基于遗传算法优化支持向量机的文本自动分类方法

胡 翔[1]
HU Xiang

摘 要 在实际文本自动分类任务中，因文本的多样性和复杂性，常会遇到一些特殊的、不符合常规分类模式的文本。这些文本往往使得标引深度和标引专指度难以达到理想的平衡状态。这种不平衡导致在处理复杂模型和大规模数据时，支持向量机（SVM）模型在参数的选择上很难找到最优参数，造成分类精度较低的结果。为此，文章提出一种基于遗传算法优化支持向量机的文本自动分类方法。通过预处理来提高文本数据的质量，并引入 TF-IDF（词频 - 逆文档频率）和词共现分析，构建出高效的文本数据特征向量。利用遗传算法对 SVM 模型参数进行优化，自动搜索最佳的参数配置，提升模型的分类性能。将预处理后的文本数据输入到优化后的 SVM 模型中，模型通过学习文本数据的特征向量与类别标签之间的映射关系，实现对新文本的自动分类。实验结果表明，该方法在分类精度、Kappa 统计量和汉明损失 3 个关键指标上，均显著优于对比方法，有效提高了文本自动分类的准确性和稳定性。

关键词 遗传算法；支持向量机；文本自动分类；特征向量；优化模型参数

doi：10.3969/j.issn.1672-9528.2025.02.039

0 引言

文本因其多样性与复杂性，常常会出现一些特殊文本，这些文本不符合常规的分类模式。在处理特殊文本时，往往难以在标引深度（即详细程度）和标引专指度（即精确性）之间达成理想的平衡。对此，有学者提出基于改进 KNN 算法的档案信息文本自动分类方法 [1]。在实际的文本自动分类任务中，该方法的清洗步骤可能会删除一些看似不规则但实际上包含重要语义的内容。在处理大规模数据时，这种弊端会被放大，从而影响分类的准确性。还有学者提出基于改进卷积神经网络的多标签文本自动化分类研究 [2]。一旦模型过度拟合，将难以泛化到这些特殊文本的实际应用中，无法深入且精准地挖掘特殊文本的语义信息，同时也难以准确地对这些特殊文本进行分类。

鉴于上述不足，本文提出一种基于遗传算法优化支持向量机的文本自动分类方法。

1 基于 TF-IDF 与词共现的特征向量构建

预处理过程始于原始的待处理文本数据，首先利用分词工具将连续的字序列分解为独立的词汇单元。通过进一步的分词和清洗步骤，包括去除停用词、标点符号等噪声数据，以及可能的词形还原和词干提取，以增强文本数据的纯净度 [3]。具体预处理方法如图 1 所示。

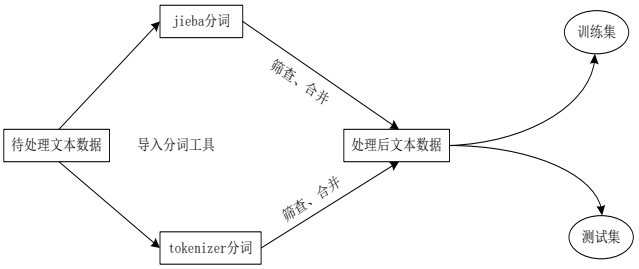

图 1 文本预处理方法

在预处理完成后，为了将文本数据转化为能够输入到机器学习模型中的数值特征向量，采用了 TF-IDF（词频 - 逆文档频率）权重计算方法。TF-IDF 首先通过词频 T 来量化词汇在单一文本中的频繁性 [4]，这一数值直接体现了词汇对于该文本的局部重要性，公式表示为：

$$T(w,d)=\frac{f_{w,d}}{\sum_{w'\in d} f_{w',d}} \tag{1}$$

式中：$T(w,d)$ 表示词汇 w 在文本 d 中的频率；$f_{w,d}$ 表示词汇 w 在文本 d 中出现的次数；$\sum_{w'\in d} f_{w',d}$ 表示文本 d 中所有词汇出现次数的总和 [5]。

逆文本频率 I 衡量了词汇在文本集合中的普遍重要性，通过调整高频但普遍存在的词汇权重，能够突出在文本中频繁出现但在整个文本集合中分布稀疏的词汇。I 用公式表示为：

$$I(w,D')=\log_2\left(1+\frac{\left|\left\{d\in D'\middle|w\in d\right\}\right|}{\left|D'\right|}\right) \tag{2}$$

1. 安庆师范大学数理学院 安徽安庆 246133

式中：$I(w,D')$表示词汇 w 在文本集合 D' 中的稀有程度；$\left|\{d\in D'|w\in d\}\right|$表示文本集合中出现了 w 的文本数量；$|D'|$ 表示文本集合的总文本数[6]。

结合词频和逆文本频率，得到 TF-IDF 权重：

$$T-I(w,d,D')=T(w,d)\cdot I(w,D') \tag{3}$$

通过计算每个词汇的 TF-IDF 权重，可以从文本集合中抽取权重最高的词汇作为关键词，构建关键词集合 K，表示为：

$$K=\{w_1,w_2,,w_k\}\in W' \quad \arg \quad \max\sum_{i=1}^{k} T-I(w_i,d,D') \tag{4}$$

为了深入分析文本内部的词汇关系，本文引入词共现与频繁词共现的概念。设$G(w_i,w_j,d)$为词汇 w_i 和 w_j 在文本 d 中共现的次数，则词共现的支持度$S(w_i,w_j,D')$用公式表示为：

$$S(w_i,w_j,D')=\frac{|D'|}{\left|D_{w_i,w_j}\right|} \tag{5}$$

利用 Apriori 算法挖掘频繁词共现集，设定最小支持度阈值 $S_{\min}$，频繁词共现集 F 表达式为：

$$F=\left\{(w_i,w_j)\middle|S(w_i,w_j,D')\geq S_{\min}\right\} \tag{6}$$

结合频繁词共现集 F 和关键词集合 K，将文本 d 表示为特征向量 $\boldsymbol{v}_d$，该向量包含了词汇的 TF-IDF 权重以及频繁词共现的支持度信息，为后续的文本分类任务提供了有效的输入。

2 基于遗传算法优化支持向量机模型参数

文本具有多样性与复杂性，这就致使在处理文本时，经常会遭遇一些特殊文本，造成标引深度与标引专指度难以达成理想的平衡状态。使得支持向量机（SVM）模型很难找到最优参数，进而造成分类精度比较低的结果。为此，本文提出利用遗传算法对 SVM 的模型参数进行优化。具体流程如下：

将 SVM 模型参数 C 和 γ 编码为遗传算法中的染色体。每个染色体 x_o 代表一个可能的参数组合 (C_o,γ_o)，其中，o 为染色体索引。在初始化阶段，随机生成一个包含多个染色体的初始种群，每个染色体都代表一个潜在的解决方案。为了评估这些潜在解决方案的优劣，本文定义适应度函数为 $f(x_o)$[7]。通过训练每个染色体所代表的支持向量机模型，并使用一个独立的验证集来评估这些模型的性能，计算出每个染色体的适应度值。

在选择过程中，本文采用轮盘赌选择法，基于每个个体 x_o 的适应度值来确定其被选为父代以产生下一代的概率 p_o：

$$p_o=\frac{f(x_o)}{\sum_{j=1}^{N} f_j(x_o)} \tag{7}$$

式中：N 表示种群中的个体总数。具有更高适应度的个体在遗传过程中拥有更大的机会被选中作为繁殖下一代的候选者，从而确保优秀基因在种群中的传递。采用单点交叉策略，通过交换两个选中染色体 x_a 和 x_b 的部分基因来生成新的染色体。这一过程可以用公式表示为：

$$\begin{cases} x_c=(\tau C_a+(1-\tau)C_b,\beta\gamma_a+(1-\beta)\gamma_b) \\ x_d=((1-\tau)C_a+\tau C_b,(1-\beta)\gamma_a+\beta\gamma_b) \end{cases} \tag{8}$$

式中：x_c 和 x_d 表示交叉后生成的新个体；C_a 和 C_b 分别表示选中进行交叉的染色体 x_a 和 x_b 的前半部分；γ_a 和 γ_b 分别表示选中进行交叉的染色体 x_a 和 x_b 的后半部分；τ 和 β 表示控制交叉点两侧基因交换的系数。

变异操作以一定的概率 P_m 随机改变染色体中的基因值，以增加种群的多样性并避免早熟收敛。公式为：

$$\begin{cases} C_o^{'}=C_o+\delta_C \\ \gamma_o^{'}=\gamma_o+\delta_\gamma \end{cases} \tag{9}$$

式中：$C_o^{'}$ 和 $\gamma_o^{'}$ 表示变异后的基因值；C_o 和 γ_o 表示变异前的基因值；δ_C 和 δ_γ 表示小的随机扰动，用于改变基因值，使得新生成的个体可能具有更好的性能。

引入自适应交叉概率机制，依据当前种群中个体的平均适应度与最高适应度之间的相对关系，自适应调整交叉概率 $P_c^{'}$：

$$P_c^{'}=P_c(1+\frac{kg_f-f_{\max}}{\varepsilon+g_f}) \tag{10}$$

式中：P_c 表示调整前的交叉概率；k 表示调整因子；g_f 表示当前种群平均适应度；$f_{\max}$ 表示当前种群最大适应度；ε 表示为避免除零错误的小正数。

同时，为了确保每一代中适应度最高的个体 x_b 能够保留到下一代，本文还采用最优保存机制。将每一代中最优的个体直接复制到下一代种群中，从而能够避免最优解的丢失。可以通过公式表示为：

$$x_b=\arg\max_{x_o\in l_{\text{all}}} f(x_o) \tag{11}$$

式中：l_{all} 表示当前种群中所有个体的集合。

持续进行优化过程，直到满足最大迭代次数 T 的终止条件：

$$\begin{cases} t\geq T \\ or \\ f(x_b^t)-f(x_b^{t-1})<\theta \end{cases} \tag{12}$$

式中：t 表示当前迭代次数；x_b^t 表示第 t 代中的最优解；θ 表示适应度变化阈值。

在完成所有遗传算法的迭代步骤后，从最终生成的种群中挑选出适应度值达到最高的个体，并将其进行解码处理，得到优化后的参数 C_b 和 γ_b。

3 优化 SVM 模型实现文本数据的自动分类

将预处理后的文本数据的特征向量和对应的类别标签输入到参数优化后的支持向量机模型中。支持向量机模型通过学习特征向量与类别标签之间的映射关系，调整模型内部的参数（如支持向量等），使得模型能够对输入的特征向量进

行准确地分类。对于新的文本，首先进行与训练数据相同的预处理操作，构建特征向量。然后将该特征向量输入到训练好的支持向量机模型中，模型根据学习到的映射关系输出新文本的类别。具体分类流程如图 2 所示。

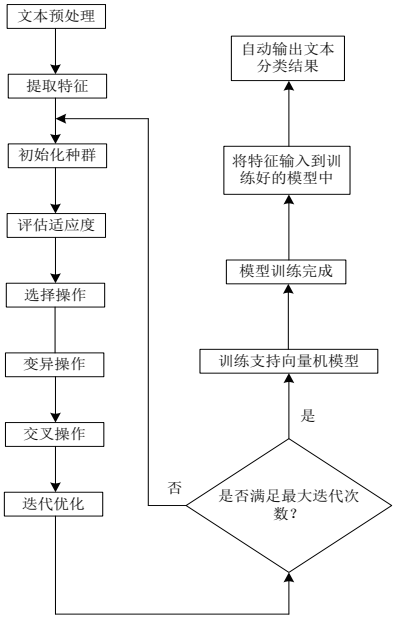

图 2 文本自动分类流程

在模型训练阶段，通过对预处理后的文本数据通过 TF-IDF 权重计算及词共现分析技术，转换成包含词汇 TF-IDF 权重和词共现支持度的特征向量。这些特征向量结合遗传算法搜索得到的最优 SVM 模型参数，用于启动 SVM 模型的训练流程。训练过程中，SVM 模型经过多次迭代与参数调整，优化其内部结构。训练完成后，SVM 模型具备将未知文本自动归类到预定义类别的能力。

对于待分类的新文本，首先进行与训练数据相同的预处理步骤，构建特征向量。然后，该特征向量输入到训练好的 SVM 模型中。模型根据学习到的特征向量与类别标签的映射关系，计算新文本最可能归属的类别标签，并输出分类结果。

4 实验

4.1 实验准备

实验的硬件环境包括 Intel Core i7 CPU、32 GB DDR4 3200 MHz 内存以及 1 TB SSD（系统盘）+2 TB HDD（数据盘）的存储配置；软件环境则采用了 Ubuntu 20.04 LTS 操作系统、Python 3.9.12 版本，并使用了 PaddleNLP NLP 库以及 Pandas、NumPy、Scikit-learn 等数据处理库。实验参数设置包括批量大小为 32、学习率为 0.001 以及训练轮次为 10。

4.2 实验数据集

构建一个多样化的文本分类数据集，覆盖了多个领域，包括新闻、科技、娱乐、体育、教育、健康以及法律等，旨在全方位考察模型在不同主题与语境下的分类能力。表 1 详细列出了各类别文本在训练集、验证集和测试集中的样本分配情况。

表 1 文本分类数据集样本分配表

类别名称	训练集样本数	验证集样本数	测试集样本数
新闻	3500	500	1000
科技	2100	300	600
娱乐	2800	400	800
体育	1750	250	500
教育	1050	150	300
健康	1400	200	400
法律	700	100	200

4.3 运行效果测试

在以上实验平台采用本文方法对测试集文本进行分类，抓取每段文本的关键词作为标签，关联分析确定类别，这一过程如图 3 所示。

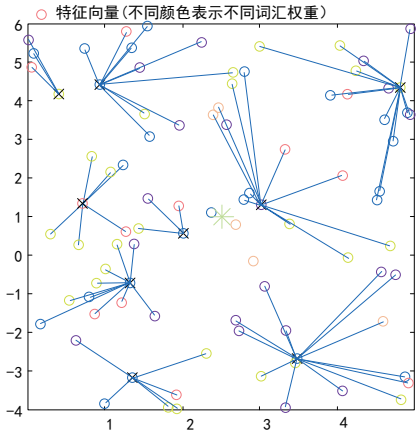

图 3 文本分类过程示意图

分析图 3，基于 TF-IDF 与词共现自动抓取每段文本的关键词，经过计算获得权重，构建出包含词汇重要性和词汇关联信息的特征向量。通过遗传算法优化支持向量机，关联分析这些关键特征向量，进一步确定每段文本所属的类别。

以测试集中某段文本为例，利用本文方法进行文本自动分类，输出结果如图 4 所示。

图 4 文本分类输出结果示意图

根据图 4 可知，运用本文所提出的方法对文本进行自动分类。通过系统的处理，该段文本被成功归类到了预定的类别中，并且分类的结果以直观的输出。表明本文方法具有有效性，可实际应用。

4.4 对比实验

实验中，将本文提出的方法作为方法 1，基于改进 KNN 算法作为方法 2，基于改进卷积神经网络方法作为方法 3，对比结果如表 2 所示。

表 2 不同方法在文本自动分类中的性能对比

类别名称	分类精度			Kappa 统计量			汉明损失		
	方法 1	方法 2	方法 3	方法 1	方法 2	方法 3	方法 1	方法 2	方法 3
新闻	0.81	0.68	0.72	0.56	0.39	0.45	0.14	0.22	0.19
科技	0.89	0.76	0.85	0.78	0.61	0.72	0.06	0.11	0.08
娱乐	0.74	0.59	0.63	0.42	0.27	0.31	0.21	0.29	0.26
体育	0.91	0.82	0.88	0.83	0.70	0.77	0.04	0.09	0.07
教育	0.61	0.55	0.48	0.25	0.18	0.09	0.31	0.34	0.41
健康	0.84	0.78	0.71	0.73	0.60	0.54	0.09	0.12	0.18
法律	0.93	0.87	0.90	0.86	0.79	0.81	0.02	0.04	0.05

根据表 2，本文提出的文本分类方法相较于其他方法展现出显著优势。在新闻、科技、体育及法律等类别上，其分类精度与 Kappa 统计量均达高水平，表明分类效果稳定且准确。从汉明损失看，本文方法损失值低，进一步验证其优越性。该方法采用遗传算法优化 SVM 参数，自动搜索最佳配置，显著提升分类性能，使模型表现更稳定可靠。

5 结语

本文提出了一种创新的文本自动分类方法，该方法结合了遗传算法与支持向量机的优势。借助遗传算法的全局搜索和优化能力，成功解决了 SVM 参数调优的复杂问题，从而显著提升了文本分类的准确性。此方法在处理大规模文本数据时表现优异，为文本自动分类领域带来了新的视角和技术革新。展望未来，将进一步深化遗传算法与 SVM 结合机制的研究，探索更多创新性优化策略，以期提升分类模型的性能和稳定性。

参考文献：

[1] 潘国炀 . 基于改进 KNN 算法的档案信息文本自动分类方法研究 [J]. 信息与电脑 (理论版), 2024, 36(4): 71-73.
[2] 刘影 , 余进 , 陈莉 . 基于改进卷积神经网络的多标签文本自动化分类研究 [J]. 自动化与仪器仪表 ,2023(11):62-66.
[3] 李淑红 , 邓明明 , 孙社兵 , 等 . 基于注意力机制和 CNN-BiLSTM 模型的在线协作讨论交互文本自动分类 [J]. 现代信息科技 , 2023,7(13):26-31.
[4] 索南多杰 , 官却多杰 , 拉玛杰 , 等 . 基于深度学习的藏文文本自动分类研究 [J]. 青海科技 ,2023,30(3):192-196.
[5] 刘蕾 , 田鑫宇 , 朱大洲 . 基于 SSA-SVM 的营养健康信息文本分类研究 [J]. 计算机时代 ,2023(6):82-86.
[6] 徐涯昕 , 何泽恩 , 徐绪堪 . 基于 CNN-BiLSTM 网络的数控机床故障文本自动分类 [J]. 计算机与现代化 ,2023(4):7-14.
[7] 陈玉天 , 陈洋 , 梁恒瑞 , 等 . 基于 TI-LSTM 的文本自动分类算法及应用 [J]. 长春理工大学学报 (自然科学版), 2023, 46(1):130-136.

【作者简介】

胡翔（1980—），女，安徽马鞍山人，硕士，讲师，研究方向：计算数学。

（收稿日期：2024-10-12）

（上接第 163 页）

候，完全可以基于规则模板的算法为主，贝叶斯算法为辅助（作为补充算法来解决某些特殊的语义理解问题），以保障系统整体的健壮性。

4 结语

本文为解决基于小样本短句意图识别问题，设计了基于规则模板的意图识别算法和通过 TextRank 优化分类器训练流程而改进的朴素贝叶斯意图识别算法。通过实验分析算法性能，结果表明改进后的朴素贝叶斯算法确实得到了一定程度的性能优化；基于规则模板的算法在不需要样本训练的前提下，搭建灵活性更强，代码维护性和复用性更高，更适合中小企业应用。两种算法都可以为中小企业智能问答系统的研发提供一种实现途径，有一定的实践和参考价值。

参考文献：

[1] AHO A V, CORASICK M J. Efficient string matching: an aid to bibliographic searach [J]. Communications of the ACM, 1975, 18(6): 333-340.
[2] 何晗 . 自然语言处理入门 [M]. 北京 : 人民邮电出版社 ,2019.
[3] 冯建周 . 自然语言处理 [M]. 北京 : 中国水利水电出版社 , 2022.
[4] 刘挺 , 秦兵 , 赵军 , 等 . 自然语言处理 [M]. 北京 : 高等教育出版社 , 2022.
[5] 周志华 . 机器学习 [M]. 北京 : 清华大学出版社 , 2016.
[6] JAMES H M, DANIEL J. 语音与语言处理自然语言处理、计算语言学和语音识别导论 [M]. 北京 : 人民邮电出版社 , 2010.

【作者简介】

王炳翔（1988—），男，陕西西安人，硕士研究生，高级工程师，研究方向：系统规划与管理、知识图谱、自然语言处理等。

（收稿日期：2024-11-05）