DOI:10.19353/j.cnki.dzsj.2019.06.009

基于卡方统计改进的TF-IDF的文本分类的研究

北京工业大学信息学部软件学院
王露瑶 张涛 陈才 朱安琥 罗启明

TF-IDF算法作为最常见的特征权重计算方法被广泛使用。传统TF-IDF特征提取方法在文本分类任务中缺乏对类之间分布差异的体现。基于此种情况，立足于传统TF-IDF算法中根据词频来选择特征词的特性，本文提出一种新的基于卡方统计的特征词提取算法并通过改进后的新方法对文本分类模型进行评估。实验结果表明，新方法在查准率、查全率、F1值和ROC_AUC等评估结果上较传统特征提取方法有明显优化。

1.引言

近年来，文本分类任务主要包括文本预处理、特征选择、模型训练以及效果评估（蒋健，文本分类中特征提取和特征加权方法研究：重庆大学，2010）。文本预处理阶段特别是特征词提取部分很大程度的影响了分类的准确性，所以国内外研究学者致力于改进特征词的提取工作。Salton（Salton G，McGill M J.Introduction to modern information retrieval：Communications of the ACM，1983）在1988年提出结合词频权重和反文档频率计算方法，即TF-IDF算法。

传统TF-IDF权重计算方法存在以下两个问题：①TF仅仅从词频的角度考虑文本信息，缺乏对特征词上下文的处理，忽略了文本的结构；②IDF忽略了特征项在类别之间的分布情况的不足。本文提出用卡方统计方法来描述特征词在类间的分布信息和分类能力，并且以农产品舆情信息为例，通过对比改进前后的TF-IDF在褒义，中性，贬义分类上的评估指标，验证了改进特征提取方法的正确性以及有效性。

2.传统TF-IDF算法介绍

2.1 传统TF-IDF的计算方法

传统文本特征在经过特征选择之后，会选择出具有类别区分度高的特征词汇。具体的体现为：对类别区分度高的特征赋予较高的权重，这样的优势在于最大程度地体现文本类别之间的差异，提高分类的性能。

在TF-IDF算法中，词频表示特征词在文章中出现的次数，通常用TF表示。它的特点是：如果特征词在某篇文章中出现的次数较多，那么该特征就能很好表达出该文本的主要信息，从而适合进行文本分类。IDF表示反文档频率，其特点表现为：如果含有同一词语的文档的数量越多，那么该特征词的分类能力越差。例如：虽然一些人称代词、感叹词、介词出现的频次很多，但是却并没有对文本区分能力。该算法的计算方法如公式（2-1）所示。

$$\mathrm{IDF}=\log\left(\frac{N}{n_k}+1\right) \tag{2-1}$$

公式（2-1）中N表示整个训练样本中文本总数，nk表示包含特征词的文本数。

传统TF-IDF计算方法中，TF词频统计方法用来描述高频特征，而这些高频特征往往是对文本分类没有帮助的噪声词，有些低频词能很好地表示文本信息却因为出现频率低的原因被忽略掉。IDF增强出现频率低的特征词的权重，在一定程度上弥补TF的不足。TF-IDF加权方法将两方法结合在一起，公式如下：

$$W=TF*\log\left(\frac{N}{n_k}+1\right) \tag{2-2}$$

2.2 传统TF-IDF的不足

传统的TF-IDF忽略了特征词在类别间的分布情况（Leilei Chu，Hui Gao，Wenbo Chang.A New Feature Weighting Method Based on Probability Distribution in Imbalanced Text Classification：2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery，2010）。例如：特征词a在积极类文本中出现的次数为m，在消极类文本中出现的次数为n，则特征词在文本集合中出现的总次数记为l=m+n。当m数值很大，即特征词在积极类的文本中出现次数多，此时特征词a能很好反映积极类文本。但在TF-IDF中，由公式（2-1）可知，当m和l越大时，IDF反而越小，故特征词a不再具有代表性。

不仅如此，当特征词a均匀地分布在每个类别之中时，其并不具备类别代表性，故特征词a应该赋予较低的权重。但是由公式（2-2）可知，此时IDF的值却很大，这并不符合实际情况。出现这种结果的原因是没有考虑TF-IDF在类别间的情况。

最后，TF-IDF没有考虑特征词在类别内部的分布情况。正常情况下，分布不稳定的特征词语应该比分布比较稳定的特征词语赋予更低的权重，同样如果某个特征词仅仅出现在某一类文本中，该词汇对文本分类没有区分度，应赋予较低权重，可是在传统的TF-IDF中却没有体现。

3.基于卡方改进的TF-IDF

3.1 卡方计算方法

卡方统计算法(CHI Square，CHI)用于计量两个变量的偏差程度，其中理论值和实际值相互独立。如果偏差较大的话，那么这两个变量相互独立。其特征与类别关系表如表1所示。

表1 特征与类别关系表

	属于类别 Ci	不属于类别 Ci	总数
包含特征词t的文本	A	B	A+B
不包含特征词t的文本	C	D	C+D
总数	A+C	B+D	N=A+B+C+D

基于特征词t的计算方法如公式（3-1）、（3-2）、（3-3）、（3-4）所示：

$$E_{11}=\left(\frac{A+B}{N}\right)*(A+C) \tag{3-1}$$

$$E_{12}=\left(\frac{A+B}{N}\right)*(B+D) \tag{3-2}$$

$$E_{21}=\left(\frac{C+D}{N}\right)*(A+C) \tag{3-3}$$

$$E_{22}=\left(\frac{C+D}{N}\right)*(B+D) \tag{3-4}$$

根据偏差计算方式计算出类别Ci包含特征词t的偏差如公式（3-5）所示。

$$D_{11}=\frac{(A-E_{11})2}{E_{11}} \quad (3\text{-}5)$$

同理可求出D12、D21、D22，带入并化简得特征词汇t和类别Ci的CHI值分方法如公式（3-6）所示。

$$CHI(t,Ci)=\frac{(A*D-C*B)^2*N}{(A+C)*(B+D)*(A+B)*(C+D)} \quad (3\text{-}6)$$

由公式（3-6）可知，当特征词汇t和类别Ci相互独立时，A*D-C*B等于0，CHI也等于0。当特征词汇t和类别Ci越相关，CHI的值越大。

Yang（Yiming Yang，Jan O.Pedersen. A Comparative Study on Feature Selection in Text Categorization：Proceedings of the 14th International Conference on Machine Learning(ICML)，1997）的研究表明，卡方统计是现有的文本分类中效果最好的特征选择方法之一。一个特征项在某类别中卡方值越大，则该特征词汇越具有区分性。但卡方统计也有不足之处，下面将对其缺点进行详细分析。

（1）传统的卡方统计方法忽略对低频词的考虑。具体表现为如果一篇文本中出现的特征词的次数比较少，而在多个文档中反复出现，这样的词并不能区分文本类别，因此所进行的分类不准确。

（2）传统的卡方未考虑特征词在类别内部分布均匀的状况。具体体现为卡方统计公式中并没有体现特征词在某类文档中均匀出现时应该被赋予的高权重，而只关注在其他类特征词出现频率高而在该类出现频率低时权重赋予方式（李原，中文文本分类中分词和特征选择方法研究：吉林大学，2011）。

（3）传统的卡方统计方法更倾向于选择与类别负相关的特征词。由公式（3-6）可知，特征词汇在其他类别中出现多，而在指定类出现得少，即B与C较大，A与D较小，可得BC>>AD。这样计算出来的CHI值比较高，而此时用对其他类区分程度的词作为特征词很不可靠，故不能给文本分类的性能带来提高。

3.2 基于卡方改进的TF–IDF

针对传统的TF-IDF忽略了特征词在类别间的分布情况，本节进行改进。

由上节可知，传统的TF-IDF权重计算方法忽略了特征词汇在类别之间分布差异，特征词的卡方统计值能够很好地描述特征词在类别之间的分布信息。具体体现为特征词汇分类能力与它的卡方值成正比。基于此种情况，我们引入卡方统计方法来提升特征词在类间的分类能力。基于卡方统计的特征词权重改进方法如公式（3-7）所示。

$$TFCHI=TF*\log\left(\frac{N}{n_k}+1\right)*\frac{(A*D-C*B)^2*N}{(A+C)*(B+D)*(A+B)*(C+D)} \quad (3\text{-}7)$$

由公式（3-7）可知，基于卡方统计的TF-IDF权重计算方法弥补了未考虑特征词汇在类间分布的缺陷，可以将在类间分布均匀但并不具备类别区分能力的词赋予较低的权重，在一定程度上改进了传统计算方法，有效地提高了权重计算的准确性和文本分类的正确性。

4.实验结果

4.1 实验数据来源

本文以对农产品舆情信息进行分类为例，通过爬虫技术从网上获取农业新闻的舆情信息，将舆情信息按褒义、贬义、中性分成三类，各取20000条作为训练集，5000条作为预测集。

4.2 评价指标

本实验采取查准率、查全率、F1测量值和ROC_AUC作为评价文本分类效果的主要指标。这些指标不仅能够反映文本分类模型的分类效果，而且在保持其它功能模块一样的情况下，有利于分析特征提取方法对分类任务的影响（Sebastiani F.Text Categorization：Encyclopedia of Data-base Technologies & Applications，2005）。

（1）查准率与查全率

查准率和查全率基于表2进行计算。查准率P是指文本经过分类模型后，在某一类所有文本中，文本真实类别与判定类别相同的文本所占的比率，其重点考虑分类的准确率，查准率的计算方法如公式（4-1）为：

$$P=\frac{TP}{TP+FP} \quad (4\text{-}1)$$

表2 特征与类别关系表

	预测为正例	预测为负例
真实为正例	TP	FN
真实为负例	FP	TN

查全率是指真实类别为某类别的所有文本中，其经过分类模型后仍判定为此类别所占的比率，重点考虑分类的完整性。其计算方法如公式（4-2）所示。

$$R=\frac{TP}{TP+FN} \quad (4\text{-}2)$$

（2）F1测量值

在实际应用中一般使用F1测量值（F-Measure）来全面考虑查全率和查准率，其计算方法如公式（4-3）所示。

$$F_\beta=\frac{(1+\beta^2)*P*R}{(\beta^2*P)+R} \quad (4\text{-}3)$$

其中，β>1时，查全率更有影响；β=1时，退化为标准的F1；查准率更有影响。

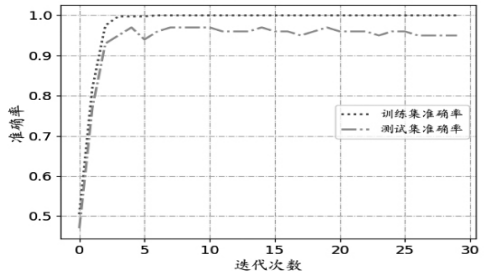
（3）ROC_AUC

ROC_AUC是“真正例率”与“假正例率”的特性曲线。其横轴为“假正例率”，纵轴为“真正例率”。AUC表示ROC曲线下的面积。ROC曲线的优势在于：当数据集中出现正负比例失调，测试样本随时间变化的情况下，ROC曲线依然保留原本特性。AUC值是一个概率值，当随机选择一个样本，经过分类算法后根据评价指标会得到一个分数，它表示正样本在负样本前的概率。AUC值越大，当前分类算法越有可能将正样本排在负样本前面，从而能够更好地分类。

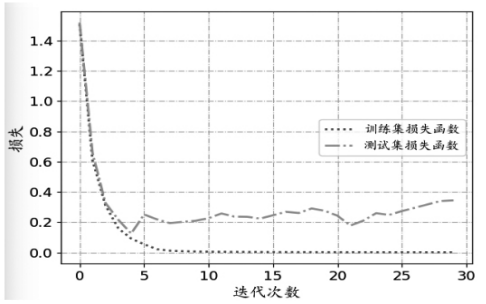
4.3 实验结果

本实验基于传统的TF-IDF和改进的TF-IDF的两种特征提取方法分别对实验数据进行特征提取，再利用分类算法进行训练并验证得到分类结果。

（下转第28页）



(a) 识别准确率变化情况



(b) 损失变化情况

图6 小波变换提取特征向量的训练和识别情况

3.3 对比实验分析

同时为验证该方法采用声发射信号参数作为神经网络模型输入样本特征向量的可靠性，对比分析了采用小波变换方法（李学军，廖传军，褚福磊，适于声发射信号故障特征提取的小波函数：机械工程学报，2008；王楠，基于应力波与小波分析的低速滚动轴承故障诊断研究：振动工程学报，2007）提取声发射信号的特征，作为CNN模型输入样本特征向量的方法。小波变换方法提取声发射信号特征时，采用db10小波基函数，将信号分解到8个频段上，计算出各个频段上的能量值，并将8个频段上的能量作为输入样本的特征向量。如图5所示，为采用小波变换方法提取特征向量作为输入样本的训练和识别情况。

从图6中可以观察到经过小波变化提取特征向量后作为输入样本后滚动轴承的三种运行状态测试集的识别准确率仅为95.0%，并且从图6（b）中可以发现测试集的损失值随着迭代的进行一直没有下降，反而逐渐上升，向过拟合的趋势发展。说明在提取特定的频域特征向量的同时会丢掉一定量的声发射信号特征，致使采用强大的CNN网络模型时也没法学习到全面且表征性强的故障声发射信号特征，并且因为需要利用小波变换提取特征，花费的时间也比较多。因此也验证了在本文学习分类任务中所提出方法相比提取特征向量作为CNN网络模型输入样本时的性能优势。

4.总结

本文提出了一种利用原始声发射信号与卷积神经网络相结合的滚动轴承内圈与外圈故障的智能化诊断方法。在故障诊断过程中避免了需要人工使用复杂信号处理技术进行特征提取或形式变换的步骤，使解决实际工程问题更加便捷。同时真正的利用了CNN强大的自适应特征提炼与浓缩表达，在对滚动轴承的不同故障声发射信号的自适应识别与诊断中，准确率达到97.2%，使该方法具有较高的准确性和有效性。对基于声发射检测技术的旋转机械设备健康状态在线监测与故障诊断研究具有较大的指导意义和实用价值。

（上接第25页）

文本分类查准率的对比效果数据如表3所示。文本分类查全率的对比效果数据如表4所示。文本分类F1测量值的对比效果数据如表5所示。

表3 文本分类查准率的对比效果数据

方法	褒义文本	中性文本	贬义文本
改进前TF-IDF	72.3%	77.5%	78.7%
改进后TF-IDF	87.6%	92.8%	86.5%

表4 文本分类查全率的对比效果数据

方法	褒义文本	中性文本	贬义文本
改进前TF-IDF	78.3%	75.4%	73.2%
改进后TF-IDF	89.8%	88.3%	87.8%

表5 文本分类F1测量值的对比效果数据

方法	褒义文本	贬义文本
改进前TF-IDF	78.9%	75.7%
改进后TF-IDF	88.6%	86.9%

从表中数据可知，经过改进TF-IDF算法较传统TF-IDF算法在文本分类查准率，查全率，以及F1测量值方面均有明显改善。具体的体现为：文本改进后的算法在褒义、中性、贬义效果提升查准率方面至少提高8%，在查全率方面至少提高11%，在F1度量值方面至少提高10%。

文本褒义，中性，贬义分类ROC_AUC曲线分别如图1，图2，图3所示：

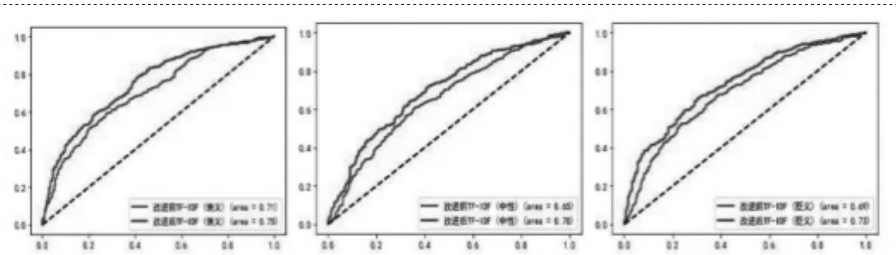


图1 褒义分类　图2 中性分类　图3 贬义分类

由图可知，改进后的AUC面积总是大于改进前的AUC面积，由此可见改进后的TF-IDF特征提取算法能更好地区分正负样本，提高分类模型的准确性。

5.结语

本文在传统的TF-IDF算法的基础之上提出一种新的卡方统计的特征词提取算法，从查准率、查全率、F1测量值和ROC_AUC四个方面进行了评估。新算法弥补了未考虑特征词汇在类间分布的缺陷，可以将在类间分布均匀但并不具备类别区分能力的词赋予较低的权重，有效地提高了权重计算的准确性和文本分类的正确性。