化两个阶段，但是由于信贷特征多为单一变量，且本文研究的信贷特征基于文本形式；又由于信贷特征文本具有连续化和非结构化的特性，因而包含多种噪声。因此，本文的预处理包括提取特征、构建文本、清理及标准化 4 个阶段。具体步骤如下：

（1）收集数据并提取影响评估结果的信贷特征。郭小波等[18]在微企业信用风险的识别因子研究中指出，企业财务、企业定性指标以及与企业业主有关的指标等都可纳为信贷特征；肖斌卿等[19]认为，信用评估需考虑的因素包括借款企业经营环境、所有制与经营权、管理水平、营运价值、盈利能力以及风险程度等因素。因此，信贷特征的选择需根据收集的数据进行综合考虑。

（2）构建信贷特征文本。按照时间序列和统计语言模型规律，将信贷特征构建为文本形式，如‘餐饮业’‘相对控股’公司因为‘融资’原因，使用‘抵押’门面店和店内硬件设备的方式贷款‘100 万元’，银行对贷款公司进行调查显示，资产总额为‘1 500 万元’，得出信用评价评级‘AA’。

（3）清理信贷特征文本。包括创建词典 ϕ、中文转英文、切片、去停用词、去标点、去非词典词以及重组语句 7 个步骤。

（4）标准化。将清理后的信贷特征文本分为训练集与测试集，同时确定对应的标签集。

1.3　基于 Word2Vec 的词向量化

Word2Vec 是根据特征文本生成词向量（简称词向量化）的高效算法，可有效地解决维数灾难和词汇鸿沟的困扰[20-21]。本文根据 Huffman 二叉树思想，并结合信用贷款实际情况，在 WV-CNN 模型中使用基于 Word2Vec 的 CBOW（Continuous Bag-of-Words）模型进行词向量化操作。

CBOW 模型由输入层、投影层和输出层组成，通过求解词典 ϕ 内任意词 ω 的词向量完成信贷特征文本的词向量化，CBOW 模型运作流程如图 1 所示。图 1 展示了词向量化模型进行信贷特征的词向量化过程，揭示了信贷特征词向量化过程与社会经济发展的联系性。当词向量化过程越复杂，表示信贷数据更全面，社会经济发展趋于多元性；当词向量化过程较平缓，社会经济发展趋于公平性。同时，图 1 还展示了 CBOW 模型词向量化的内部处理过程，

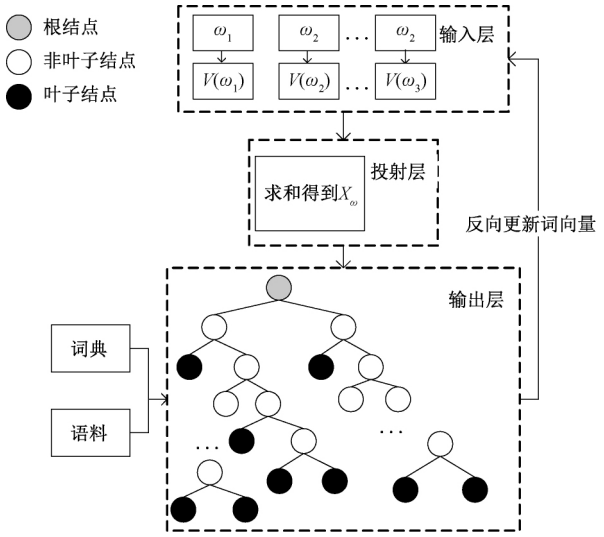

图 1 CBOW 模型运作图

方便后期进一步基于信贷特征数据对当前社会经济状态、贷款机构财务等研究进行信贷特征数据的关联性判断。

CBOW 模型输入层根据词 ω 在信贷文本的出现位置 Context(ω)，选择词 ω 前后 $2c$ 个词所对应的词向量作为输入，如信贷文本 A 为：张先生从银行贷款 12 万元，打算两年后分 24 个月分期归还银行，则将词 120 000 作为词 ω，词 ω 出现位置标记为 Context(ω)，经过去停用词和分词处理后，分别得到在 12 000 元前的 $2c$ 个词为：ω_1 = 张先生，ω_2 = 银行，ω_3 = 贷款，得到 12 000 元后的 $2c$ 个词为：ω_1 = 打算，ω_2 = 两年，ω_3 = 24 月，ω_4 = 分期，ω_5 = 归还，ω_6 = 银行。进一步，分别假设每个词的词向量为 $\boldsymbol{V}$(Context(ω))，即输入为 V(Context (ω)$_1$)，V(Context (ω)$_2$)，…，V(Context (ω)$_{2c}$)，每个词向量具有相同的维度 m。

由此结合图 1，CBOW 模型首先遍历出信贷文本内的词语，将每个词语假设为已知的词向量，通过求和与反向迭代的方式更新词向量。因为多数信贷特征处理模型在研究信贷特征时，往往是基于离散的角度对信贷特征进行处理，如去除缺失值、填充众数平均数等，或基于降维的方式减少信贷特征之间的联系。而 CBOW 模型直接将整个信贷文本作为信贷特征进行整体处理，没有人为的主观干预并简化了信贷特征前期处理过程，可进一步提高对信贷结果的评估准确率。

投射层接收所有词向量并做累加求和操作：

$$X_\omega = \sum_{i=1}^{2c} V(\text{Context}(\omega)_i) \tag{1}$$

当进行结点分类时，约定分到二叉树左边为负类，分到右边为正类，则由 sigmoid 函数知，结点被分到正类的概率为

$$\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}) = \frac{1}{1+\mathrm{e}^{-\boldsymbol{X}_\omega^{\mathrm{T}}}} \tag{2}$$

输出层基于 Huffman 二叉树和编码原则，联立路径 p^ω 内 $l^\omega - 1$ 个分支的概率乘积和式(1)，构建条件概率公式和条件概率的对数似然函数为：

$$p(\omega \mid \text{Context}(\omega)) = \prod_{j=2}^{l^\omega} p(d_j^\omega \mid \boldsymbol{X}_\omega, \theta_{j-1}^\omega) \tag{3}$$

$$\zeta = \sum_{\omega \in \zeta} p(\omega \mid \text{Context}(\omega)) \tag{4}$$

式中，$p(d_j^\omega \mid \boldsymbol{X}_\omega, \theta_{j-1}^\omega)$ 为二叉树各分支的概率。结合式(2)正类概率并联立负类概率得

$$p(d_j^\omega \mid \boldsymbol{X}_\omega, \theta_{j-1}^\omega) = [\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)]^{1-d_j^\omega}\ [1-\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)]^{d_j^\omega} \tag{5}$$

联立式(3)～(5)，可得条件概率的对数似然函数恒等式为

$$\begin{aligned}\zeta &= \sum_{\omega\in\zeta} \log \prod_{j=2}^{l^\omega} \{[\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)]^{1-d_j^\omega} \times [1-\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)]^{d_j^\omega}\} = \\ &\sum_{\omega\in\zeta}\sum_{j=2}^{l^\omega} \{(1-d_j^\omega)\log[\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)] + d_j^\omega \log[1-\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)]\}\end{aligned} \tag{6}$$

根据式(6)可知，条件概率对数似然函数与花括号内的函数成正比关系，设 $\zeta(\omega, j)$ 为花括号内函数，故只需最优化 $\zeta(\omega, j)$，则可得对数似然函数的最优解：

$$\zeta(\omega, j) = (1-d_j^\omega)\log[\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)] + d_j^\omega \log[1-\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)] \tag{7}$$

由随机梯度上升算法知，最优化函数需求解函数在其参数上的方向梯度，$\zeta(\omega, j)$ 有两个参量 θ_{j-1}^ω 和 $\boldsymbol{X}_\omega$，依次求解如下：

$$\begin{aligned}\frac{\partial \zeta(\omega, j)}{\partial \theta_{j-1}^\omega} &= \frac{\partial}{\partial \theta_{j-1}^\omega}\{(1-d_j^\omega)\log[\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)] + d_j^\omega \log[1-\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)]\} = (1-d_j^\omega) \times \\ &[1-\sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)]\boldsymbol{X}_\omega - d_j^\omega \sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)\boldsymbol{X}_\omega = \\ &[1-d_j^\omega - \sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)]\boldsymbol{X}_\omega\end{aligned} \tag{8}$$

因为 θ_{j-1}^ω、$\boldsymbol{X}_\omega$ 在 $\zeta(\omega, j)$ 函数上是对称关系，所以由式(8) 可得 $\dfrac{\partial \zeta(\omega, j)}{\partial \boldsymbol{X}_\omega}$ 的梯度为

$$\frac{\partial \zeta(\omega, j)}{\partial \boldsymbol{X}_\omega} = [1-d_j^\omega - \sigma(\boldsymbol{X}_\omega^{\mathrm{T}}\theta_{j-1}^\omega)]\theta_{j-1}^\omega \tag{9}$$

由上述推导可知，输出层得到映射层的累加向量 $\boldsymbol{X}_\omega$ 后，基于式(1)、(8) 和式(9) 之间的联系，更新词 ω 的词向量 $\boldsymbol{V}(\omega)$。即每次更新参数 $\boldsymbol{X}_\omega$ 时，$\boldsymbol{V}(\omega)$

也伴随更新，故输出层的输出量为

$$\boldsymbol{V}(\omega) := \boldsymbol{V}(\omega) + \eta \sum_{j=2}^{l^{\omega}} \frac{\partial \zeta(\omega, j)}{\partial \boldsymbol{X}_{\omega}} \tag{10}$$

式中：η 为设定的学习率；$\sum_{j=2}^{l^{\omega}} \frac{\partial \zeta(\omega, j)}{\partial \boldsymbol{X}_{\omega}}$ 表示梯度更新后，将映射层的累加求和分散回每个词向量。

1.4　基于 CNN 的神经网络训练模型

深度学习是基于数据进行表征学习的方法，与传统机器学习算法相比，其具有更优异的特征提取能力，特别是深度学习分支下的卷积神经网络（Convolutional Neural Networks，CNN）模型，在目标检测、自然语言处理等领域发挥了重要作用[22-25]。本文将 CNN 应用于信贷评估领域，在 WV-CNN 模型内构建一维 CNN，将词向量化后的信贷特征文本作为该网络的输入进行训练并评估。

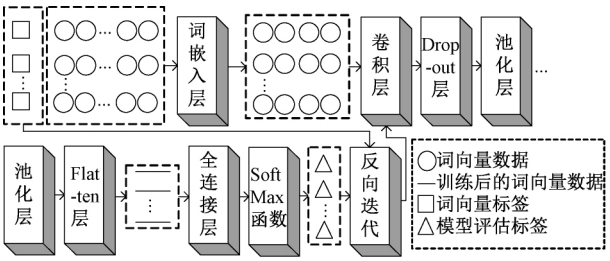

图 2　WV-CNN 模型内 CNN 结构图

由图 2 可知，WV-CNN 模型的 CNN 部分首先将 CBOW 计算出的词向量嵌入至 CNN 输入层，称为词嵌入层。其次，构建多重卷积层和池化层，由于信贷特征在时间维度上以一维向量的形式存在，因而卷积层与池化层的过滤器也是一维向量。同时，为了防止过拟合现象，在卷积层与池化层中加入 Dropout 层；Flatten 层是基于多次卷积池化操作后进行扁平化处理。最后，通过全连接层（Dense 层）输出信贷评估结果。

1.5　基于 Keras 深度学习框架的模型实现

本文基于 Tensorflow 为后端的 Keras 实现 WV-CNN 模型。Keras 是一个高级深度神经网络实现框架，具有模块化、可扩展性、快速部署神经网络等优点[26-27]。图 3 所示为 WV-CNN 模型在 Keras 中的部署图。根据图 3，Keras 根据 WV-CNN 模型的信贷特征文本预处理、词向量化、模型训练与信贷评估 3 个过程，依次调用 numpy、nltk 和 gensim 等处理包进行部署实现。结合图 3，Keras 不仅是高级深度神经网络的实现框架，同时也提供多种模型和函数之间的衔接函数，如 Flatten 数据扁平化处理函数、Dense 数据全连接函数以及 join 语句连接函数等，

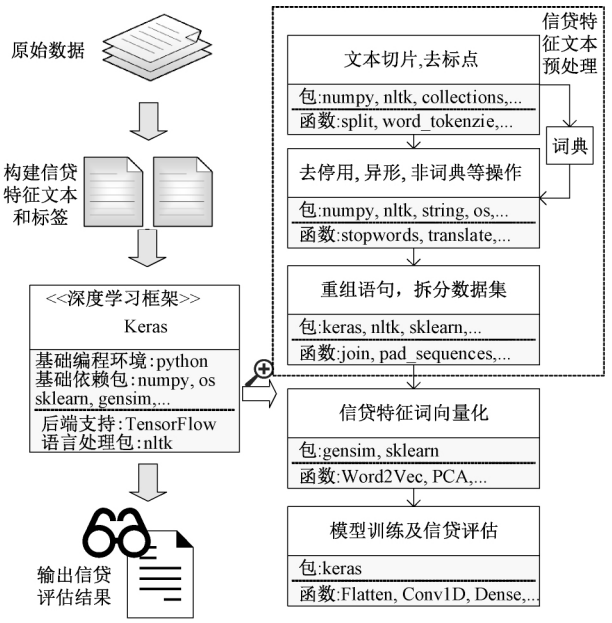

图 3　WV-CNN 模型在 Keras 中的部署图

这类衔接函数可以将文本预处理、词向量化、训练及最终评估所产生的数据进行数据衔接，从而形成整体的 WV-CNN 模型，进一步完善了 WV-CNN 模型处理信贷数据的衔接性问题，提高了信贷评估的流畅性。另外，根据部署图所示，将文本预处理、词向量化、模型训练与信贷评估封装在 Keras 内，不仅简化评估的中间过程，若侧重研究信贷特征的变化过程，同样也可根据部署图先探讨信贷评估的内部认知结构，进而调取信贷特征影响信贷评估结果的自主性，方便后续展开不同信贷特征对整个评估结果影响的研究。

综合而言，WV-CNN 模型由于使用 Word2Vec 算法，因而可将单词进行相似度聚类，相比于其他特征处理方法，词向量转化更贴合实际定性表述，且 CNN 与传统机器学习相比，具有更高效的特征分析与特征关联判断能力。因此，评估结果更高效准确。同时，WV-CNN 模型将 Word2Vec 和 CNN 通过词嵌入层进行衔接，使特征处理与模型评估有效结合，实现了端到端的整体化评估，避免多次操作带来的繁琐冗长。

2　模型实验验证

2.1　数据采集及分析

本文以黑龙江省某商业银行在 2008～2018 年期间个人信用贷款数据为例进行实验。由王娟等[28]根据 2005～2011 年多个地区贷款因素对消费信贷约束及其影响的研究发现，个人受教育程度、年收入和支出等对信用贷款有重要影响；胡毅

等[29]在银行客户贷款违约风险预警研究中指出，应将客户外部因素、客户经营水平及客户交易水平 3 个一级指标作为个人信用贷款的重要考虑因素。因此，结合上述文献，表 1 给出了贷款人基本信息、收支情况和贷款信息 3 个一级指标划分出 11 个信贷特征。

表 1 个人信贷特征及对应解释

自变量所属分类	编号	自变量名称及简写	自变量类别	英语表示
贷款人基本信息	A1	性别，Ge	男	male、man、he
			女	madam、lady、she…
	A2	受教育程度，Ed	研究生	graduate
			本科	bachelor
			本科以下	junior、senior、primary…
	A3	婚姻状况，Ma	已婚	married
			未婚	bachelordom、spinsterhood…
	A4	居住地区，Pa	农村	village、countryside…
			郊区	suburb、outskirts…
			城市	downtown、center…
	A5	家庭人员数量，De	1 个	阿拉伯数字或英语表示
			2 个	
			3 个及以上	
收支情况	A6	年收入，Ai	8 000～120 000 元	同 A5
	A7	年支出，Ci	4 000～78 000 元	同 A5
	A8	信用卡还贷情况，Ch	无信用卡	no credit card
			记录良好	sound credit history
			有逾期未还记录	credit card delinquencies
	A9	雇佣关系，Se	被雇佣	employed…
			自雇佣	self employed…
贷款信息	A10	贷款金额，La	8 000～30 000 元	同 A5
	A11	还款日期，Lt	1～24 月	同 A5

表 1 同时列出各信贷特征的英语表示及各自变量名称的简写，方便后续将离散型数据转变为连续性语句提供参考。在剔除异常和特征缺失严重的数据后，本实验共使用 3 560 组有效数据，其中，成功贷款数量 1 720 组，贷款失败 1 840 组。采用交叉验证法将训练数据集分为 3 000 组，测试数据集为 560 组。

为了方便观察各特征的聚类相关度和特征变化，使用热图对表 1 共 11 个特征进行分析。热图起源于分子生物学领域，主要通过颜色的不同表达出多个特征的全局变化；同时，热图也使用数字大小关系表示特征之间的聚类关系，数字越大，则两类特征之间的相关度越高[30]。在经济领域中，热图一般做前期的关联分析，如影响股票波动的影响因子之间的联系、政府投资前采集多个试验点进行投资分析等。通过热图分析信贷特征之间的聚类相关度，从而剔除聚类相关度高的信贷特征，减少多个相似特征对信贷结果的影响，提高信贷评估的准确率。本文使用 seaborn、matplotlib 包构建特征热图。

由图 4 并结合表 1 各特征简写知，贷款人年收入与贷款人还贷周期的聚类相关度最高，达到 0.57，但多数特征之间相关度不超出 0.2，满足特征之间相关度不超过 0.6 的要求。因此，个人信贷特征可用于后续预处理环节。

2.2 数据预处理

根据理论构建过程知，WV-CNN 模型的操作对象是连续性文本，而非离散型特征，因此需将表 1 的各信贷特征按照语言叙述的规律进行复现。本文使用中英语结合的叙述方法对个人信贷数据进行语言复现，结合表 1 特征值对应的英语表述，表 2 给出部分语言复现后的文本。如表 2 编号 S1 所示，中文信贷文本记录张先生借贷 120 000 元，两年内还款等文本信息，在满足信贷特征不变的前提下，可采用任意的叙述方式（如表 2 中文信贷文本采用第三人称叙述，连续性英语表述采用第一人称叙述）表现借贷过程，采用任意叙述方式是为了真实还原借贷过程，节约大量前期信贷数据整理的评估时间，提高信贷评估效率。

原则上，复现每组数据耗时严重，因此，首先复现出 30 组语言模板，然后基于 python 拆分每组连续性文本并随机打乱文本中语句的顺序后替换特征，完成语言复现操作。但复现后的连续性文本由于存在大量冗余信息，故需进行预处理操作。其中，冗余信息如连续性中文信贷文本中“于、的、且”等无实质性意义的词语，连续性英语表述中标点美元符号和“a、for、in”等停用词，去除这类词或标点符号后，可降低信贷评估过程中计算资源的使用率，

图 4　个人信贷特征的热图分析

表 2　个人信贷数据的部分语言复现

编号	连续性中文信贷文本	连续性英语表述贷款结果	
S1	张先生打算借贷 120 000 元，并计算在两年内还款。张先生 3 年前本科毕业后，从事通信公司会计一职，年薪 85 000 元、每年可存款 40 000 元且无不良信用贷款记录。其两年前结婚且刚有孩子，房子位于市中心	Hi，I am Mr Zhang. I am planning to borrow $120 000 and repay it in 24 months. I have a bachelor degree from Harbin Institute of Technology. After graduation，I work as a clerk in commerciai company for about 3 years. My annual salary is around $85 000. I married Li Na two years ago. We have an adorable baby. Our house located in downtown. The mortgage costs half of my salary and I can save $40 000 every year. Since I have never been overdue the credit repayment，I must have a great credit score.	成功(1)
S2	32 岁的王先生就职于手机 APP 开发公司，担任行政职位，年薪 90 000 元，无资产赤字且有两套房。打算借款 200 000 元，3 年内还清	Dear who it may concerns my is Mr Wang，34 years old working as an IT executive.I am looking to start an IT start up software company developing apps company for mobile phones，I have an annual income of $90 000 when I graduated. I am looking to borrow $200 000 with the intent to return the money within 3 years. I am a reliable person with no unpaid debt to my name. I have a family with tow houses so I can put one up for collateral. I look forward to hearing from you to move forward on the investment，yours respectively Mr Wang.	失败(0)

提高信贷评估的分析效率。因此，WV-CNN 模型首先根据图 3 的操作步骤对连续性文本进行预处理。本文通过 numpy、NLTK 和 collections 等包实现预处理操作，表 3 所示为预处理完成后的部分文本数据。

2.3　模型训练

WV-CNN 模型基于 Word2Vec 算法将预处理后的文本数据词向量化。通过 gensim 包调用 Word2Vec 的 CBOW 模型并通过多次实验验证，最终使用词向量维度为 100、词向量上下文最大距离为 8 的词向量化数据。图 5 所示为词向量化训练完成后，使用 PCA 算法对数据降维并调用 matplotlib 可视化出的部分词向量图。

结合表 3 和图 5 分析知，WV-CNN 模型进行词向量化操作后，使得文本数据中距离相近的单词在空间距离上也更小，如同属类别的单词 institute 与 bachelor，repay 与 repayment 等在图 5 中更接近。

WV-CNN 模型下一步将词向量数据输入 CNN，本实验在 Keras 深度学习框架中构建 CNN，网络模型见图 3。同时，设定模型训练周期为 3 000次，优化函数为 Adam 算法，损失函数为二元交叉熵。最终训练过程如图 6 所示。由图 6 可知，随着训练周期不断增加，WV-CNN 模型的训练准确率虽有小幅不稳定波动，但总体呈现上升趋势；而损失值的变化趋势和训练准确率相反，虽小幅不稳定波动，但总体呈现下降趋势。当 WV-CNN

表 3 预处理后的文本数据

编号	预处理后的中文信贷数据	预处理后的英文文本数据
D1	张先生打算借贷 120 000 元，计算两年还款。张先生 3 年前本科毕业，从事通信公司会计一职，年薪 85 000 元、每年存款 40 000 元，无不良信用贷款记录。两年前结婚刚有孩子，房子位于市中心	mr zhang planning borrow 120 000 repay 24 months. have bachelor degree harbin institute technology, after graduation, work clerk commercial company 3 years. annual salary around 85 000. married li na two years ago. have adorable baby. house located downtown, mortgage costs half salary save 4 000 every year. Since never overdue credit repayment, have great credit score.
D2	32 岁的王先生就职于手机 APP 开发公司，担任行政职位，年薪 90 000 元，无资产赤字，有两套房。打算借款 200 000 元，3 年内还清	dear who may concerns mr wang, 34 years old working it executive.looking start it start up software company developing apps company for mobile phones, have annualincome 90 000 when graduated. looking borrow 200 000 intent return money 3 years. reliable person no unpaid debt.

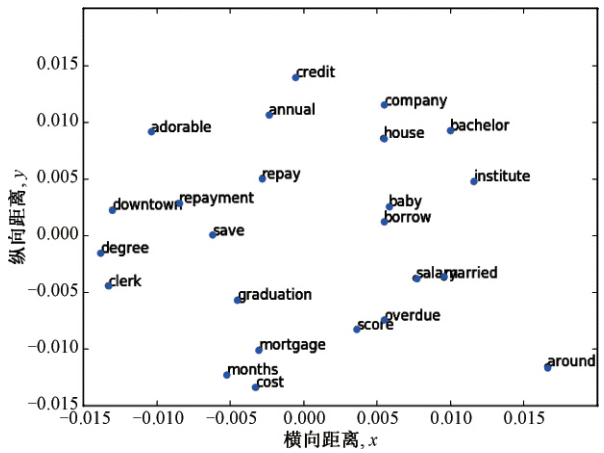

图 5 PCA 降维后的词向量可视化

模型达到 3 000 次的训练周期时，本实验的训练准确率最终稳定在 98.2% 附近，损失函数维持在 0.07。

2.4 模型测试及对比

基于 WV-CNN 模型的训练完成，利用前文所述的 560 组个人信贷评估数据集进行测试，其中 560 组个人信贷评估数据集包括缺失特征数据集 127 组。使用 Logit 函数、支持向量机（Support Vector Machine，SVM）、BP 神经网络和多元线性回归进行对比分析。

根据表 4 分析可知，由缺失特征的数据处理与训练可以看出，WV-CNN 模型与多元线性回归都无需处理可直接评估，降低了人为填充特征的主观因素的干扰，但 WV-CNN 对于缺失特征的鲁棒能力更强，最终测试准确率可达 85.8%。由最终训练与测试结果可以看出，WV-CNN 训练准确率为 98.2%，测试准确率为 91.7%，较其他常用模型都有小幅提高。因此，WV-CNN 模型在提高信贷评估准确率的同时，也优化了对缺失特征的处理能力。

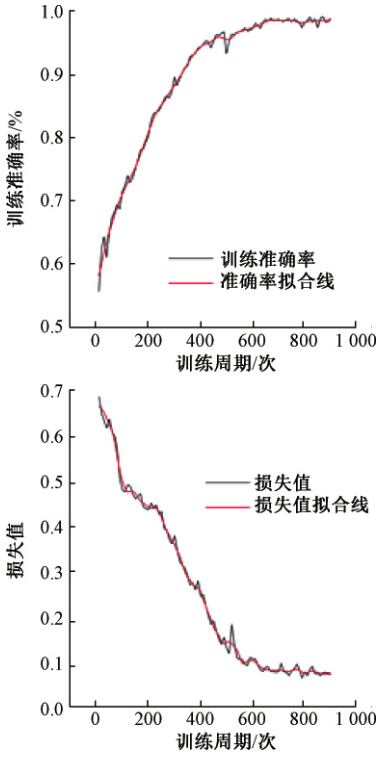

图 6 模型准确率与损失值随训练周期的变化规律

表 4 多模型各参数比对 %

模型	训练准确率	是否可以对缺失特征的数据不处理直接评估	缺失特征处理方法	缺失特征的测试准确率	最终测试准确率
Logit	86.5	否	填充法。填充平均数、众数、中数等	74.1	82.3
BP 神经网络	96.4	否		84.9	90.3
SVN	92.6	否		82.7	88.4
多元线性回归	88.1	可处理		79.4	84.2
		可不处理		72.1	
WV-CNN	98.2	是	无需处理，直接评估	85.8	91.7

3 结　论

信用贷款是提高社会再生产能力的有效手段，而信贷评估模型可实现客观的信用放贷，并减少时间成本的投入。本文使用自然语言处理技术清洗连续性信贷特征文本、Word2Vec 算法进行词向量化，结合卷积神经网络（CNN）构建出 WV-CNN 模型，通过 Keras 框架实现该模型，并以个人信用贷款评估进行实验验证。得出：

（1）特征处理。多数模型首先对特征做定性到定量的改变，然后归一化与标准化，若特征维度过高，还需进行降维处理。而 WV-CNN 模型将特征转换为连续性文本，并基于自然语言技术进行快速清洗，有效减少了时间的投入比。

（2）评估精度，相比于常用模型的测试准确率，如 BP 神经网络 90.3%、SVN 88.4%等，WV-CNN 测试准确率为 91.7%。可见，WV-CNN 模型提高了信用贷款的评估精度，鲁棒性更好。

（3）对缺失特征的处理与评估。多数常用模型采取填充众数、平均数等方法补齐缺失特征后才能进行下一步的评估工作，但 WV-CNN 模型无需人为填充可直接评估缺失特征的数据，减少了人为主观因素对评估结果的干扰；同时，相较于其他常用模型，对于缺失特征的评估准确率也提高至 85.8%。因此，WV-CNN 模型可有效地帮助贷款机构透明、公开、高效的放贷于企业和个人。

参考文献：

[1] ZHOU R X, DU S N, YU M. Pricing credit spread option with longstaff-schwartz and GARCH models in Chinese bond market[J]. Journal of Systems Science & Complexity, 2015, 28(6): 1363-1373.

[2] BO L J, YANG X W. Smooth-pasting property on reflected Lévy processes and its applications in credit risk modeling[J]. Science China (Mathematics), 2014, 57(6): 1237-1256.

[3] LENG A, XING G Y, FAN W G. Credit risk transfer in SME loan guarantee networks[J]. Journal of Systems Science & Complexity, 2017, 30(5): 1084-1096.

[4] LIU X F, ZHANG W, XIONG X. Credit rationing and the simulation of bank-small and medium sized firm artificial credit market[J]. Journal of Systems Science & Complexity, 2016, 29(6): 991-1017.

[5] 石宝峰，王静. 基于 ELECTRE III 的农户小额贷款信用评级模型[J]. 系统管理学报，2018，27(5)：854-862.

[6] PRAGER D, ZHANG Q. Valuation of stock loans under a Markov chain model[J]. Journal of Systems Science & Complexity, 2016, 29(7): 171-186.

[7] 葛兴浪，刘海龙. 基于动态权重的信用评级[J]. 系统管理学报，2019，28(2)：285-293.

[8] 肖斌卿，杨旸，余哲. 小微企业信用评级模型及比较研究[J]. 系统工程学报，2016，31(6)：798-807.

[9] 曹勇，李孟刚，李刚. 基于违约状态联合概率的商业银行信贷资金行业间优化配置模型[J]. 系统管理学报，2018，27(5)：881-894.

[10] 石宝峰，刘锋，王建军. 基于 PROMETHEE-II 的商户小额贷款信用评级模型及实证[J]. 运筹与管理，2017，26(9)：137-147.

[11] 李丹. 基于违约概率与违约损失相关的贷款定价[J]. 系统管理学报，2015，24(1)：56-62.

[12] 赵志冲，迟国泰，潘明道. 基于信用差异度最大的信用等级划分优化方法[J]. 系统工程理论与实践，2017，37(10)：2539-2554.

[13] 卞世博，张熠. 开放式信用债券型基金的最优投资策略[J]. 系统管理学报，2016，25(6)：1023-1028.

[14] 王星，金淳，李延喜. 客户特征影响信用卡业务盈利水平的结构方程模型[J]. 系统管理学报，2018，27(3)：520-528.

[15] 熊志斌. 信用评估中的特征选择方法研究[J]. 数量经济技术经济研究，2016，33(1)：142-155.

[16] 郭建光，张卫杰，杨健. 基于广义规范 Huffman 树的高效编解码算法[J]. 清华大学学报（自然科学版），2009(1)：73-77.

[17] ZHANG Y S, CAO Y D, Statistical language model for Chinese text proofreading[J]. Journal of Beijing Institute of Technology (English Edition), 2003(4): 441-445.

[18] 郭小波，王婉婷，周欣. 我国中小企业信贷风险识别因子的有效性分析——基于北京地区中小企业的信贷数据[J]. 国际金融研究，2011(4)：62-67.

[19] 肖斌卿，杨旸，李心丹. 基于模糊神经网络的小微企业信用评级研究[J]. 管理科学学报，2016，19(11)：114-126.

[20] 潘博，于重重，张青川. 基于词性与词序的相关因子训练的 word2vec 改进模型[J]. 电子学报，2018，46(8)：1976-1982.

[21] ZHOU G, ZHU Q. Kernel-based semantic relation detection and classification via enriched parse tree structure[J]. Journal of Computer Science & Technology, 2011, 26(1): 45-56.

[22] ZHANG Y S, ZHENG J, JIANG Y. A text sentiment classification modeling method based on coordinated CNN-LSTM-attention model[J]. Chinese Journal of Electronics, 2019, 28(1): 120-126.

[23] LIN L K, WANG S Y, TANG Z X. Using deep learning to detect small targets in infrared oversampling images [J]. Journal of Systems Engineering and Electronics, 2018, 29(5): 947-952.

[24] ZHAO Y Y, QIN B, LIU T. Encoding syntactic representations with a neural network for sentiment collocation extraction[J]. Science China(Information Sciences), 2017, 60(11): 7-18.

[25] WANG J, SUN J Q, LIN H F. Convolutional neural networks for expert recommendation in community question answering[J]. Science China, 2017, 60(11): 19-27.

[26] ANGELOVA A, KRIZHEVSKY A, VANHOUCKE V. Pedestrian detection with a large-field-of-view deep network[C]//In Robotics and Automation (ICRA), IEEE International Conference on, [s.l.]: IEEE, 2015: 704-711.

[27] ZHEN J D. Research and application of machine learning on geographic information system [C]// Journal of Artificial Intelligence Practice (2016 Vol.1 Num.1). [s.l.]:[s.n.],2016: 6.

[28] 王娟,李锐. 农户消费信贷约束及其影响——来自10省的样本[J]. 系统工程理论与实践, 2016, 36(6): 1372-1381.

[29] 胡毅,王珏,杨晓光. 基于面板Logit模型的银行客户贷款违约风险预警研究[J]. 系统工程理论与实践, 2015, 35(7): 1752-1759.

[30] SHENG Y, XING H W, LING Y H. Transcriptome analysis of adherens junction pathway-related genes after peripheral nerve injury[J]. Neural Regeneration Research, 2018, 13(10): 1804-1810.