基于特征聚类与降维的新闻文本智能分类算法

张鸿彦 [1]
ZHANG Hongyan

摘 要 文本分类是当前智能处理新闻数据信息的一种有效手段。为了提高信息处理的准确性和有效性，提出基于特征聚类与降维的新闻文本智能分类算法。首先，利用汉语词汇分析系统 ICTCLAS 对新闻文本分词作预处理，去除其中的停用词，并区分新闻文本的词性。然后，利用权重函数对新闻文本作降维处理，约简新闻文本的关键词集。最后，采用聚类分析技术对新闻文本特征实施聚类，得到不同类别的新闻文本。实验表明，所提出的算法得到的分类结果准确率在 96% 以上，召回率在 98% 以上，说明所提出的算法能够实现对新闻文本智能精准分类，具有良好的应用前景。

关键词 特征聚类；降维处理；新闻文本；智能分类；汉语词汇分析系统；权重函数

doi：10.3969/j.issn.1672-9528.2024.04.023

0 引言

针对海量的新闻信息，如何有效实现文本分类、提高信息检索的准确性和效率，一直是计算机领域研究的热点问题之一。

传统的新闻文本分类方法主要基于人工分类或者简单的机器分类，这些方法不仅耗时、耗力，而且准确率也不高。因此，研究一种智能化的文本分类算法至关重要。文献 [1] 中对 BERT 预训练模型进行微调，利用微调后的模型获得文本的向量表示形式，在分析文本特征的全局语义关系后，将文本向量表示数据输入 CNN 模型中实现文本分类。但是文本分类任务的复杂度会受到多种因素的影响，如文本长度、语境依赖性、领域差异等。对于某些复杂的分类任务，使用简单的 CNN 模型可能无法捕捉到足够的语义信息，从而导致分类准确性低。文献 [2] 中根据文本与主题之间的关联性构建邻接矩阵，然后利用词级图注意力网络获取文本的向量表示。在通过最大池化处理提取文本目标向量后，判别文本的类别特征。然而，尽管词级图注意力网络在处理文本的上下文信息方面具有一定优势，但它也有其局限性。例如，如果输入的文本长度很长，那么词级图注意力网络可能无法捕捉到全局的语义信息，从而导致分类准确性低。文献 [3] 中首先构建了文本数据集，对文本位置信息编码并融入词嵌入向量，然后分别利用 BiGRU 和 CNN 提取文本数据间的相关性，在重新分配文本权重的基础上通过 Softmax 实现分类。然而，BiGRU 和 CNN 是用于文本特征提取的常见方法，但它们可能无法捕捉到一些细微的语义特征或上下文依赖关系。因此，模型在提取文本特征时可能没有捕捉到足够的相关性信息，导致分类结果召回率较低。文献 [4] 中利用预训练后的 MacBERT 模型得到文本中的动态词向量，首先将提取结果输入 BiLSTM 模型中提取上下文关系特征，然后结合注意力机制分配不同的权重值，最后通过 Softmax 分类器得到分类结果。尽管该方案使用了 BiLSTM 模型来捕捉文本中的上下文关系特征，但该模型可能无法捕获较长文本中的全局依赖关系。特别是对于需要考虑大范围上下文的任务，使用 BiLSTM 模型可能会导致召回率降低。

针对上述问题，本文提出基于特征聚类与降维的新闻文本智能分类算法，为新闻文本智能分类提供参考依据。

1 新闻文本分词处理方法设计

词汇是具有语义功能的最小语言单位。通常在英文文本中，词与词之间有天然的间隔，所以不会出现分词问题。然而，由于汉语是以文字为基础的文本书写单元，词与词之间无显著区别，且汉字构成词的规律十分复杂，因此，中文词汇的形态学研究是中文信息处理的基础和核心。考虑到此研究算法是以特征聚类为重点，为了更好地对新闻文本特征聚类分析，将中国科学院计算机研究所开发的汉语词汇分析系统 ICTCLAS 作为新闻文本中文自动分词工具。利用 ICTCLAS 完成对新闻文本分词、关键词提取、停用词过滤以及关键词列表输出等处理。

在原始语料库的基础上，利用 ICTCLAS3.0 中文分词系统对新闻文本进行分词处理。该系统在分析文本时，不仅可

1. 河南工程学院 河南郑州 450000

[基金项目] 河南省高等学校重点科研项目（22A520022）；河南工程学院横向科研项目：HKJ2021139

以将各个词汇进行单独的分词处理，而且得出每个关键词的详细信息，包括词语本身、该词语的词性，以及系统自动为该词计算出的权重。然而，考虑到某些词性的关键词对于新闻文本智能分类的实际作用并不大，比如介词、分词、量词、代词和连词等，本文决定在后续的处理过程中，只保留那些词性为名词、动词和形容词的关键词[5]。

在文本分词过程中，“是”“为”“啊”等高频词的使用频率很高，但它们在语义上不起到决定性的作用，这种词被称作“停用词”。若不对停用词进行筛选，则会产生大量的噪音。所以，当有一个停用词出现时，就将这个词从分词结果集中删除，然后再构建一个停用词表过滤文本，就能对文本切分结果进行初步筛选。

为了实现上述目的，对新闻文本中的每个关键词进行词性判断。如果一个关键词的词性满足要求，那么就将其记录并保存下来。反之，如果其词性不在所需的范围内，那么就将其丢弃。这样，就能得到一个经过筛选和优化后的关键词集合，这些关键词都是名词、动词或形容词，对于后续的新闻文本分类工作将起到重要的帮助作用。

2 新闻文本特征降维处理设计

相比于其它语言，汉语言更加强调词义，句式结构比较复杂，并且文本中含有大量的语义相同的词语。这一特点使得新闻文本中蕴含着多种信息，比如“这个人好说话”在不同的上下文中所表示的意义是不一样的，可以表示“此人爱讲话”与“此人待人和善”两种含义，“好”一词在两种语境中所表现出来的意义也各不相同。再比如，词语“电脑”与“计算机”通常是可以互换的。由此可以看出，新闻文本中蕴含着丰富的、可理解的语义信息，具有复杂性与多样性。因此，在处理新闻文本时，必须充分考虑词汇的语义信息对新闻文本表示、文本分类等方面的影响，从而达到更好的分类结果[6]。

在新闻文本智能分类过程中，首先通过降低新闻文本特征维数，使得分词后得到的新闻文本关键词集更能准确地从语义层次上表达文本含义。经过分词预处理过程后，新闻文本可以被表示成一个特征词 - 文本矩阵，可用公式表示为：

$$\boldsymbol{A}=\left[a_{ij}\right]_{m\times n} \tag{1}$$

式中：$\boldsymbol{A}$ 表示新闻特征词 - 文本矩阵；a_{ij} 表示第 i 个特征词在第 j 个文档中出现的频度，其为非负数；m 表示新闻文本集中包含的所有不同的特征词数量；n 表示新闻文本集中的文本数量。矩阵中，每一个特征词对应着特征词 - 文本矩阵的一行，每一个新闻文本对应着矩阵的一列。对于任意一个新闻文本，其是由数量明确的关键词而不是所有关键词都参与构成的，因此，可以将新闻关键词 - 文本矩阵视为一个稀疏矩阵[7]。

为了降低矩阵的稀疏性，根据词语权重对矩阵降维，词语矩阵由词语全局权重和词语局部权重组成，用公式可以表示为：

$$W\left(i,j\right)=LW\left(i,j\right)+GW\left(i,j\right) \tag{2}$$

式中：$W(i,j)$ 表示新闻文本词语权重函数；$LW(i,j)$ 表示新闻文本词语局部权重，表征词语 i 对于新闻文本 j 的重要程度；$GW(i,j)$ 表示新闻文本词语全局权重，表征词语 i 对于新闻文本集中所有文本的重要程度。本文拟采用对数词频法对新闻文本词语局部权重计算，以熵权重作为新闻文本词语全局权重，二者的计算公式为：

$$\begin{cases} LW\left(i,j\right)=\log\left(1+\dfrac{p}{k}\right) \\ GW\left(i,j\right)=1-\sum\limits_{j=1}\dfrac{L\log\left(L\right)}{\log N} \end{cases} \tag{3}$$

式中：p 表示词语在新闻文本中出现的次数；k 表示新闻文本中次数的总数；L 表示词语在新闻文本集中出现的频率；N 表示新闻文本集中的文本总数。根据新闻文本词语权重函数，对新闻关键词 - 文本矩阵降维，其用公式表示为：

$$\boldsymbol{A}^{*}=W\left(i,j\right)\times\boldsymbol{A} \tag{4}$$

式中：$\boldsymbol{A}^{*}$ 表示降维后的新闻关键词 - 文本矩阵。

通过对 $\boldsymbol{A}$ 进行降维处理，破坏了原新闻关键词 - 文本矩阵塌陷，从而生成了规模明显缩小的近似矩阵，从众多原始特征中选择最能够反映新闻文本类别统计特性的关键词，降低新闻文本空间的维数，本质上是对新闻文本关键词的约简。

3 基于特征聚类的文本分类方法设计

在上述基础上，采用聚类分析技术对新闻文本特征聚类，将新闻文本划分到不同类别中，从而实现对新闻文本的智能分类[8]。特征聚类过程包括奇异值分解、计算词语语义关系、语义特征聚类三部分，虽然上文利用权重函数对新闻文本特征进行了降维，但是文本中仍旧会存在一些奇异特征，因此，采用隐性语义索引（LSI）的方式对新闻文本奇异值分解，将新闻文本词语映射到低维的向量空间。根据 LSI 定理，任何一个文本的矩阵可以被分解为一个对角矩阵和两个正交矩阵的乘积，用公式表示为：

$$\boldsymbol{A}^{**}=\boldsymbol{T}\times\boldsymbol{D}\times\boldsymbol{S}\times\boldsymbol{A}^{*} \tag{5}$$

式中：$\boldsymbol{T}$ 表示新闻关键词 - 文本矩阵的左奇异矩阵，为正交矩阵；$\boldsymbol{D}$ 表示新闻文本关键词奇异值矩阵，为对角矩阵；$\boldsymbol{S}$ 表示新闻关键词 - 文本矩阵的右奇异矩阵，为正交矩阵[9]。可以将矩阵 $\boldsymbol{A}^{**}$ 视为新闻关键词 - 文本矩阵的 SVD 式，通过保留对角矩阵中最大的对角元素，同时删除左奇异矩阵与右奇异矩阵中对应的行和列，去除新闻关键词 - 文本矩阵中奇异值，保留矩阵大部分语义结构，同时去除不该保留的具有

奇异的关键词，构建新闻文本潜在语义空间。

在对新闻文本奇异值分解的基础上，利用词语语义关系矩阵表征出新闻文本词语的语义相似度，其用公式表示为：

$$\eta = \boldsymbol{A}^{**} \times \boldsymbol{R}^{T} \tag{6}$$

式中：η 表示新闻文本中词语的语义相似程度；$\boldsymbol{R}^{\mathrm{T}}$ 表示词语语义关系矩阵 [10]。

根据新闻文本语义相似度特征，将具有共同特征的新闻文本聚成一个文本簇。簇是一个类别单位，簇中的文本相似度较高。将具有相同特征的新闻文本划分为一个簇中，自动形成一个词典。该词典可以很好地解释关键词与文本、主体与关键词之间的关系 [11]。

本研究采用 K-means 聚类，其核心思想是根据新闻文本词语语义相似特征通过迭代将新闻文本集中的文本分为多个聚类，使不同类簇的新闻文本之间的相似度很大，同一类簇内部的新闻文本相似度很小。假设每个新闻文本对象代表着一个类簇中心，根据新闻文本词语语义相似度确定新闻文本对象与聚类中心的距离，将其赋给距离最近的类簇中。将该过程不断迭代，直到聚类算法收敛，特征聚类准则采用误差平方准则，其计算公式为：

$$E = \eta \left| f - d \right|^2 \tag{7}$$

式中：E 表示新闻文本集中所有对象的误差平方和；f 表示新闻关键词 - 文本矩阵空间中的点，即给定的新闻文本对象；d 表示类簇的中心 [12]。

假设聚类簇数为 α，即新闻文本类别数量为 α，待分类的新闻文本数量为 β，则可以得到聚类阈值，其用公式表示为：

$$\varpi = \frac{\alpha}{\beta \times E} \tag{8}$$

式中：ϖ 表示新闻文本特征聚类阈值。

将计算到的新闻文本对象与聚类中心的误差平方和聚类阈值比较，如果大于阈值，则表示该新闻文本对象不在类簇范围内；如果小于阈值，则表示该新闻文本对象属于类簇范围内，将其划分到该类簇中。按照上述流程对 β 个新闻文本划分到 α 个类簇中，实现基于特征聚类与降维的新闻文本智能分类，进而完成文本分类。

4 实验论证

为了验证基于特征聚类与降维的新闻文本智能分类算法的实际应用性能，设计如下实验。

4.1 实验数据与环境

在 TU95、YU75、OP954、ER9W7 四个基本数据集上评估本文设计算法的有效性，四个数据集具体情况如表 1 所示。

表 1 新闻文本数据集信息

数据集	大小 /GB	总词数 / 个	标签数 / 个
TU95	1.26	7589	32
YU75	1.56	8216	25
OP954	1.35	7862	21
ER9W7	1.42	8016	19

TU95 数据集是大约 8000 个新闻文档，由路透社新闻专线文本组成，包含 32 个类别。

YU75 数据集隶属于 IYHFAH 数据库，包括 5000 个文本，数据集文本相似度较高，分类难度大。

OP954 数据集属于情感分析类数据，包含 21 个类别。

ER9W7 数据集来源于医学文献数据库，包含 19 类临床疾病医学文本。

实验中，新闻文本智能分类算法应用基于 Python 框架，操作系统为 Windows2010，Intel Core i8-2907 CPU，32 GB 内存，采用 PyCharm 开发工具，使用 Python3.6 开发语言。实验过程中，算法随机选取数据集中的 50% 数据作为验证集，对四个数据集中的数据进行分类，记录分类结果。

4.2 实验结果与讨论

为了验证基于特征聚类与降维的新闻文本智能分类算法的优越性，使用准确率（*Acc*）与召回率（*Recall*）作为评估指标，两个指标计算公式为：

$$Acc = \frac{T_P + T_N}{T_P + F_P + F_N + T_N} \tag{9}$$

$$\mathrm{Re}\,call = \frac{T_P}{T_P + F_N} \tag{10}$$

式中：*Acc* 表示新闻文本智能分类准确率；T_P 表示被正确分类的正样本数量；T_N 表示被正确分类的负样本数量；F_P 表示被错误分类的负样本数量；F_N 表示被错误分类的正样本数量；*Recall* 表示新闻文本智能分类召回率。通常准确率与召回率数值越大，新闻文本智能分类效果越好，算法收敛性越强。

为了验证本文算法的有效性，在同样实验环境中将其与文献 [1]、文献 [2] 中的算法进行对比实验。在四个公开数据集上，对算法的准确率与召回率展开检验。

以 ER9W7 数据集为例，给出不同算法的准确率、召回率结果分别如表 2、图 1 所示。

表 2 新闻文本智能分类的准确率

文本数量	本文算法 /%	文献 [1] 算法 /%	文献 [2] 算法 /%
1000	98.95	84.15	75.62
2000	98.26	83.26	75.15
3000	98.41	82.46	74.26
4000	97.89	81.26	74.03
5000	97.86	81.03	73.26
6000	97.56	80.64	73.56
7000	97.42	80.42	73.61
8000	97.15	80.23	72.15
9000	97.02	80.15	71.25
10 000	96.25	80.03	71.62

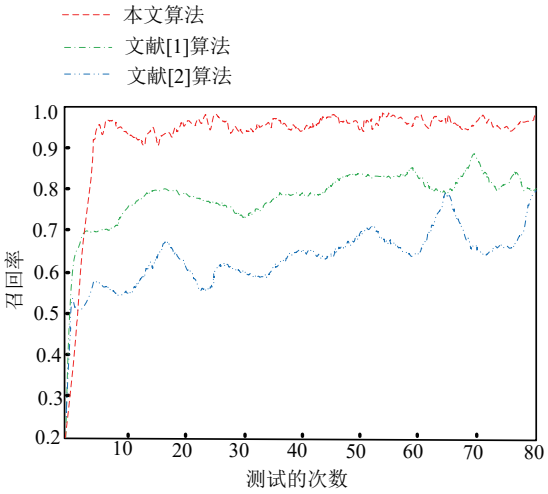

图 1 新闻文本智能分类结果的召回率

表 2、图 1 分别给出了本文设计算法与两种主流算法在不同数据集上的准确率与召回率的对比实验结果。通过与两种文献方法进行比较可以发现，本文设计算法在准确率、召回率两个方面取得了显著的提升。针对分类结果的准确率，本文设计算法比文献 [1] 算法高出将近 14%，比文献 [2] 算法高出将近 18%；针对分类结果的召回率，本文设计算法的召回率接近 1.0，而文献 [1] 算法、文献 [2] 算法的召回率曲线始终处于本文设计算法之下。

通过以上对比，本文设计算法在新闻文本智能分类上更有优势，分类精度取得了明显的提升，这是因为本文设计的算法使用基于特征聚类与降维，可以有效捕捉新闻文本单词的语义特征和语法特征，根据特征对新闻文本分类，在一定程度上保证了分类精度，从而获得了更好的分类效果。

5 结语

本文提出了一种基于特征聚类与降维的新闻文本智能分类算法。该算法首先对新闻文本进行特征提取，然后利用 K-means 算法进行特征聚类，以实现特征降维。之后，使用降维后的特征进行新闻文本分类。

实验结果表明，该算法能够有效地提高新闻文本分类的准确率。然而，尽管本文提出的算法取得了一定的成果，但仍存在一些问题，需要进一步改进和完善。例如，在接下来的研究中，可以考虑使用更加先进的聚类算法，如谱聚类、层次聚类等，以提高特征聚类的效果。此外，也可以尝试将其他类型的特征，如语义特征、情感特征等，引入到新闻文本分类中，以期进一步提高分类性能。

参考文献：

[1] 景永霞，苟和平，刘强．基于 BERT 语义分析和 CNN 的短文本分类研究 [J]. 洛阳理工学院学报 (自然科学版), 2023, 33 (4): 78-83.

[2] 杨春霞，马文文，徐奔，等．融合标签信息的分层图注意力网络文本分类模型 [J]. 计算机工程与科学，2023, 45 (11): 2018-2026.

[3] 周俊杰，许鸿奎，卢江坤，等．引入位置信息和 Attention 机制的诈骗电话文本分类 [J]. 小型微型计算机系统，2023, 44 (11): 2502-2509.

[4] 王道康，张吴波．基于 MacBERT-BiLSTM 和注意力机制的短文本分类研究 [J]. 现代电子技术，2023, 46 (21): 123-128.

[5] 郭顺利，苏新宁，房旭辉．融合 NER 和 Apriori 算法的游记文本关联知识挖掘及推荐服务研究 [J]. 现代情报，2023, 43 (11): 123-134.

[6] 陆潜慧，张羽，王梦灵，等．基于改进循环池化网络的核电质量文本分类方法 [J]. 计算机应用 ,2023(7):1-9.

[7] 刘爱琴，郭少鹏，张卓星．基于 LDA 模型融合 Catboost 算法的文本自动分类系统设计与实现 [J]. 国家图书馆学刊，2023, 32 (5): 84-92.

[8] 杨春霞，黄昱锟，闫晗，等．融合 GAT 与头尾标签的多标签文本分类模型 [J]. 计算机工程与应用 ,2023(10): 1-12.

[9] 刘勇，杜建强，罗计根，等．基于语义筛选的 ALBERT-TextCNN 中医文本多标签分类研究 [J]. 现代信息科技，2023, 7 (19): 123-128.

[10] 杨兴锐，赵寿为，张如学，等．改进 BERT 词向量的 BiLSTM-Attention 文本分类模型 [J]. 传感器与微系统，2023, 42 (10): 160-164.

[11] 王宣军，于虹，祁兵，等．基于注意力机制的混合神经网络电力设备缺陷文本挖掘方法 [J]. 电力信息与通信技术，2023, 21 (9): 44-51.

[12] 刘新忠，赵澳庆，谢文武，等．基于 BERT-GAT-CorNet 多标签中文短文本分类方法 [J]. 计算机应用，2023, 43(S2): 18-21.

【作者简介】

张鸿彦（1977—），女，山西平遥人，硕士，副教授，研究方向：计算机应用。

（收稿日期：2024-03-05）