互信息特征选择法在《中图法》内容相似类目中的运用及改进
——以E271和E712.51为例

李湘东[1,2]，阮涛[1]
(1.武汉大学信息管理学院，武汉 430072；2.武汉大学电子商务研究与发展中心，武汉 430072)

摘要：针对内容相似的两个类目间存在大量共同特征而难以自动区分的特点，提出一种改进的互信息特征选择法，以提高两类文本自动分类的效果。以《中国图书馆分类法》中E271（中国陆军）和E712.51（美国陆军）两个类别的书目信息作为文本分类的对象，首先针对传统互信息特征选择法未考虑负相关特征、类间集中度和类内分散度等问题，引入改进的互信息特征选择法DNCF_MI；其次，针对DNCF_MI未区分不同特征对类别的贡献程度等不足，引入领域无关特征和领域相关特征，提出一种改进的互信息特征选择法DNCF_DI_MI；最后，使用knn分类器进行分类，并采用宏平均F1值和微平均F1值对分类结果进行评价。实验结果表明，本文提出方法的宏平均F1值和微平均F1值比传统互信息特征选择法分别提升24.1%和28.5%，比DNCF_MI均提升4.5%，证明本文方法对内容相似类目的分类更有效。

关键词：内容相似类目；中国图书馆分类法；两类分类；互信息；特征选择
中图分类号：G250.7
DOI：10.3772/j.issn.1673-2286.2018.01.008

随着现代互联网技术的蓬勃发展，人工智能开始普及并逐渐渗透到各个应用领域。基于机器学习的自动文本分类技术（也可简称为自动分类），作为人工智能的主要技术之一，是对文本信息资源进行组织与管理的重要手段[1]。在信息管理领域，人工智能技术的应用也逐渐成为可能和必要。《中国图书馆分类法》（以下简称《中图法》）是图书馆对信息资源进行分类组织的主要分类体系，该分类体系下存在大量内容相似的类目，本文将其称为内容相似类目。在信息管理领域开展自动分类研究，对内容相似类目的处理是其中一个重要课题。

从机器学习角度看，内容相似类目指两个及两个以上类别间的文本在用词上非常接近，以不分词序和语义特征的词袋模型所表示的文本间差异非常小的类目或类别。对其采取自动分类技术研究指将内容相似类目中的任意两个类别看作两个类的文本内容或特征非常接近时的两类文本分类问题。内容相似类目的分类问题对数据集和相关分类技术都有更高的要求。

本文以《中图法》中E271（中国陆军）和E712.51（美国陆军）两个类别的书目信息作为内容相似类目的分类对象，针对分类技术中的特征选择环节，引入领域无关特征和领域相关特征对DNCF_MI（DNC Frequency and Mutual Information）特征选择法进行优化，以提高对内容相似类目的分类精度，为今后对《中图法》中更多内容相似类目开展基于机器学习的自动分类研究抛砖引玉。

1 研究现状和意义

1.1 研究现状

在《中图法》等科学分类体系下开展的自动分类研究中，分类对象的文本主要是由题名和摘要构成的书目信息。一部分研究主要针对单层分类[2-6]，使用的类目为

不同类别区分明显的同层类目，目的是探索基于机器学习的自动分类是否适用于书目信息所组成的文本。单层分类明显不符合《中图法》等级层次明确的科学分类体系。目前已有学者围绕《中图法》或DDC等科学分类体系对书目信息进行层次分类[7-9]，但这些研究多通过缩减等级的层次深度、合并类目或缩短类号等方式对原科学分类体系进行改造和重构，在3—4级层次的深度上对书目信息开展自动分类时的有效性进行检验，其本质是将内容相似类目合并为一个类。这种做法依然不能满足《中图法》中11层的科学分类体系，同时，也从实证角度说明在《中图法》等科学分类体系下开展自动分类的难度，即随着层次和类别数量的增加，类别间差异较小或类别间极为相似而难以区分，现有研究不得不通过缩减等级的层次深度或合并类别等方式忽略内容相似类目的自动分类，将内容相似类目合并为一个较大的类别。黄莉[10]和薛春香[11]等指出《中图法》体系庞大，存在不同的主题划分，且各主题下层次纵横，存在大量上下位关系，同位类上也存在多个不同类别，同时类别间差异较小，难以区分。何琳等[12]认为《中图法》层级较多且类别间极为相似，容易造成错分、误分。如E271和E712.51属于主题相同而地区不同的内容相似类目，在对其区分时首先要确定上位类E（军事）下的文本，再进一步确定下一层（二级类目）应归入E2（中国）还是E712（美国），以此类推，直到将文本划入E271或E712.51。因此，如何确定内容相似类目中待分类文本的所属类别是《中图法》分类体系下自动分类的一个难点。然而，已有研究只是分析并指出问题所在[10-12]，尚没有提出解决内容相似类目自动分类的具体办法。

在解决内容相似类目的自动分类研究中，目前两类分类是主要的解决途径。两类分类指对类别间主题相似或相同，又有区分必要的两类文本进行基于机器学习的自动分类。两类分类的主要问题在于商品评价信息分类、情感分类、敏感信息过滤等，其主要特点为分类对象间内容极为相似、难以区分。有研究探讨了微博评论的情感倾向分类问题[13-15]，指出微博评论的分类实际是一种将评论内容分为积极或消极的两类分类问题，区分的难点主要在于对那些相似性高，却从属于不同情感词的划分。研究指出网络不良信息过滤问题[16-17]，实际是对网络信息进行区分的两类分类问题，即对具有相似内容却表现出不同倾向的信息进行正确过滤。已有文献均指出开展内容相似类目的两类分类研究的难点是在内容极为相似的前提下，如何为内容相似类目赋予合适的特征[13-17]。如李亚南[13]采用最大匹配方法重新组合特征；杨欢[14]在文本分类时，将主题与情感相关联进行特征值组合；Pan等[18]提出领域相关特征和领域无关特征的思想及相应决定方法，并将其应用到情感的两类分类问题中。这些研究为本文的展开提供思路，可以借鉴到利用《中图法》开展两类分类的研究中，即通过优化特征选择过程，选取更具类别代表性的特征以提高分类效果。

特征选择是一种特征降维的方式，特征选择的主要任务是根据某种规则选择出更具类代表性的特征，而剔除与类联系性弱的特征。常用的特征选择法有x^2统计、信息增益、互信息、期望交叉熵、文档频率等[19]。互信息在实际运用中得到多数研究人员的青睐。已有研究指出互信息是一种基于信息熵的特征选择法，在文本分类中，互信息可以表示任意特征词与类别的共现关系，互信息值越大，则表示它们间的共现概率越大，相关性也越大，本文将其称为传统互信息[20-22]。邓彩凤[23]对传统互信息方法进行了深入研究分析，总结出传统互信息方法的特点及不足，并引入类内特征频度和类内分散度指标，对传统互信息进行改进，通过选取复旦语料库中艺术、计算机、经济等10个类目开展分类对比实验，结果表明改进的方法较传统互信息要好。辛竹等[24]针对传统互信息中的负相关现象以及偏向于选择低频词等特点，在综合考虑负相关特征、类间集中度和类内分散度等因素的基础上，提出一种改进的互信息特征选择法DNCF_MI，同样通过选取复旦语料库中艺术、计算机、经济等10个类目开展分类对比实验，结果表明DNCF_MI在分类性能上比传统互信息要好。邓彩凤[23]和辛竹等[24]提出的互信息方法都取得了一定的效果，但这些改进的互信息方法仍然存在不足：传统互信息[20-22]及改进后的互信息[23-24]都是以类别区分明显的三个类或以上为分类对象，没有考虑内容相似类目的处理方法。因此，可以结合情感分类和敏感信息过滤问题的处理思路，对互信息作进一步改进。但该思想及方法没有区分领域无关特征和领域相关特征对两类分类的贡献，由于在实际情况中，领域无关特征的贡献度应小于领域相关特征，因此将其应用到《中图法》内容相似类目的分类中时，需要进一步优化，以提高《中图法》相似类目下两类分类的精确度。

综上所述，针对《中图法》现有自动分类研究中忽略内容相似类目的不足，本文以《中图法》中内容相似类目为研究对象，将领域无关特征和领域相关特征的思想引入互信息特征选择法中，并进一步区分领域无关特征和领域相关特征对两类分类的贡献，对DNCF_MI

进行改进，提出改进的互信息特征选择法DNCF_DI_MI（DNC Frequency and Domain Independent Mutual Information），使其能较好地适应内容相似类目的特征选取。此外，《中图法》中3个或3个以上的类目，其分类也可以转化为两类分类，如对E512.51（俄罗斯陆军）、E353.51（巴基斯坦陆军）和E712.51进行分类时，可以将E512.51/E353.51看作一个类，从而转换为E712.51与E512.51/E353.51的两类分类问题，进而再对E512.51与E353.51的两类分类分析。因此，对《中图法》中多个内容相似类目的划分问题，都可以看作对任意两个内容相似类目的文本进行两类分类的问题。

1.2 研究意义

目前以《中图法》为对象的自动分类研究主要包括综述性、分析性文章，与分类方法相关的研究主要针对界限清晰的类别。然而，《中图法》中存在大量内容相似、难以区分的类别。本文将相似类目的分类转换为两类分类问题，对两类分类中的特征选择法提出具体改进措施，为《中图法》中内容相似类目的自动分类提供解决方向和途径。

此外，特征选择是降低特征空间维度、减少空间复杂度的过程，是数据挖掘和机器学习中的基础技术，也是自动分类的主要环节。本文对传统互信息进行改进，提高其提取特征的能力，使其更加适应内容相似类目的分类需求，提高对《中图法》开展自动分类的信心和效果，进一步细化特征选择法的相关研究和应用能力。

2 研究方法

2.1 传统互信息特征选择法及其不足

互信息特征选择法的主要思想是判定特征词与类别的共现关系，特征词与类别的互信息越大，共现概率越大，即二者间的相关性越大，则更能代表该类的特征[25]。在进行特征选择时，传统互信息存在3点不足[24]。（1）忽视负相关特征对特征筛选环节的影响。根据互信息计算结果可知，部分特征的互信息值为负，称为负相关特征，而负相关特征在特征选择过程中通常会被剔除，但实际情况下负相关特征对分类结果也会产生影响。（2）没有考虑特征词的集中程度。部分重要特征词通常集中在一类中。（3）没有考虑特征词的分散程度。当两个特征词的互信息值相同时，根据互信息原理认为两个特征词同样重要，但在实际情况下，若一个特征词在某类少数文本中出现，而另一个特征词在该类大多数文本中出现，则前者更具有类别区分性。

2.2 DNCF_MI特征选择法及其不足

辛竹等[24]针对传统互信息的不足，引入DNC参数抵消传统互信息的负相关现象，并引入类内分散度、类间集中度分别表示特征词在类别内的分散程度以及在类别间的聚集程度，提出一种改进的互信息特征选择法DNCF_MI。

2.2.1 DNC、类间集中度和类内分散度

传统互信息方法存在负相关现象，在进行特征选择时，通常先剔除值为负的特征。但在实际情况下，负相关现象并不能被忽视。因此，引入DNC参数抵消负相关现象对传统互信息计算结果的影响。DNC的计算公式[24]为：

$$DNC = \frac{f_i(t) - \overline{f(t)}}{\overline{f(t)}} \quad (1)$$

其中，$f_i(t)$表示类别c_i中包含特征t的文档数，$\overline{f(t)}$表示平均每个类别含有特征t的文档数。

类间集中度指特征词在不同类别间分布的集中程度。类间集中度用特征词t对于类别c的后验概率表示。

类内分散度指特征词在某一类别内部不同文档间的分散程度。类内分散度用特征词t对于类别c的先验概率表示。

通过分析这些参数对传统互信息的影响，可以得出对于某一特征词来说，该特征为正相关特征且集中度越强、分散度越大，则更具有类别区分能力。DNCF_MI的计算公式[24]为：

$$DNCF\_MI(t) = DNC \times C \times D \times MI(t) \quad (2)$$

2.2.2 DNCF_MI的不足

DNCF_MI克服传统互信息的不足，结合类间集中度和类内分散度，并抵消负相关现象对特征选择的影响。DNCF_MI特征选择法能够选择出更具类别代表性的特征词，但对于内容相似类目的两类分类问题，无论

是传统互信息，还是DNCF_MI，都没有区分领域无关特征和领域相关特征对其的贡献，也没有将在两类中同时出现且具有不同贡献程度的特征与只在其中一个类中出现的特征加以区别。

2.3 改进的DNCF_DI_MI特征选择法

类别是反映领域概念的相关文本集合。因此，本文引入领域无关特征概念表示对两类贡献度低的特征，引入领域相关特征表示对两类贡献度高的特征[18]。

2.3.1 领域无关特征和领域相关特征

领域无关特征指在两类中同时出现且出现频次高于阈值δ的特征，反之，则为领域相关特征。对于特征t与两类间的贡献度，用κ来表示，定义为特征t对类别c_1的条件概率与特征t对类别c_2的条件概率比值，当κ越接近1，表明特征t在类别c_1和类别c_2中的条件概率越接近，特征t在类别c_1和类别c_2中的联系越紧密，特征t对c_1和c_2类的区分贡献度就越小，即对分类的作用越小。因此，可以使用特征t与两类间的贡献度κ值的大小来决定领域无关特征和领域相关特征。

2.3.2 DNCF_DI_MI特征选择法

本文针对上述问题，引入领域无关特征和领域相关特征概念，并调整领域无关特征和领域相关特征的贡献权重，即减少领域无关特征的权重ω，增大领域相关特征的权重ω，以此反映特征词对两类分类的不同重要性。在此基础上，提出一种优化的特征选择法DNCF_DI_MI，其计算公式如下：

$$DNCF\_DI\_MI\ (t) = DNC \times C \times D \times MI\ (t) \times \omega \quad (3)$$

其中，ω为权重，初始值为1，当t为领域无关特征时，则ω可用ω-α表示；当t为领域相关特征，则ω可用ω+α表示，其中α为调节参数，用来更新权重ω。

3 基于改进互信息特征选择法的分类框架

为解决内容相似类目的分类问题，提高分类效果，本文对DNCF_MI进行改进，提出一种优化的特征选择法DNCF_DI_MI，该方法引入领域无关特征和领域相关特征概念，并给出明确的定义和计算方式，在特征选择过程中赋予领域无关特征和领域相关特征不同权重，以选择出更具类别代表性的特征[24]。具体分类框架如图1所示。

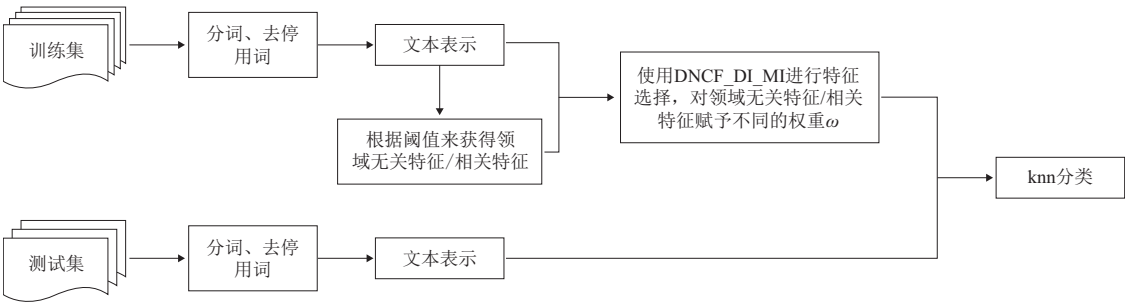

图 1 基于改进的DNCF_DI_MI特征选择法的分类框架

DNCF_DI_MI的过程主要分为五步。第一步，分词、去停用词。分别对训练集和测试集进行分词、去停用词等处理，获得每篇文章的初始特征集合。第二步，文本表示。使用向量空间模型对初始特征集合进行文本表示，将特征词表示为空间向量。第三步，根据上述内容，选取阈值参数δ获得领域无关特征和领域相关特征。第四步，结合领域无关特征和领域相关特征集合，使用本文提出的DNCF_DI_MI特征选择法，对领域无关特征赋予较低权重ω，而对领域相关特征赋予较高权重ω，以此来进行特征选择，选择出更具类别区分能力的特征集合。第五步，采用knn分类器（k-Nearest Neighbor，knn），计算待分类文本与经过特征选择后的特征集合相似度，将相似度最高的类别分配给待分类文本后，再计算宏平均F1值和微平均F1值，评估分类效果。

4 实验设计与分析

4.1 实验材料

本文的对象是内容相似的两类分类，因此，对语料集要求较高。为保证实验过程和结果满足公开原则，从维普数据库中提取《中图法》分类号E（军事）下的E271

与E712.51两个内容极为相近的语料作为实验的数据来源。其中，E271的文档共616篇，E712.51文档共1 366篇。每篇文档包括题名、关键词和文摘三部分信息，且两类文本集不存在交叉现象。为避免实验随机性对结果造成影响，实验共分为5组，每组在E271和E712.51中随机抽取200篇文档作为训练集，再随机抽取100篇文档作为测试集（每次共600篇文档且每次抽取的测试集与训练集不重复）进行实验，记录每组实验数据，对5组实验数据取平均值作为最终实验结果。为避免不平衡数据对实验结果的影响，本文所使用的测试集和训练集均采用平衡数据，即训练集和测试集中的文本数量一致。

4.2 实验方法与测评方法

实验目的是验证本文所提出的DNCF_DI_MI与DNCF_MI、传统互信息方法对分类性能的影响。由于分类效果易受特征数目、分类器参数的影响，根据预备实验的结果，将特征数目设为1 500；由于相似类目的特征项极为相似，而knn分类器是通过计算样本间的相似度来实现分类过程，因此knn分类器对内容相似类目的特征更加敏感，受噪声和非相关性特征向量的影响较小，且knn分类器具有实现简单、时间复杂度低、准确率高的特点，因此本文采用knn分类器[26]。knn分类器的性能与k值的选取以及相似度计算有关。故实验采取固定k值的方式以消除k值对分类结果的影响，根据预备实验的结果选取k=10，相似度计算采取杰卡德相似系数[27]。领域无关特征和领域相关特征的选取由阈值δ决定，即当特征t在两类中同时出现且出现频次高于δ时，同时结合贡献度κ，确定为领域无关特征，否则为领域相关特征。随后在DNCF_DI_MI特征选择过程中对其赋予权重ω。

为验证本文提出的方法对内容相似类目的有效性，综合考虑查准率和查全率，使用F1值表示对分类效果的评价，F1值能够体现自动分类的整体分类效果。与传统F1值不同的是，本文在查准率、查全率和F1值3种评价指标的基础上，使用宏平均F1值和微平均F1值对分类效果进行进一步评价[28]。

4.3 实验结果与分析

实验分为3个部分。（1）预备实验。主要通过一系列参数设置来决定领域无关特征和领域相关特征的阈值δ和权重调节参数α的取值。（2）特征选择法的效果比较。即对比传统互信息、DNCF_MI和本文方法得出的20个特征词，分析本文的方法是否能选择出更具代表性的特征。（3）分类效果比较。对比传统互信息、DNCF_MI和本文方法在采取knn分类器时的分类效果。

4.3.1 预备实验

领域无关特征和领域相关特征是根据特征词在两类中同时出现频次决定的，因此可以通过调整阈值δ以分析其对分类结果的实际影响，设置δ值分别为3、5、7、10，宏平均F1值的变化情况如图2所示。

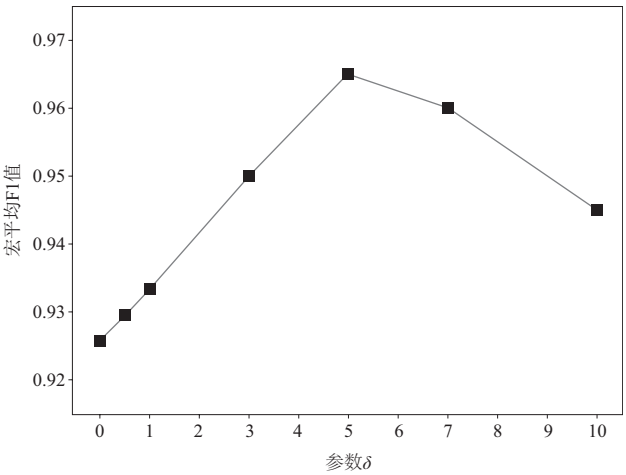

图 2 阈值δ对分类效果（宏平均F1）的影响

领域无关特征和领域相关特征确定后，需在DNCF_DI_MI特征选择过程中对其赋予权重ω，即对领域无关特征减小其权重，对领域相关特征增大其权重。由于不同的α值对最终的分类效果会产生影响，因此实验通过设置不同值（0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9），观察宏平均F1值的变化情况，实验结果如图3所示。

由图2和图3可知，当δ=5，α=0.7时，分类效果均最佳，故根据预备实验结果，选取阈值δ为5，权重ω的调节参数α为0.7进行后续实验验证。

4.3.2 特征选择法的效果比较

对训练集和测试集进行分词、去停用词和文本表示后，使用传统互信息、DNCF_MI和DNCF_DI_MI特征选择法进行特征选择，对每一个特征词计算互信息值，按照从大到小排序后，得出前20个特征词。

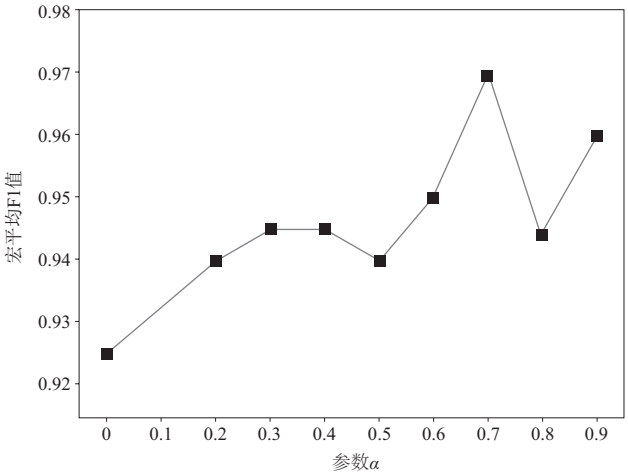

图 3 权重调节参数α对分类效果（宏平均F1）的影响

根据结果显示，传统互信息特征选择后的特征词明显不具有类别区分能力，不能够很好地代表E271和E712.51两个内容相似类目信息；而经过DNCF_MI特征选择法选取出来的特征词，诸如“美军”“美国陆军”和“中国陆军”等词，具有强类别区分能力，但是仍然存在一些诸如“美”“21世纪”等弱类别区分能力的特征词存在，可见DNCF_MI特征选择法虽然具有类别区分能力，能够选取出大部分有用特征，但仍有改进和优化空间；本文提出的DNCF_DI_MI特征选择法不仅选择出更具类别区分能力的特征，而且还将诸如“美国陆军”“中国陆军”等特征词排在前位，这符合E271和E712.51的实际情况。

经过分析得知，相较于传统互信息和DNCF_MI特征选择法，本文提出的DNCF_DI_MI特征选择法在对特征的选取上更符合实际。

4.3.3 分类效果比较

根据不同特征选择法选取的前20个特征词可以发现，使用DNCF_DI_MI特征选择法能够选择出更具类别区分度的特征集合。本文进一步采用knn分类器验证自动分类效果，实验结果如图4所示。

由图4可知，本文提出的DNCF_DI_MI特征选择法结合knn分类器的分类效果在宏平均F1值和微平均F1值上比传统互信息、DNCF_MI均有所提高。DNCF_DI_MI特征选择法的分类效果在宏平均F1值和微平均F1值上比传统互信息特征选择法分别提高24.1%和28.5%，比DNCF_MI均提高4.5%。由此可见，本文提出的方法在knn分类器下可以有效地提升分类效果。

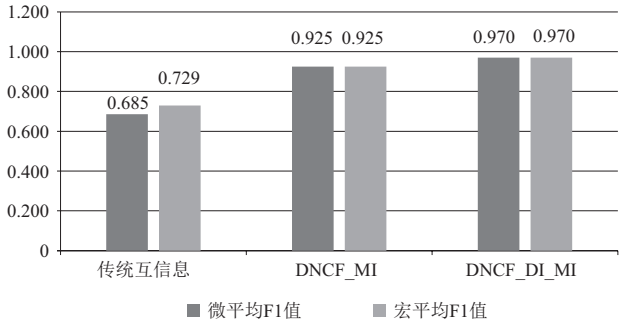

图4 knn分类器下传统互信息、DNCF_MI和DNCF_DI_MI的宏平均F1值和微平均F1值比较

5 结语

本文旨在针对《中图法》下内容相似类目难以实现自动分类的实际情况，引入两类分类的思想为内容相似类目提供解决思路和途径，并通过改进互信息特征选择算法以提高内容相似类目的分类效果。本文探讨了传统互信息和DNCF_MI方法及其不足，在此基础上，引入领域无关特征和领域相关特征，并在特征选择过程中进行权重调整以适应分类情况。最后本文使用《中图法》中E271和E712.51两个内容相似的类目信息作为实验语料，通过实验论证了本文提出的DNCF_DI_MI方法与传统互信息和DNCF_MI特征选择法相比，不仅能够提取出更具有类别区分度的特征词，而且能进一步改善自动文本分类的性能，提高分类效果。本文仅以《中图法》中主题相近但地区不同的两个内容相似类目作为有效性的检验对象，未来需要从更多角度（如对主题相近但时代不同等）对《中图法》中其他类型的内容相似类别开展相关研究。

参考文献

[1] 李湘东,巴志超,高凡.数字文本自动分类中特征语义关联及加权策略研究综述与展望[J].现代图书情报技术,2016,32(9):17-26.

[2] 李森,马军,赵嫣,等.对数字化科技论文的自动分类研究[C]//全国搜索引擎和网上信息挖掘学术研讨会,2006.

[3] 赵纪元,罗霄.面向中图法的学术文献自动分类研究[C]//中国计算机语言学研究前沿进展,2009.

[4] 王东波,苏新宁,朱丹浩,等.基于支持向量机的医学期刊文章自动分类研究[C]//全国计算机信息管理学术研讨会,2010.

[5] 薛春香,夏祖奇,侯汉清.基于语料和基于标引经验的自动分类模式比较[J].南京农业大学学报(社会科学版),2005,5(4):37-43.

[6] 刘红梅.图书馆多种类型文献自动分类研究[D].武汉:武汉大学,2012.

[7] 王军.数字图书馆的知识组织系统:从理论到实践[M].北京:北京大学出版社,2008.

[8] PONG Y H,KWOK C W,LAU Y K,et al.A comparative study of two automatic document classification methods in a library setting[J]. Journal of Information Science,2008,34(2):213-230.

[9] 王昊,严明,苏新宁.基于机器学习的中文书目自动分类研究[J].中国图书馆学报,2010,36(6):28-39.

[10] 黄莉,李湘东.基于《中图法》的自动分类研究现状与展望[J].图书情报知识,2012(4):30-36.

[11] 薛春香,何琳,侯汉清.基于《中图法》知识库的自动分类相关问题探析[J].图书馆建设,2015(6):16-20.

[12] 何琳,刘竟,侯汉清.基于《中图法》的多层自动分类影响因素分析[J].中国图书馆学报,2009,35(6):49-55.

[13] 李亚南.微博评论情感倾向性分类研究[D].天津:天津科技大学,2015.

[14] 杨欢.基于文本分类的微博情感倾向研究[D].重庆:重庆师范大学,2016.

[15] LI J,FONG S,ZHUANG Y,et al.Hierarchical classification in text mining for sentiment analysis of online news[J].Soft Computing,2016,20(9):1-10.

[16] 彭昱忠,元昌安,王艳,等.基于内容理解的不良信息过滤技术研究[J].计算机应用研究,2009,26(2):39-44,53.

[17] XIE X L,LONG Z.Implementation of bad information filtering system based on SVM algorithm[J].International Journal of Security and Its Applications,2016,10(9):45-54.

[18] PAN S J,NI X,SUN J T,et al.Cross-domain sentiment classification via spectral feature alignment[C]//International Conference on World Wide Web,World Wide Web 2010,Raleigh,North Carolina:DBLP,2010:751-760.

[19] DASGUPTA A,DRINEAS P,HARB B,et al.Feature selection methods for text classification[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.ACM,2007:230-239.

[20] 范小丽,刘晓霞.文本分类中互信息特征选择方法的研究[J].计算机工程与应用,2010,46(34):123-125.

[21] 刘佳.基于互信息特征选择算法的文本自动分类研究[D].淮南:安徽理工大学,2015.

[22] ESTÉVEZ P A,TESMER M,PEREZ C A,et al.Normalized mutual information feature selection[J].IEEE Transactions on Neural Networks,2009,20(2):189.

[23] 邓彩凤.中文文本分类中互信息特征选择方法研究[D].重庆:西南大学,2011.

[24] 辛竹,周亚建.文本分类中互信息特征选择方法的研究与算法改进[J].计算机应用,2013,33(s2):116-118.

[25] 孙建军.信息检索技术[M].北京:科学出版社,2004.

[26] 张宁,贾自艳,史忠植.使用KNN算法的文本分类[J].计算机工程,2005,31(8):171-172.

[27] 卢盛祺,管连,金敏,等.LDA模型在网络视频推荐中的应用[J].微型机与应用,2016,35(11):74-79.

[28] 奉国和.文本分类性能评价研究[J].情报杂志,2011(8):66-70.

作者简介

李湘东，男，1963年生，博士，副教授，研究生导师，研究方向：文本分类、机器学习、信息检索，E-mail：xli_xiao@hotmail.com。
阮涛，男，1992年生，硕士研究生，研究方向：文本分类、机器学习，E-mail：1159830216@qq.com。

The Application and Improvement of Mutual Information Feature Selection Method in the Similar Categories of Classification in CLC: Take E271 and E712.51 as an Example

LI XiangDong[1,2], RUAN Tao[1]
(1. School of Information Management, Wuhan University, Wuhan 430072, China;
2. Center for Electronic Commerce Research and Development at Wuhan University, Wuhan 430072, China)

Abstract: An improved mutual information feature selection method is proposed to improve the effect of automatic classification of two kinds of text, which is characterized by the existence of a large number of common features in text, which is difficult to distinguish automatically. The E271 (Chinese army) and E712.51 (American army) bibliographic information in CLC are used as the object of two types of text classification. Firstly, the traditional mutual information feature selection method, which does not consider the negative correlation feature, however the DNCF_MI feature selection method has overcome the weakness. Secondly, the DNCF_MI does not consider the difference between the two types of features in two categories, because the features that will appear simultaneously in two categories, have different degrees of contribution to characteristics that appear only in one of the classes. So, this paper introduces the field-independent features, domain-related features and proposes an improved DNCF_DI_MI feature selection method. Finally, the knn classifier is used for classification, and the Marco-F1 value and the Mirco-F1 value are used to evaluate the classification results. The experimental results show that the Marco-F1 and Mirco-F1 values of the proposed method are 24.1% and 28.5% higher than that of the traditional mutual information respectively, and 4.5% higher than that of DNCF_MI, which proves that the method is valid.

Keywords: Similar Content Category; Chinese Library Classification; Two Categories of Classification; Mutual Information; Feature Selection

（收稿日期：2017-09-27）