计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(1):141−145 [doi: 10.15888/j.cnki.csa.007744]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 基于 Word2Vec 词嵌入和聚类模型的安全生产事故文本案例分类①

吴德平[1,2], 华　钢[1]

[1](中国矿业大学 信息与控制工程学院, 徐州 221008)
[2](江苏安全技术职业学院 网络与信息安全学院, 徐州 221011)
通讯作者: 吴德平, E-mail: 840892469@qq.com

**摘　要**: 安全生产事故的分析对应急管理能力提升具有重要意义. 通过对安全生产案例的语义分析, 利用 Word2Vec 词嵌入技术和聚类模型, 选用 CBOW+负采样技术实现词向量, 并结合安全生产事故案例分类的数据特点, 通过基于半监督学习的聚类模型算法, 根据事故性质的认定特点, 提出了一种优化初始聚类中心的算法, 并利用 K-means 聚类算法实现安全事故文本案例的分类. 实验表明该方法较好实现安全生产的事故案例分类, 并对安全生产事故的多个维度分析具有很好借鉴意义.

**关键词**: Word2Vec 词嵌入; 聚类; 半监督学习; 安全生产事故; 案例分类

引用格式: 吴德平,华钢.基于 Word2Vec 词嵌入和聚类模型的安全生产事故文本案例分类.计算机系统应用,2021,30(1):141–145. http://www.c-s-a.org.cn/1003-3254/7744.html

## Text Case Classification of Safety Production Accidents Based on Word2Vec Word Embedding and Clustering Model

WU De-Ping[1,2], HUA Gang[1]

[1](School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221008, China)
[2](School of Network and Information Security, Jiangsu College of Safety Technology, Xuzhou 221011, China)

**Abstract**: The analysis of safety production accidents is of great significance to the improvement of emergency management ability. Based on the semantic analysis of safety production cases, Word2Vec embedding technology and clustering model are used, CBOW + negative sampling technology is used to realize word vector, and the data characteristics of safety production accident cases classification are combined, through semi-supervised learning based clustering model algorithm, according to the characteristics of the accident nature, an optimized initial clustering center algorithm is proposed, and K-means clustering algorithm is used to classify the text cases of safety accidents. The experimental results show that the proposed method can realize the classification of accident cases, and can be used for reference in the multi-dimensional analysis of accident.

**Key words**: Word2Vec word embedding; clustering; semi-supervised learning; safety production accidents; case classification

安全生产事关生命财产安全. 通过对安全生产事故划分, 对安全生产事故发生的行业、时间、地域、原因、教训等多个维度展开大数据分析, 采用语义分析技术, 从客观的数据中挖掘安全生产事故的特点与规律, 为安全生产的应急管理提供科学决策具有重要技术意义和参考价值. 本文旨在通过 NLP 技术实现安

① 基金项目: 国家自然科学基金 (51574232)
Foundation item: National Natural Science Foundation of China (51574232)
收稿时间: 2020-05-04; 修改时间: 2020-06-10, 2020-06-23; 采用时间: 2020-07-03; csa 在线出版时间: 2020-12-31

全生产事故大数据分析. 图 1 是安全生产事故分类的实现流程, 通过该流程实现安全生产事故的分类. 准备大量的安全生产案例作为语料, jieba分词工具实现语料分词, 将分词后的单元输入 Word2Vec模型获得词向量, 通过 K-means 聚类对词向量实现聚类实现安全生产事故的分类[1].

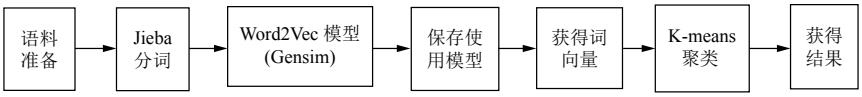


图 1 安全生产事故分类实现流程

## 1 文本表示

文本表示是把字词处理成向量或矩阵, 以便计算机能进行处理. 文本表示是自然语言处理的开始环节. 目前常用的文本表示模型有: 词袋模型、主题模型和词嵌入模型等. 词袋模型主要有 One-Hot (独热编码)、n-gram、TF-IDF. 本例采用 One-Hot 编码.

One-Hot 编码, 又称一位有效编码, 其方法是使用 $N$ 位状态寄存器来对 $N$ 个状态进行编码, 每个状态都有它独立的寄存器位, 并且在任意时候, 其中只有一位有效. 本质上是用一个只含一个 1, 其他都是 0 的向量来唯一表示词语. 表 1 中安全生产事故性质分类为例 (仅考虑死亡人数), 死亡人数 1–9 的一种 One-Hot 编码如表 1.

表 1 One-Hot 编码示意

| 事故性质 | 死亡(人) | 编码 |
|---|---|---|
| 一般事故 | 0 | 0000000001 |
| | 1 | 0000000010 |
| | 2 | 0000000100 |
| 较大事故 | 3 | 0000001000 |
| | … | … |
| | 9 | 1000000000 |

## 2 利用 Word2Vec 实现词向量

### 2.1 分词

分词是实现中文文本词性标注、关键词抽取等功能. jieba 分词包是 Python 中很好的分词组件, 通过加载大量安全生产案例的文本文件, 先基于词典分词, 然后进行词性标注和发现新词, 同时进行关键词提取完成分词. 同时可使用 jieba.suggest_freq('事故', True) 调节单个词语的词频, 使"事故"能被分出来, 提高分词效果[2–5].

### 2.2 CBOW 模型和负采样

Word2Vec 是 Google 推出的用于获取词向量的工具包. Word2Vec 作为神经概率语言模型, 采用两种模型 (CBOW 和 Skip-gram) 与两种方法 (Hierarchical Softmax 和 Negative Sampling) 的组合. CBOW 是根据某个词前面的 $N$ 个词或前后 $N$ 个词计算某个词概率的模型, 其模型如图 2. Skip-gram 是根据某个词计算它前后出现某几个词的各个概率.

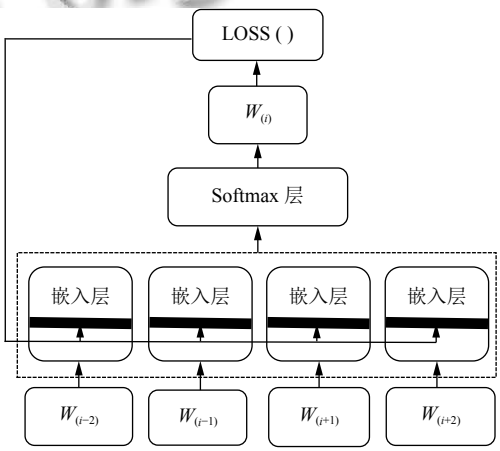


图 2 CBOW 模型

CBOW 模型一般采用 3 层神经网络结构, 分为输入层, 隐藏层 (嵌入层) 和输出层 (Softmax 层). CBOW 模型输入上下文词的 One-Hot 编码, 然后连接一个全连接层, 再连接若干个层, 最后接 Softmax 分类器, 再通过梯度优化和反向传播让模型逼近最小误差就可以得到词向量. 由于神经网络模型训练中生成的词汇往往数万以上, 这大大降低了神经网络的训练速度, 本例选用 CBOW+负采样提高训练速度, 该组合具有运算快的特点. 任何采样算法应该保证频次越高的样本越容易被采样出来. 负采样的本质就是每次让一个训练样本更新神经网络的部分权重. CBOW 模型中词向量的数量大, 神经网络则有庞大的权重数, 不同于原本每个训练样本更新所有的权重, 负采样每次让一个训练样本仅仅更新一部分的权重, 其他权重全部固定, 这样即可以减少计算量, 同时在一定程度上增加随机性, 降低了损失值. 具体代码中 loss 函数定义如下:

loss=tf.reduce_mean(tf.nn.nce_loss(weights=nce_we

ights, biases=nce_biases, labels=train_labels, inputs=embed, num_sampled=num_sampled, num_classes=words_size)).

train_inputs 是中心词, train_label 是该中心词在滑动窗口内的上下文词. train_inputs 中会有连续 $n-1$ ($n$ 为滑动窗口大小) 个元素是相同的. 即同一中心词. embddings 是要学习的词向量的存储矩阵[6–13].

### 2.3 利用 Gensim 实现 Word2Vec

Gensim 是一款开源的第 3 方 Python 工具包, 用于无监督地学习到文本隐层的主题向量表达. 主要用于主题建模和文档相似性处理, 在获取单词的词向量等任务中非常有用. Gensim 中集成有 CBOW+负采样算法, Word2Vec 相关的 API 都在包 gensim.models. Word2Vec 中. 本例中设置的主要参数如表 2.

表 2 Gensim 模型主要参数设定

| 主要参数 | 意义 | 取值 | 含义 |
|---|---|---|---|
| size | 指词向量的维度 | 100 | 大的size需要更多的训练数据, 效果会更好 |
| window | 词向量上下文最大距离 | 10 | 某一个中心词可能与前后多个词相关 |
| min_count: | 词向量最小词频 | 5 | Negative Sampling时负采样的个数 |
| cbow_mean | 上下文词向量和 | 1 | 上下文的词向量的平均值 |
| sample | 高频词汇阈值 | 0.0001 | 范围是(0, 1e−5) |

### 2.4 词向量实验结果

实验基于 Tensorflow 和 Python3.7, 取样 1000 个安全生产案例, 对 100 个词向量可视化结果如图 3. 不难看出这些词向量以安全生产为主题. 主题拟合较好, 如发生、事故、经过、事故、原因等; 同义词检测如年月日时语义相似度很高, 数字的一致性检查很好.

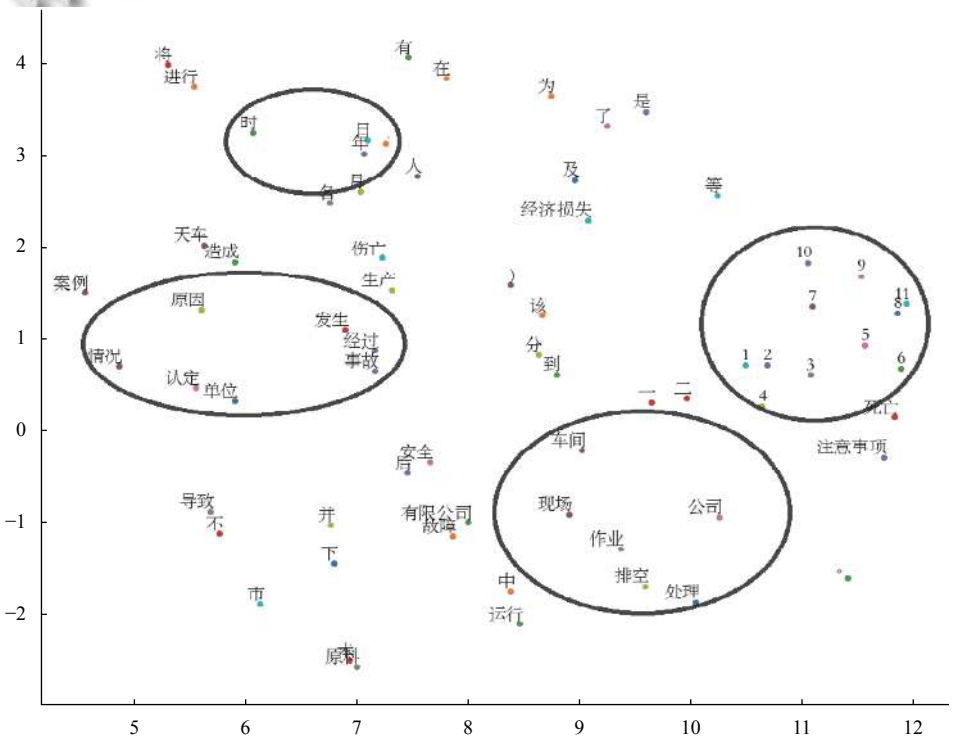


图 3 安全生产案例词向量可视化

## 3 K-means 聚类

### 3.1 K-means 聚类算法

K-means 算法是一种迭代型无监督学习聚类算法, 采用距离作为相似性指标, 从而发现给定数据集中的 $K$ 个类, 且每个类的中心是根据类中所有值的均值得到, 每个类用聚类中心来描 K-means 算法是一个重复移动类中心点的过程, 把类的中心点, 移动到其包含成员的平均位置, 然后重新划分其内部成员. $K$ 是算法计算出的超参数, 表示类的数量; K-means 可以自动分配样本到不同的类, 但是不能决定究竟要分几个类, $K$ 必

须是一个比训练集样本数小的正整数. 对于词向量集 $D=\{X_1, X_2, \cdots, X_m\}$, K-means 算法针对聚类的分类 $C=\{C_1, C_2, \cdots, C_k\}$最小化平方误差为

$$E=\sum_{i=1}^{K}\sum_{x\subseteq c_i}\|x-\mu_i\|_2^2$$

$$\left(\text{其中}\mu_{\mathrm{i}}=\frac{1}{|C_i|}\sum\nolimits_{X\subseteq C_i}\right) \quad (1)$$

其中, $\mu_i$ 是第 $K$ 个聚类的均值向量. 每个类的畸变程度等于该类重心与其内部成员位置距离的平方和. 若类内部的成员彼此间越紧凑则类的畸变程度越小, 反之, 若类内部的成员彼此间越分散则类的畸变程度越大. 求解成本函数最小化的参数就是一个重复配置每个类包含的观测值, 并不断移动类重心的过程. 其算法如图 4.

### 3.2 半监督学习初始化聚类中心

由于安全生产事故分类, 如重大事故指死亡 10 人以上, 30 人以下或重伤 50 以上, 100 以下; 或直接经济损失 5000 万以上, 1 亿元以下. 分类中死亡、重伤人数, 特别是财产损失数值范围很大, 特征空间会变得非常稀疏. 为了解决这个问题, 可通过线性回归模型, 利用半监督学习, 即用已有的词向量确定伤害与死亡、重伤较少人数 (取 30 人以下) 的关联度 $X_{1i}$、$X_{2i}$ 和作为标签, 令相应的权重值分别为 $W_{1i}$、$W_{2i}$, 把经济损失与伤害人数关联度 $B_i$ 视为偏移量, 线性回归的预测函数为:

$$f(x)=W_{1i}X_{1i}+W_{2i}X_2i+B_i \quad (2)$$

利用已有的样本训练式 (2) 可确定相应的学习参数, 如表 3. 如对于特大事故, 利用学习好的参数 $W_{1i}$、$W_{2i}$, 再利用预测函数 (2) 和大量样本确定 $X_{1n}$、$X_{1n}$ 和 $B_n$[14–16].

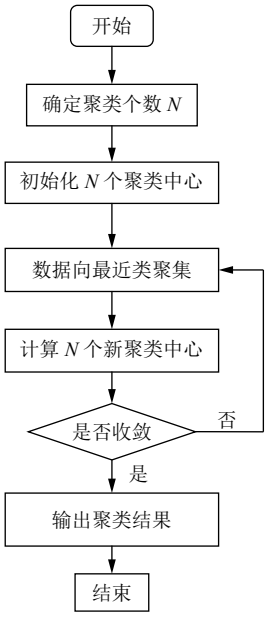


图 4 K-means 聚类算法流程图

表 3 事故性质与关联参数

| 事故性质 | 伤害人数 | 死亡与人数关联度$X_{1i}$ | 重伤与人数关联度$X_{2i}$ | 经济损失与伤害人数关联度$B_i$ |
|---|---|---|---|---|
| 一般事故 | 0 | 0.849 202 871 322 631 8 | 0.850 775 837 898 254 4 | $B_0$ |
| | 1 | 0.822 886 436 462 402 3 | 0.849 523 842 334 747 3 | $B_1$ |
| | 2 | 0.815 686 931 610 107 4 | 0.846 503 615 379 333 5 | $B_2$ |
| 较大事故 | 3 | 0.795 924 186 706 543 | 0.842 493 653 297 424 3 | $B_3$ |
| | … | … | … | … |
| | 9 | 0.730 000 298 023 223 9 | 0.829 389 810 562 133 8 | $B_9$ |
| 重大事故 | 10 | 0.719 612 927 436 828 6 | 0.826 938 390 731 811 5 | $B_{10}$ |
| | … | … | … | … |
| | 29 | 0.700 842 847 824 096 7 | 0.826 416 254 043 579 1 | $B_{29}$ |
| 特大事故 | ≥30 | $X_{1n}$ | $X_{2n}$ | $B_n$ |

对 4 类安全事故, 聚类簇数 $K$=4, 算法开始均值向量取值如下:

$$\begin{cases}\mu_1=\left\{\frac{1}{3}\sum_{i=0}^{2}X_{1i},\frac{1}{3}\sum_{i=0}^{2}X_{2i},\frac{1}{3}\sum_{i=0}^{2}B_i\right\}\\ \mu_2=\left\{\frac{1}{7}\sum_{i=3}^{9}X_{1i},\frac{1}{7}\sum_{i=3}^{9}X_{2i},\frac{1}{7}\sum_{i=3}^{9}B_i\right\}\\ \mu_3=\left\{\frac{1}{30}\sum_{i=10}^{29}X_{1i},\frac{1}{30}\sum_{i=10}^{29}X_{2i},\frac{1}{30}\sum_{i=10}^{29}B_i\right\}\\ \mu_4=\{X_{1n},X_{2n},B_n\}\end{cases}$$

将 $\mu_1$、$\mu_2$、$\mu_3$、$\mu_4$ 作为初始化聚类中心, 然后按照图 4 中算法计算, 得到最终分类.

### 3.3 K-means 算法实验结果

取 1000 个安全生产事故为样本, 把样本的词向量作为聚类的输入, 按照上述实验, 图示化结果如图 5. 图中, 绿色为特大事故, 蓝色为重大事故, 黄色为较大事故, 红色为一般事故。通过得到的词向量和上述聚类算法, 较好的实现安全生产事故分类. 在样本数万时, 分类正确率达 93% 以上. 同时该模型对安全生产事故开展多个维度数据分析也有很好的借鉴意义.

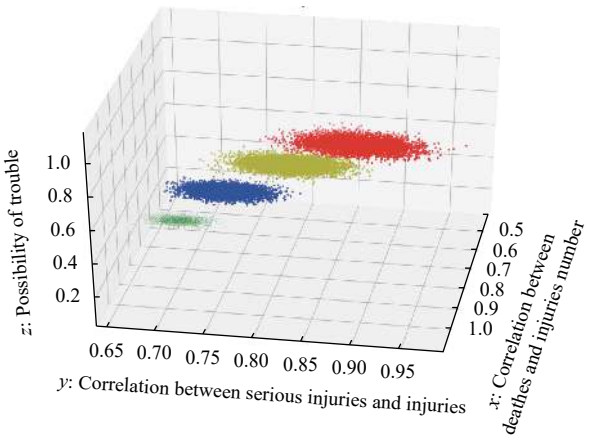


图 5 安全生产事故分类图

## 参考文献


1 易高翔, 魏利军, 吴宗之, 等. 全国安全生产调查信息系统设计与实现. 中国安全生产科学技术, 2009, 5(4): 60–63.

2 Rong X. Word2Vec parameter learning explained. arXiv preprint arXiv: 1411.2738, 2014.

3 Zhang W, Qu CF, Ma L, *et al*. Learning structure of stereoscopic image for no- reference quality assessment with convolutional neural network. Pattern Recognition, 2016, 59: 176–187.

4 Mansoor HH, Shaker SH. Using classiffcation technique to SMS spam filter. International Journal of Innovative Technology and Exploring Engineering, 2019, 10(8): 56–62.

5 李金洪. 深度学习之 TensorFlow 入门、原理与进阶实战. 北京: 机械工业出版社, 2019. 279–296.

6 李孟全. TensorFlow 与自然语言处理应用. 北京: 清华大学出版社, 2019. 77–120.

7 杨楠, 李亚平. 基于 Word2Vec 模型特征扩展的 Web 搜索结果聚类性能的改进. 计算机应用, 2019, 39(6): 1701–1706. [doi: 10.11772/j.issn.1001-9081.2018102106]

8 蒋振超, 李丽双, 黄德根. 基于词语关系的词向量模型. 中文信息学报, 2017, 31(3): 25–31.

9 孙佳伟, 李正华, 陈文亮, 等. 基于词模式嵌入的词语上下位关系分类. 北京大学学报 (自然科学版), 2019, 55(1): 1–7.

10 Rubin TN, Chambers A, Smyth P, *et al*. Statistical topic models for multi-label document classification. Machine Learning, 2012, 88(1–2): 157–208. [doi: 10.1007/s10994-011-5272-5]

11 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 3111–3119.

12 Zheng XQ, Chen HY, Xu TY. Deep learning for Chinese word segmentation and POS tagging. Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, WA, USA. 2013. 647–657.

13 张克君, 史泰猛, 李伟男, 等. 基于统计语言模型改进的 Word2Vec 优化策略研究. 中文信息学报, 2019, 33(7): 11–19. [doi: 10.3969/j.issn.1003-0077.2019.07.002]

14 王千, 王成, 冯振元, 等. K-means 聚类算法研究综述. 电子设计工程, 2012, 20(7): 21–24. [doi: 10.3969/j.issn.1674-6236.2012.07.008]

15 周志华. 机器学习. 北京: 清华大学出版社, 2016. 197–224.

16 周爱武, 于亚飞. K-Means 聚类算法的研究. 计算机技术与发展, 2011, 21(2): 62–65. [doi: 10.3969/j.issn.1673-629X.2011.02.016]