计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700029

基于改进 TF-IIGM 算法的畜禽疫病诊断模型研究

郭晓利[1,2,3]　李奇峰[1,3]　刘　羽[1,3]　张　俊[1,3]　赵红涛[2]　杨　淦[1,3]　蒋瑞祥[1,3]　余礼根[1,3]
1 北京市农林科学院信息技术研究中心　北京 100097
2 华北电力大学数理学院　北京 102206
3 国家数字畜牧业创新中心　北京 100097
(guoxiaoli06@163.com)

摘　要　针对畜禽疫病文本中特征项权重分配不准导致诊断准确率较低的问题,利用提出的 TF-IIGM-NW(Term Frequency-Improved Inverse Gravity Moment With Normalization and Weighting)改进算法结合 Word2vec 词向量进行文本向量化表示。该方法在 TF-IIGM(Term Frequency-Improved Inverse Gravity Moment)算法的基础之上,对其进行归一化处理并结合基于关键词抽取算法设定的规则,进一步提升文本内核心关键词权重,然后将其与结合 Word2vec 词向量获取的文本向量化表示结果输入支持向量机(Support Vector Machine,SVM)进行畜禽疫病诊断。为了验证算法的有效性,基于自建的羊疫病文本数据集,将改进算法与现有词向量常见处理方式进行对比分析。结果表明,基于 TF-IIGM-NW 算法的 macro-F1 值与 micro-F1 值分别达到 96.73%,96.76%;与传统经典算法 TF-IDF(Term Frequency-Inverse Document Frequency)相比,分别提升 2.25%,2.26%;与 TF-IIGM 算法相比,分别提高 0.90%,0.97%。改进算法能够有效提升疫病诊断性能。通过 SVM 在每类疫病上的实验结果分析表明,羊口疮疫病类别最易被错判。

关键词:TF-IIGM;权重;向量化表示;疫病诊断;SVM
中图分类号　TP391

Study on Diagnosis Model of Livestock and Poultry Disease Based on Improved TF-IIGM Algorithm

GUO Xiaoli[1,2,3],LI Qifeng[1,3],LIU Yu[1,3],ZHANG Jun[1,3],ZHAO Hongtao[2],YANG Gan[1,3],JIANG Ruixiang[1,3] and YU Ligen[1,3]
1 Research Center of Information Technology,Beijing Academy of Agriculture and Forestry Sciences,Beijing 100097,China
2 School of Mathematics and Physics,North China Electric Power University,Beijing 102206,China
3 Innovation Center of National Digital Livestock,Beijing 100097,China

Abstract　In order to deal with the problem of low diagnostic accuracy caused by inaccurate weight allocation of feature items in livestock and poultry diseases texts,the improved TF-IIGM-GW algorithm combined with Word2vec word vector is used to realize the text vectorization. On the basis of the TF-IIGM weighting method,the method is normalized and combined with the rule based on the keyword extraction algorithm to further improve the weight of core keywords in the texts. Finally,the text vectorization results obtained by combining the weight with Word2vec word vector are inputted into the support vector machine(SVM) for diagnosis of livestock and poultry diseases. In order to verify the effectiveness of the improved algorithm,based on the self-built text datasets of livestock and poultry diseases,the improved algorithm is compared with the commonly used methods of word vector. Results show that the macro-F1 value and micro-F1 value based on the TF-IIGM-GW algorithm are 96.73% and 96.76%,respectively,which are 2.25% and 2.26% higher than those of the commonly used algorithm TF-IDF,and 0.90% and 0.97% higher than those of TF-IIGM weighting method. The improved algorithm could effectively improve the performance of disease diagnosis. The analysis of the experimental results of SVM on each type of diseases shows that sheep oral aphthae is most easily misjudged.

Keywords　TF-IIGM,Weighting,Vectorization,Disease diagnosis,SVM

基金项目:国家重点研发计划(2023YFD1300805);云南省重大科技专项计划(202102AE090039);北京市农林科学院创新能力建设专项(KJCX20230204);内蒙古现代畜牧业发展战略研究(2023NM2N-01)
This work was supported by the National Key R&D Program of China(2023YFD1300805),Yunnan Province Major Science and Technology Special Program(202102AE090039),Beijing Academy of Agriculture and Forestry Sciences Innovation Capacity Building Project(KJCX20230204) and Research on the Development Strategy of Modern Animal Husbandry in Inner Mongolia(2023NM2N-01).
通信作者:余礼根(yulg@nercita.org.cn)

1 引言

目前，我国畜牧业正逐渐朝着集约化、规模化与智能化方向发展，畜牧养殖密度的持续增加使得畜禽疫病诊断难度显著加大[1]。随着新一代信息技术的发展，研究者们[2-4]大多将机器视觉、音频分析、红外感知、智能诊断系统等自动化技术与传统畜禽养殖业相结合，较大地提高了疫病诊断水平。疫病诊断多基于病例文本数据[5]。本文以提升文本编码效率为目标，构建更优化的诊断模型，以有效提高疫病诊断效率，降低畜禽死亡率。

研究表明，特征项权重分配作为文本向量化表示过程中的一个重要步骤，对疫病诊断准确率有较大影响，特征项权重越大，说明其对疫病文本诊断结果的影响程度越高[6-7]。基于畜禽疫病领域内的专业词汇相较于通用词汇对疫病诊断结果的影响更大，本文将通过提高文本内核心关键词权值，进而有效改善特征词权重分配不准的问题。

受 Ao 和 Lan 等[8-9]将 TF-IDF 与 TextRank 关键词抽取算法相结合进一步突显文本内关键信息思想的启发，本文提出 TF-IIGM-NW 改进算法，在考虑了类间分布信息的 TF-IIGM 算法基础之上，与基于 Salience Rank 关键词抽取算法设定的规则算法相结合，对不同级别的关键词赋予不同的权值，利用改进算法提升文本内核心关键词权重，以有效提高疫病诊断性能。

2 相关工作

文本作为一种非结构化数据，无法直接被计算机识别。在疫病诊断的过程中，通常需要将文本形式转换为可被计算机成功识别的向量化形式，这一过程被称为文本表示[10-11]。

目前，向量空间模型(Vector Space Models，VSM)是一种应用较为广泛的通过将文本的特征项加权进行向量化表示的方法[12]。而在提出的诸多加权法中，TF-IDF 最为常用，但其对特征项加权时并没有利用训练文本的已知类信息，计算出的权重不能充分反映该特征项在文本分类中的重要性[13]。基于这一问题，Debole 和 Sebastiani[14]提出了监督项加权法，通过考虑与文本类别相关的已知信息对特征项加权，利用特征选择方法，如卡方统计量(Chi-Square Statistic，CHI2)、信息增益(Information Gain，IG)和增益比(Gain Ratio，GR)替换 IDF 因子，相应地提出了 TF-CHI，TF-IG 和 TF-GR 算法。后续学者也提出了类似的监督项加权法，即基于 TF-IDF 算法，利用其他考虑了类信息的权重因子替换 IDF 全局因子，如基于互信息(Mutual Information，MI)的 TF-MI 算法、基于概率(Probability，PB)的 TF-PB 算法和基于逆类频率(Inverse Category Frequency，ICF)的 TF-ICF 算法等。

此外，Chen 等[15]基于对特征项的类间分布集中度和不均匀性的考虑，提出了 TF-IGM(Term Frequency-Inverse Gravity Moment)算法。实验结果表明，与传统加权法 TF-IDF，TF-CHI 以及 TF-PB 等相比，基于 TF-IGM 的类间分布信息表示效果更好。但在某些特殊情况下，TF-IGM 对不同特征项分配的权重相同。基于此，Dogan 等[16]提出 IIGM 算法，在 IGM 的基础上添加了一个比率值，有效提高了标准 IGM 在特殊场景下的加权性能。也有研究者在 TF-IDF 传统算法的基础上直接加入其他权重因子以优化算法。Xu 等[17]针对新闻文本数据集，通过引入去中心化词频因子和特征词位置因子加强了新闻文本中的各特征词权重的准确性。Xu[18]为从大量英语语料库中获取所需信息，在 TF-IDF 算法的基础之上，考虑了类间、类内特征词的分布密度情况，显著提升了模型效率。Jing 等[19]利用特征项在类内、类间的分布关系和位置信息来改进 TF-IDF 算法，进一步突出了特征项的重要性，同时结合了 Word2vec 词向量对文本进行向量化表示。结果表明，结合改进 TF-IDF 的分类模型在 THUCNews 和 online_shopping_10_cats 数据集上的准确率分别达到 97.38%和 91.33%。Tang 等[20]基于 TF-IDF 算法引入累计残差熵与比例失真函数，与现有的其他 9 种加权法相比，分类效果更优，在 Reuters-21578 数据集上，micro-F1 值达 97.20%。

上述研究分别在 TF-IDF 传统算法的基础之上，通过替换 IDF 全局因子或引入新的权重因子来增强算法性能。考虑到疫病文本内症状相关词汇对模型诊断结果的影响更大，本文提出 TF-IIGM-NW 算法，基于考虑了特征项类间分布信息的 TF-IIGM 算法，引入基于关键词抽取算法设定的权重因子，进一步提升文本内关键词权重，获取更优的文本向量化表示结果，进而提升基于文本数据的畜禽疫病诊断模型性能。

3 基于 TF-IIGM-NW 的 SVM 诊断模型

3.1 TF-IIGM

特征项在文本中的权重大小由其在文本中的重要性及其对文本分类的贡献度来确定，分别对应于特征项的局部和全局加权因子。本研究基于考虑特征项类间分布信息的 TF-IIGM 算法，与 TF-IDF 相同，TF-IIGM 算法的局部加权因子为 TF，即特征项在文本中的频率。特征项对文本分类的贡献度取决于类别区分能力，区分度越高，赋予特征项的权重越大，对应 TF-IIGM 中的 IIGM 因子。TF-IIGM 的计算式为：

$$TF\text{-}IIGM = tf(w_i, d_k) \times \left(1+\lambda \times \frac{f_{i1}}{\sum_{j=1}^{m} f_{ij} \times j + \log_{10}(D_{\text{total}}(w_{i\_\max})/f_{i1})}\right) \tag{1}$$

其中，λ 为可调系数，用来保持特征项权重中局部与全局因子之间的相对平衡，通常将其范围设定为[5,9]；m 为总类别数；f_{ij} 为每类中包含特征项 w_i 的文本总数按降序排列后，第 j 个类中包含该特征项的文本总数，j 为秩，$j=1,2,\cdots,m$；$D_{\text{total}}(w_{i\_\max})$ 表示特征项 w_i 出现次数最多的类中的总文本数。

3.2 权重因子

3.2.1 Salience Rank 算法

Salience Rank[21]算法由 Teneva 等提出，是对 Topical PageRank[22]算法的优化和改进。Topical PageRank 算法首先基于隐含狄利克雷分布(Latent Dirichlet Allocation，LDA)主题模型，计算出每个单词在各个主题中出现的概率，随后在每个主题下运行 PageRank 算法，按得分值由高到低排序获得每个主题下的关键词。然而，Topical PageRank 算法的计算复杂度较高，因此，Sterckx 等[23]进一步提出了 Single Topical PageRank 算法，通过仅考虑单词在所有主题下的概率分布，保证了模型性能，同时简化了算法的计算过程。Salience Rank 采用了与 Single Topical PageRank 类似的方法来改进 Topical PageRank，同时通过计算单词在不同主题下的特异

性，有效避免了与主题相关性较弱的特征项干扰，从而提高了关键词提取的准确性。

为了同时提升算法的性能与灵活性，Salience Rank 算法将基于 LDA 得到的 K 个潜在主题组合成一个定义为 Word Salience 的单词度量值，将其作为每个特征项 $w_i \in W'$ 的偏好值（$W'=\{w_1,w_2,\cdots,w_n\}$ 表示所有文本中出现的特征项集合）。因此，Salience Rank 仅需执行一次 PageRank 就可以获取每条文本中特征项按重要程度由高到低排序的结果。

定义 Word Salience 前，需要对特征项的主题特异性和语料库特异性进行计算。

特征项 w_i 的主题特异性用来衡量其在多个主题间的共享程度，共享程度越低，主题特异性值越高，计算式为：

$$TS(w_i)=\sum_{t\in T}p(t/w_i)\log\frac{p(t/w_i)}{p(t)} \tag{2}$$

其中，条件概率 $p(t/w_i)$ 表示特征项 w_i 由主题 t 产生的可能性，$p(t)$ 表示任意随机选取的特征项由主题 t 产生的可能性。由于 $TS(w_i)$ 是非负且无界的，因此将其 Min-Max 归一化为[0,1]的值。

$$TS(w_i)^*=\frac{TS(w_i)-\min_u TS(u)}{\max_u TS(u)-\min_u TS(u)} \tag{3}$$

通过计算语料库中的特征项频率来评估特征项的语料库特异性，计算式为：

$$CS(w_i)=p(w_i/corpus) \tag{4}$$

特征项 w_i 的 Salience 值 $S(w_i)$ 定义为其主题特异性与语料库特异性的线性组合。计算式为：

$$S(w_i)=(1-\alpha)CS(w_i)+\alpha TS(w_i)^* \tag{5}$$

其中，α 为权衡主题特异性与语料库特异性的参数，取值范围为[0,1]。

Salience Rank 将每个特征项 $w_i \in W'$ 按其得分 $R(w_i)$ 的值由高到低进行排列，计算式为：

$$R(w_i)=\lambda\sum_{j;w_j\to w_i}\frac{e(w_j,w_i)}{Out(w_j)}R(w_j)+(1-\lambda)S(w_i) \tag{6}$$

$$Out(w_j)=\sum_{j;w_j\to w_i}e(w_j,w_i) \tag{7}$$

其中，λ 取值范围为[0,1]，$e(w_j,w_i)$ 表示文本内两个特征词 w_j 与 w_i 之间的相关性，$W'=\{w_1,w_2,\cdots,w_n\}$ 表示语料库中所有特征词的集合。实验过程中，默认 α 值为 0.3，λ 值为 0.85。

3.2.2 规则算法

Salience Rank 算法基于 LDA 主题模型。本文研究对象为羊的 3 种典型疫病：羊口疮、羊疥螨及羊传染性胸膜肺炎。研究类别为 3 类，故直接将 LDA 的主题数 topic_num 设置为 3；另外将参数 top_k 设置为 15，表示基于 Salience Rank 算法计算每条文本内特征词的得分值后，抽取前 15 个得分值最高的特征词；每个主题下按得分值由高到低排序后，选取前 100 个关键词，设定规则如下。

基于每个主题下提取的前 100 个关键词，构建 3 个主题中共有的特征词列表 list1、两个主题中共有的特征词列表 list2（去除重复项及在 list1 中包含的特征词），以及只有某一主题中包含的特征词列表 list3。统计 list1，list2 及 list3 中包含的特征词数，分别表示为 n_1，n_2，n_3。按重要程度排序，从高到低为：list3 中特征词、list2 中特征词、list1 中特征词。根据 n_1，n_2，n_3 的取值容易发现，list3 中特征词数＞list2 中特征词数＞list1 中特征词数。3 个列表中的特征词重要程度与特征词数量整体分布趋势相同。

引入对数函数 $\log_2(2+X)$，其中，X 表示列表中特征词所占的比例，取值范围为(0,1)。对数函数中设定了常数值 2 使得权重因子值大于 1。为防止引入的权重因子过小或过大，本文采用对数函数，在确保权重因子 W 的权值大于 1 的同时又小于类间分布因子 $1+7*IIGM$ 的最小值（$1+7*IIGM$ 最小值为 1.8309。依据平衡系数评估结果，在 $\lambda=7$ 时，TF-IIGM 各评价指标值最高）。对权重因子 W 设定规则为：

$$W=\begin{cases}\log_2\left(2+\dfrac{n_2}{n}\right), & w_i\in \text{list2}\\ \log_2\left(2+\dfrac{n_3}{n}\right), & w_i\in \text{list3}\\ \log_2\left(2+\dfrac{n_1}{n}\right), & \text{其它}\end{cases} \tag{8}$$

其中，n_1 表示 list1 中特征词数，即 3 个主题中共有的特征词数；n_2 表示 list2 中特征词数，即两个主题中共有的特征词数（去除重复项和在 list1 中包含的特征词）；n_3 表示 list3 中特征词数，即仅在某一主题中包含的特征词数；n_1，n_2，n_3 分别取值为 18，47，152。$n=n_1+n_2+n_3=217$。

3.3 L2 范数归一化

L2 范数归一化指将向量中的每一个元素除以向量的 L2 范数（又称欧几里得范数）。具体来说，对于一个 n 维向量 $x=(x_1,x_2,\cdots,x_n)$，其 L2 范数 $norm(x)$ 定义为：

$$norm(x)=\sqrt{x_1^2+x_2^2+\cdots+x_n^2} \tag{9}$$

对 x_i 进行 L2 范数归一化，定义为：

$$x_i'=\frac{x_i}{norm(x)} \tag{10}$$

L2 范数归一化常用于对特征向量进行预处理，可以防止某些特征对模型训练结果产生太大影响，同时可以加快收敛速度，提高模型的稳定性及泛化能力。本文对算法进行归一化处理，避免了由于权值差别过大导致结果偏向长文本的情况。

3.4 文本表示

在进行疫病诊断任务的过程中，需要将文本转换为向量形式。本文首先基于先前构建的 BERT-BiLSTM-CRF 畜禽疫病文本分词模型[24]对文本数据进行分词并进行去停用词及正则化处理。其次，利用 Word2vec 编码工具将预处理后的文本转化为词向量，词向量维度可以根据实际需求人为设置，常见设定范围为 100～300 维。一般情况下，向量维度越高，所包含的文本信息就越丰富，本文设定词向量维度为 300 维。同时，选取 Word2vec 的 Skip-Gram 模型获取词向量，这是因为，与 CBOW 相比，Skip-Gram 虽然训练时间长但学习的词向量更细致，精度更高，且当语料库中含有生僻词时，Skip-Gram 更适用[25]。获取文本内各特征项的词向量后，需要通过一定的处理方式将整条文本进行向量化表示。常见处理方法有累加法、平均法和加权法，TF-IDF 是应用最为广泛的特征项加权法。累加法虽然简单直观易于实现，却极易受到文本长度的影响。平均法将词向量的叠加和除以特征项总数，消除了长度差异的影响，但其同等看待每个特征项对文本的贡献程度。事实上，特别是对于特定领域文本来说，领域内专业词汇对文本分类结果的影响要大于通用词汇。TF-IDF 算法考虑了不同特征项间的贡献度不同，在计算特征项的权重因子时，综合了特征项的词频及逆文档频率因子，但并没有考虑到特征项在每个类别中的分布情况。TF-IIGM 在 TF-

IDF 的基础上进行了改进。

本文基于考虑了类间分布信息的 TF-IIGM 算法引入权重因子 W，提出 TF-IIGM-NW 算法计算每个特征项的权重，并与 Word2vec 词向量相结合，获取文本向量化表示结果。假设预处理后的每条文本表示为 $D=\{d_1,d_2,\cdots,d_n\}$，n 表示文本 D 内包含的总特征词数，文本表示结果如式(11)所示：

$$V_{\text{words}}=\sum_{i=1}^{n}D_i * TF\text{-}IIGM\text{-}NW(d_i) \tag{11}$$

其中，$TF\text{-}IIGM\text{-}NW(d_i)$表示计算得到的特征项 d_i 的权值；D_i 表示特征项 d_i 基于 Word2vec 编码获取的词向量。疫病文本进行向量化表示的整体流程如图 1 所示。

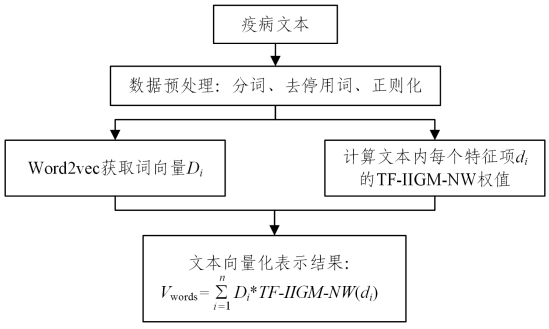

图 1 文本向量化表示流程图

Fig. 1 Flow chart of text vectorized representation

3.5 SVM 算法

支持向量机[26]的核心思想是寻找一个超平面，使两类间的间隔最大化。这个超平面被称为最大间隔超平面，能够实现更好的分类效果。

如图 2 所示，平面内具有两类样本点的线性可分数据集，分别用黑点、白点表示。存在多个分离超平面可将两类样本点正确切分，而支持向量机寻求的最优分离超平面 $w^{\mathrm{T}}x+b=0$，不仅能够将样本数据集正确分类，并且能够确保几何间隔最大。图 2 中，$w^{\mathrm{T}}x+b=1$ 与 $w^{\mathrm{T}}x+b=-1$ 表示离超平面最近的点线，两条线相互平行；$\frac{b}{\|w\|}$表示超平面到原点的距离；$\frac{2}{\|w\|}$表示两条线的间隔距离。

支持向量机最开始提出是为了解决二分类问题，但在现实生活中常见的是多分类问题，类似新闻分类、自动邮件归类等。本文研究的也是多分类任务，此时可以将多分类问题拆分，通过转化为多个二分类的组合以达到多分类的目的。

支持向量机模型参数大多依据经验进行人为选取，没有具体的理论基础，而参数的选择在很大程度上影响支持向量机分类性能的好坏。本文通过网格搜索算法在人为设定的取值范围内寻求核函数 g 与惩罚系数 c 的最优解。

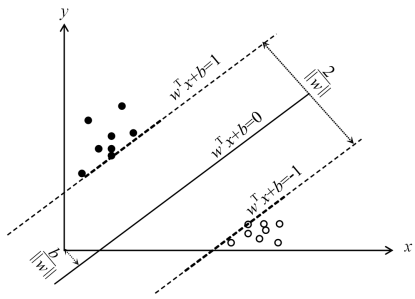

图 2 支持向量机

Fig. 2 Support vector machine

4 实验与分析

4.1 实验设置与数据集

4.1.1 实验环境

本文实验环境采用 Jupter Notebook 6.4.8，Python 版本为 3.8，使用 Scikit-learn、Pandas、Numpy 等工具包，处理器为 Intel(R) Core(TM) i5-7400 CPU @ 3.00 GHz，内存为 12 GB。

4.1.2 数据集

为解决畜禽疫病诊断难题，本文依据兽医专家经验知识选取羊的 3 种典型疫病：羊口疮、羊疥螨及羊传染性胸膜肺炎为研究对象。考虑到尚没有公开的羊疫病诊断病例报告，本文采用自建数据集进行实验分析。数据集由书本数据与网络数据组成。其中，书本数据来源于 13 本羊疫病相关书籍，如《羊病诊治原色图谱》《羊病诊疗原色图谱(第 2 版)》及《羊病临床诊治彩色图谱》等，网络数据来源于知网、维普、读秀等网站。提取内容主要包括羊疫病的症状、病理变化、诊断等相关描述文本，共收集数据集 1 030 条，各类别按 7∶3 划分训练集与测试集，数据分布情况如表 1 所列，数据集样例如表 2 所列。

表 1 数据集

Table 1 Datasets

类别	训练集	测试集	总计
羊口疮	266	114	380
羊疥螨	189	81	270
羊传染性胸膜肺炎	266	114	380
总计	721	309	1 030

收集的 1 030 条数据集中羊口疮、羊疥螨与羊传染性胸膜肺炎分别为 380 条、270 条、380 条。共有训练集 721 条，测试集 309 条。

表 2 数据集样例

Table 2 Samples of datasets

疫病文本描述(数据处理前)	疫病文本描述(数据处理后)	疫病类别
多见于一肢或四肢蹄部感染，通常于蹄叉、蹄冠或系部皮肤形成水疱、脓肿，破裂后形成溃疡，继发感染时形成坏死和化脓，病羊跛行，喜卧而不能站立。	多见 一肢 四肢 蹄部 感染 蹄叉 蹄冠 系部 皮肤 水疱 脓肿 破裂 溃疡 继发 感染 坏死 化脓 病羊 跛行 喜卧 不能 站立	羊口疮
病变扩展到口黏膜，如唇内、齿龈、颊黏膜、舌侧缘和软腭上，可形成红晕包围的微白色水疱，水疱继而变成脓疱，最终破裂形成烂斑。	病变 扩展 口黏膜 唇内 齿龈 颊黏膜 舌侧缘 软腭 红晕 包围 微白色 水疱 水疱 脓疱 最终 破裂 烂斑	羊疥螨
发病初期患病羊先是表现为精神状态逐渐变差，体温升高到 40 ℃以上，有时能够达到 42℃，不能正常采食，随即会出现频繁的咳嗽现象，先是出现干咳，随后变成湿咳，从鼻腔当中会流出清澈的鼻液。	发病 初期 患病羊 先是 表现 精神 状态 变差 体温 升高 不能 采食 随即 频繁 咳嗽 现象 先是 干咳 湿咳 鼻腔 流出 清澈 鼻液	羊传染性胸膜肺炎

本文基于先前研究得到的 BERT-BiLSTM-CRF 分词模型对文本数据分词，并进行去停用词与正则化操作，数据处理前后的文本形式如表 2 所列。其中，进行停用词处理时，选取比较常用的 4 个停用词表进行合并去重，并依据畜禽疫病文本数据特点，在合并后的停用词表中去除数字、否定词等。4 个停用词表分别为：中文停用词表、哈工大停用词表、百度停用词表及四川大学机器智能实验室停用词表。

4.2 评价指标

对于二分类问题，常将精确率(Precision，P)、召回率(Recall，R)和 F1 值作为评价指标。计算式为：

$$P=\frac{TP}{TP+FP} \tag{12}$$

$$R=\frac{TP}{TP+FN} \tag{13}$$

$$F1=\frac{2PR}{P+R} \tag{14}$$

除此之外，在多分类问题上有两种常用的 F1 值计算方法，分别是宏观 F1(macro-F1)和微观 F1(micro-F1)。计算式如下：

$$Macro\text{-}F1=\frac{1}{m}\sum_{i=1}^{m}F_1(c_i) \tag{15}$$

$$F_1(c_i)=\frac{2\cdot TP(c_i)}{2\cdot TP(c_i)+FP(c_i)+FN(c_i)} \tag{16}$$

$$Micro\text{-}F1=\frac{2\cdot\sum_{i=1}^{m}TP(c_i)}{2\cdot\sum_{i=1}^{m}TP(c_i)+\sum_{i=1}^{m}FP(c_i)+\sum_{i=1}^{m}FN(c_i)} \tag{17}$$

其中，$TP(c_i)$表示实际为 c_i 类，预测也为 c_i 类的样本数；$FP(c_i)$表示实际不属于 c_i 类，预测为 c_i 类的样本数；$FN(c_i)$表示实际为 c_i 类，预测为非 c_i 类的样本数。

本文在疫病诊断的多分类任务上以 macro-F1 与 micro-F1 为诊断性能的衡量指标。

4.3 IIGM 平衡系数的评估

IIGM 的平衡系数取值范围为 5.0～9.0。本文以 1 为间隔值，分别取参数值为 5，6，7，8，9 进行实验验证，寻找适合本文数据集的 IIGM 平衡系数最优值。

由表 3 和图 3 可知，macro-F1 与 micro-F1 整体升降趋势相似，在 λ 取 5，6，7 时各评价指标值相差不大，特别是 micro-F1 值保持不变，macro-F1 差值也仅在 0.1% 以内。在 λ=7 时各评价指标值最高，macro-F1 与 micro-F1 值分别为 95.83% 和 95.79%。λ=8 时，macro-F1 与 micro-F1 值均明显下降，而在 λ=9 时，评价指标值与 λ=8 时相同。实验结果表明，选取不同的平衡系数值，在某些情况下能够保持分类性能的稳定性，但个别情况下，会明显降低分类性能。因此，对 IIGM 算法平衡系数的评估具有一定的研究意义。本次实验选取 λ 值为 7。

表 3 平衡系数评估

Table 3 Balance coefficient evaluation

(%)

参数 λ	Macro-F1	Micro-F1
λ=5	95.77	95.79
λ=6	95.77	95.79
λ=7	95.83	95.79
λ=8	95.43	95.47
λ=9	95.43	95.47

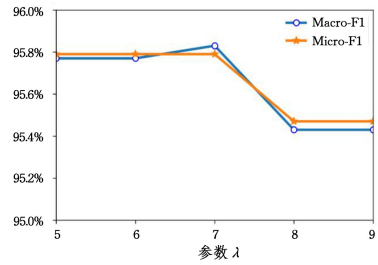

图 3 平衡系数评价指标值

Fig. 3 Value of the evaluation index of balance coefficient

4.4 不同归一化处理方式的对比分析

本文的改进算法同时引入了归一化因子与基于规则算法设定的权重因子，基于归一化不同的先后处理方式以及单一的处理方法，利用网格搜索算法优化的 SVM 模型进行疫病诊断，对比分析诊断结果。实验设计如下：

1)TF-IIGM：用反重力矩阵 IIGM 因子替换 TF-IDF 中全局因子 IDF 的算法。

2)TF-IIGM-N：在 TF-IIGM 算法基础上引入归一化因子，即通过 L2 范数对 TF-IIGM 进行归一化处理。

$$TF\text{-}IIGM\text{-}N=\frac{TF\text{-}IIGM}{\sqrt{\sum(TF\text{-}IIGM)^2}} \tag{18}$$

3)TF-IIGM-W：基于 TF-IIGM 获取的特征词权重，与基于规则算法获取的权重因子 W 相乘，乘积为特征词的最终权重。

4)TF-IIGM-WN：基于 TF-IIGM 算法的特征词权重与基于规则算法的权重因子 W 相乘以获取特征词总体权重 TF-IIGM-W，同时对 TF-IIGM-W 进行 L2 范数归一化处理。

$$TF\text{-}IIGM\text{-}WN=\frac{TF\text{-}IIGM*W}{\sqrt{\sum(TF\text{-}IIGM*W)^2}} \tag{19}$$

5)TF-IIGM-NW：对 TF-IIGM 算法进行 L2 范数归一化处理，随后，与基于规则算法获取的权重因子 W 相乘获取最终权值。

$$TF\text{-}IIGM\text{-}NW=\frac{TF\text{-}IIGM}{\sqrt{\sum(TF\text{-}IIGM)^2}}*W \tag{20}$$

由表 4 可以看出，TF-IIGM-N 算法直接对 TF-IIGM 进行 L2 范数归一化处理，macro-F1 值不升反降，这表明归一化处理并不总是能够提升任务性能也并非适用于任何场景，在个别情况下，不进行归一化处理效果更好。TF-IIGM-W 算法直接在 TF-IIGM 的基础之上引入基于规则算法设定的权重因子 W，与 TF-IIGM 的各评价指标值相比，macro-F1 有极小的提升，micro-F1 值保持不变，这说明仅仅直接引入权重因子 W，并不能有效提高诊断性能。

表 4 不同归一化处理方式的对比分析

Table 4 Comparative analysis of different normalization treatments

(%)

算法	Macro-F1	Micro-F1
TF-IIGM	95.83	95.79
TF-IIGM-N	95.80	95.79
TF-IIGM-W	95.89	95.79
TF-IIGM-WN	96.66	96.76
TF-IIGM-NW	96.73	96.76

总的来说，上述结果表明，仅引入权重因子 W 或只对 TF-IIGM 进行归一化处理，提升效果不佳。基于此，考虑在 TF-IIGM 算法的基础之上同时进行 L2 范数归一化并引入权

重因子 W，共有两种处理方式，即 TF-IIGM-WN 与 TF-IIGM-NW。由表 4 可知，TF-IIGM-WN 与 TF-IIGM-NW 算法都能够有效提升诊断结果，且 TF-IIGM-NW 优于 TF-IIGM-WN。这是因为，TF-IIGM-NW 算法通过先对 TF-IIGM 进行归一化处理，使得该算法中的权重因子 W 相较于在 TF-IIGM-WN 算法中占比更大，更关注文本内核心关键词对向量化表示结果的影响。TF-IIGM-NW 的 macro-F1 值与 micro-F1 值分别为 96.73%，96.76%，与未引入权重因子 W 且未进行归一化处理的 TF-IIGM 相比，分别提升了 0.90%，0.97%。TF-IIGM-WN 的 macro-F1 值与 micro-F1 值分别为 96.66%，96.76%，与未引入权重因子且未进行归一化处理的 TF-IIGM 算法相比，分别提升了 0.83%，0.97%，验证了对 TF-IIGM 算法同时进行归一化处理并引入权重因子 W 的有效性。

4.5 不同疫病类别诊断结果分析

为进一步探究诊断模型在各类别上的诊断结果，基于改进的 TF-IIGM-NW 算法，利用精确率、召回率与 F1 值评价指标进行分析。

图 4 中标签 0，1，2 分别表示羊的 3 类典型疫病：羊口疮、羊疥螨、羊传染性胸膜肺炎。

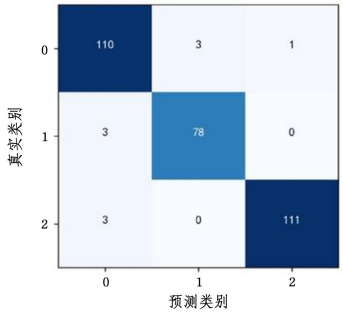

图 4 各类别诊断结果的混淆矩阵

Fig. 4 Confusion matrix of diagnostic results for various categories

由表 5 可知，SVM 诊断模型的精确率平均值、召回率平均值及 F1 值平均值分别达到 96.74%，96.72%，96.73%，这表明本文诊断模型有较好的诊断效果。其中，羊传染性胸膜肺炎类别的 F1 值最高，羊疥螨次之，羊口疮的 F1 值最低，表明在 3 类疫病文本数据中，羊口疮类别被错判的可能性最高。同时，根据图 4 的混淆矩阵可知，所有测试集样本中共有 10 个误判值，其中其他类别文本被错判为羊口疮的有 6 个，羊口疮类别文本被错判为其他类别的有 4 个，羊口疮被错判的数量占比最多即最易被错判。

表 5 SVM 模型在各类别的诊断结果分析

Table 5 Analysis of diagnostic results of SVM model in various categories

(%)

类别	精确率	召回率	F1
羊口疮	94.83	96.49	95.65
羊疥螨	96.30	96.30	96.30
羊传染性胸膜肺炎	99.11	97.37	98.23
平均	96.74	96.72	96.73

4.6 词向量不同处理方式的对比分析

为防止过拟合问题，本文采取 10 折交叉验证法进行实验。基于 Word2vec 获取词向量后，分别结合累加法、平均法、传统的 TF-IDF 算法、TF-IIGM 算法及本文提出的 TF-IIGM-NW 算法，将词向量转化为文本向量，再通过基于网格搜索算法优化的 SVM 模型进行诊断，实验结果如表 6 所列。

表 6 词向量不同处理方式的对比分析

Table 6 Comparative analysis of different processing methods of word vector

(%)

算法	Macro-F1	Micro-F1
累加法	93.99	93.85
平均法	95.01	95.15
TF-IDF	94.48	94.50
TF-IIGM	95.83	95.79
TF-IIGM-NW	96.73	96.76

通过表 6 可知，基于 TF-IIGM-NW 算法进行文本向量化表示，其诊断效果优于累加法、平均法及 TF-IDF 算法，TF-IIGM-NW 的 macro-F1 和 micro-F1 值分别为 96.73% 和 96.76%，与 TF-IIGM 算法相比分别提升 0.90%，0.97%，与传统算法 TF-IDF 相比分别提升 2.25%，2.26%，与平均法相比分别提升 1.72%，1.61%，与累加法相比分别提升 2.74%，2.91%，验证了本文算法的有效性。对 Word2vec 词向量 5 种不同处理方法的实验结果对比可知，累加法效果最差，这是因为累加法直接通过对特征词的特征向量累加获取文本向量，极易受到文本长度的影响，且没有考虑不同特征词对文本表示结果重要程度的差异。TF-IIGM 的 macro-F1 和 micro-F1 值与 TF-IDF 相比分别提升了 1.35%，1.29%，这是因为 TF-IIGM 在计算权重时考虑了特征词在类间的分布情况，更充分地利用了文本的数据信息。基于 TF-IDF 算法与通过平均化处理后的评价指标 macro-F1 和 micro-F1 值相比分别降低了 0.53%与 0.65%，这表明对词向量加权进行文本向量化表示后，其诊断效果并不总是优于平均化的简单处理结果，这可能与文本内部分通用词汇计算得到较大的权值有关，也表明了本文基于 TF-IDF 算法考虑特征词类间分布的必要性。

结束语 为解决文本内特征项权重分配不准的问题，本文在考虑了类间分布信息的 TF-IIGM 算法的基础之上，进一步结合关键词抽取算法，利用基于关键词抽取算法设定的权重因子，构建了 TF-IIGM-NW 算法，有效提升了文本内核心关键词权重。通过将本文提出的加权法与其他常见的词向量处理方式进行对比分析，结果表明，改进的 TF-IIGM-NW 算法使模型诊断性能达到最优，macro-F1 和 micro-F1 值分别达到 96.73%和 96.76%，优于其他对比方法的诊断结果。另外，SVM 在每类疫病上的精确率、召回率及 F1 值结果表明，羊口疮疫病类别最易被错判。改进的 TF-IIGM-NW 算法虽然对疫病文本的诊断结果有一定的提升，但其未考虑对文中同义词的处理。文本内意思表达一致的特征词，如瘙痒与搔痒等，通过不同的词向量表示，会在一定程度上低估这些关键词的重要程度，未来可考虑结合同义词词库匹配或相似度计算等进行研究。此外，也可通过对参数优化算法的改进进一步提升模型诊断性能。

参 考 文 献

[1] JIANG R X, YU L G, DING L Y, et al. Development Status and Prospect of Intelligent Prevention and Control Technology for Livestock and Poultry Diseases[J]. Chinese Journal of Animal Science, 2020, 56(10): 23-28.

[2] WANG H, SHEN W, ZHANG Y, et al. Diagnosis of dairy cow

diseases by knowledge-driven deep learning based on the text reports of illness state[J]. Computers and Electronics in Agriculture,2023,205:107564.

[3] MUHAMEDIYEVA D T,SAFAROVA L U,TUKHTAMURODOV N. Early diagnostics of animal diseases on the basis of modern information technologies[C]// AIP Conference Proceedings. AIP Publishing,2023.

[4] ZHENG S,ZHOU C,JIANG X,et al. Progress on infrared imaging technology in animal production: a review[J]. Sensors,2022,22(3):705.

[5] TERRADA O,CHERRADI B,RAIHANI A,et al. A novel medical diagnosis support system for predicting patients with atherosclerosis diseases[J]. Informatics in Medicine Unlocked,2020,21:100483.

[6] ALSMADI I,HOON G K. Term weighting scheme for short-text classification: Twitter corpuses[J]. Neural Computing and Applications,2019,31(8):3819-3831.

[7] LI C,LI W,TANG Z,et al. An improved term weighting method based on relevance frequency for text classification[J]. Soft Computing,2023,27(7):3563-3579.

[8] AO X,YU X,LIU D,et al. News keywords extraction algorithm based on TextRank and classified TF-IDF[C]// International Wireless Communications and Mobile Computing (IWCMC 2020). IEEE,2020:1364-1369.

[9] LAN X F,LIU Z,XU Z H,et al. A Chinese Text Keyword Extraction Method Based on the Combination of TF-IDF and TextRank——A Case Study of Sports News[J]. Software Engineering,2023,26(8):6-10.

[10] ZHAO J S,SONG M X,GAO X,et al,Research on Text Representation in Natural Language Processing[J]. Journal of Software,2022,33(1):102-128.

[11] SIEBERS P,JANIESCH C,ZSCHECH P. A survey of text representation methods and their genealogy[J]. IEEE Access,2022,10:96492-96513.

[12] SALTON G,WONG A,YANG C S. A vector space model for automatic indexing [J]. Communications of the ACM,1974,18(11):613-620.

[13] MAHDI A Y,YUHANIZ S S. Automatic Diagnosis of COVID-19 Patients from Unstructured Data Based on a Novel Weighting Scheme[J]. Computers,Materials & Continua,2023,74(1).

[14] DEBOLE F,SEBASTIANI F. Supervised term weighting for automated text categorization[C]// Proceedings of the 2003 ACM Symposium on Applied Computing. 2003:784-788.

[15] CHEN K,ZHANG Z,LONG J,et al. Turning from TF-IDF to TF-IGM for Term Weighting in Text Classification[J]. Expert Systems with Applications,2016,66:245-260.

[16] DOGAN T,UYSAL A K. Improved Inverse Gravity Moment Term Weighting for Text Classification[J]. Expert Systems with Applications,2019,130:45-59.

[17] XU T H,WU M L. An Improved Naive Bayes Algorithm Based on TF-IDF[J]. Computer Technology and Development,2020,30(2):75-79.

[18] XU J. A Natural Language Processing Based Technique for Sentiment Analysis of College English Corpus[J]. PeerJ Computer Science,2023,9:e1235.

[19] JING L,HE T T. Chinese Text Classification Model Based on-Improved TF-IDF and ABLCNN[J]. Computer Science,2021,48(S2):170-175.

[20] TANG Z,LI W,LI Y. An Improved Supervised Term Weighting Scheme for Text Representation and Classification[J]. Expert Systems with Applications,2022,189:115985.

[21] TENEVA N,CHENG W. Salience Rank:Efficient Keyphrase Extraction with Topic Modeling[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:530-535.

[22] LIU Z,HUANG W,ZHENG Y,et al. Automatic Keyphrase Extraction via Topic Decomposition[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010:366-376.

[23] STERCKX L,DEMEESTER T,DELEU J,et al. Topical Word Importance for Fast Keyphrase Extraction[C]// Proceedings of the 24th International Conference on World Wide Web. 2015:121-122.

[24] YU L G,GUO X L,ZHAO H T,et al. Text Word Segmentation of Livestock and Poultry Diseases Based on BERT-BiLSTM-CRF Model[J]. Transactions of the Chinese Society for Agricultural Machinery,2024,55(2):287-294.

[25] QIU Y,YANG B. Research on micro-blog text presentation model based on word2vec and TF-IDF[C]// IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC 2021). IEEE,2021:47-51.

[26] CORTES C,VAPNIK V. Support-Vector Networks [J]. Machine Learning,1995,20:273-297.

GUO Xiaoli, born in 1998, postgraduate. Her main research interests include NLP and diagnosis of livestock and poultry diseases.

YU Ligen, born in 1985, Ph.D, professor. His main research interests include intelligent diagnosis of livestock and poultry diseases.