文章编号:2096-1162(2019)02-0026-07 收稿日期:2018-12-25

中文科技期刊论文多标签分类研究

马 芳,黄翠玉

(烟台工程职业技术学院图书馆,山东烟台,264006)

摘 要:由于传统的人工分类不够规范、准确,而且随着期刊数字化程度的不断提高,采用文本自动分类技术很大程度上提高了分类的准确率并缓解了人工分类的压力。利用《中国图书馆分类法》建立科技期刊论文类别体系,采用组合多标签特征选择算法(CMLFS)对多标签数据进行特征选择,并采用先进的多标签随机游走算法(MLRW)对科技期刊论文样本集进行训练和测试。结果表明,对中文科技期刊论文进行多标签自动分类,能够简化科技期刊论文多标签分类的过程,提高分类效率,分类效果理想。

关键词:中文科技期刊;论文;多标签分类;特征选择算法;随机游走算法

中图分类号:G252.2 **文献标识码**:A

为了满足科技发展对情报的需求,科技期刊论文作为一种重要的情报源,已经形成了一套统一的著录标准(元数据)。而分类号作为元数据中的一员,其分类标引工作是情报加工过程中一项重要、复杂的工作。长期以来,这项工作都是由论文作者或期刊编辑手工完成的,而人工分类不可避免地存在一定的主观性,因此,科技期刊论文的分类往往不够规范、准确。为此,有必要采用新的分类技术——文本自动分类来改进手工分类的不足。

对于科技期刊论文进行自动分类比较统一的方法是:首先,通过人工标引和统计学方法构建分类库,分类库中每个类别都用一个特征词向量来表示,然后利用分类算法来判定样本数据和各个类别的特征词向量的相似度,相似度最高的类别就是该样本的类别。该方法在一定程度上减轻了传统人工分类的压力,但是仍然存在人工标引的主观性,并且事先构建的分类库需要随着知识的更新不断重新构建。为了更好地揭示论文中所包含的不同主体及其之间的相互关系,满足读者从科学分类对论文进行族性检索的需求,对论文进行准确的多标签分类也是非常必要的。为此,本文

The Exploration on Subject Service in Art Higher Vocational College Libraries: Taking Shanxi Academy of Arts as an Example

WANG Ruijun

ABSTRACT: This paper analyses the situation of subject service in the higher vocational college libraries, expounds the necessity and significance of developing subject service in art higher vocational colleges, and according to the features of subject setting in art higher vocational colleges, explores the pattern of subject service which is suitable for the library of Shanxi Academy of Arts, in order to provide reference for the art higher vocational college libraries to carry out subject service.

KEY WORDS: higher vocational college library; subject service; art

引进多标签自动分类技术，采用机器学习的理念对论文样本进行学习，构建分类库，不仅能够避免人工标引的不足，而且自动分类的准确率和效率都有了显著提高。

1　研究现状

文本自动分类技术起源于国外，经过20多年的不断发展，分类模型和分类算法逐渐完善，并广泛应用于信息检索与文本挖掘等领域。文本分类可以分为单标签分类和多标签分类两种，在实际应用中，多标签数据是普遍存在的，近年来逐渐得到人们的广泛关注。针对多标签分类问题，许多学者提出了可行的模型算法，如文献[1]提出了一种基于随机游走模型的多标签分类算法，其将多标签数据映射成为随机游走图，通过游走图中每个顶点得到的概率分布来刻画未分类数据具有每个标签的概率，该算法能够有效解决多标签分类和排序问题；文献[2]将粗糙集理论引入多标签文本分类，利用训练阶段得到的各个类别的分类规则与测试实例逐一匹配，得出实例的类标签集合，扩展了粗糙集理论在文本分类中的应用。

随着文本分类技术的不断成熟，逐渐有学者将文本分类技术引入论文分类标引中。如文献[3]提出在机器学习的计算模式下，对不同著录项进行加权构造论文特征向量，并且针对《中国图书馆分类法》(以下简称《中图法》)的特点，采用浅层次分类法构建层次分类器，来有效实现期刊论文的《中图法》分类；文献[4]采用基于支持向量机学习模型，采取基于低密度多特征的训练方法，对医学期刊R7中的9个小类进行了自动分类研究，取得了相对满意的分类结果。这些期刊论文中的自动分类方法能够有效地解决传统人工分类中存在的问题，但是实现起来有一定难度，并且以上研究都是针对期刊论文的单标签分类标引。

目前，对科技期刊论文的自动分类主要还停留在单标签分类上，主要是考虑到一篇论文同属于多个类别的多标签分类的研究较少。通过检索中国知网，仅找到一篇与之相关的研究论文，即文献[5]提出的基于本体与结构权重的中文科技论文多标签分类。该文献针对中文科技论文特殊的结构特点，提出结构权重的概念，对处于论文中不同结构部分的特征词进行加权处理，并结合领域本体技术进行特征选择，在一定程度上提高了多标签分类效果。但是随着社会的发展、科技的进步，领域本体中的概念、属性及实例也在不断更新、完善，本体的构建将是一个长期而复杂的过程，而该文献仅运用比较简便的RAKEL随机标签组合算法，没有引入其他多标签分类算法且缺少多种分类算法之间的比较分析。

针对科技期刊论文样本集中特征集合维数过高、领域本体自学习能力较差、分类性能较低等问题，本文引入SUMO本体技术，采用先进的多标签特征选择及分类算法，建立科技期刊论文多标签分类模型。该模型利用《中图法》建立科技期刊论文类别体系，针对每篇论文的分类提取与论文类别相关的信息，如题名、摘要、关键词，通过分词、特征选择、TF-IDF权重构建向量空间，然后采用多标签分类算法进行训练，构建出性能最佳的分类器。

2　科技期刊论文多标签分类模型设计

2.1　科技期刊论文类别体系

目前，我国主要采用《中图法》对科技期刊论文进行分类和标注，《中图法》是针对图书资料的分类与检索而编制的专业分类法，其标引规则的制定主要是为了达到图书排架的稳定性。由于科技期刊论文通常包含多主体要素，为了能对文章的每个主题因素都予以充分揭示，在遵循《中图法》标引规则的同时，还需针对科技期刊论文分类标引的特点，对文章标注多个分类号，这样不仅为论文增加了检索入口，也大大提高了分类的准确性。如《抑郁症的生化病理机制探讨》一文，可以标注为R749.4(情感性精神病)和R362(病理化学)2个分类号。

《中图法》共有22个大类，标引深度一般为6级，采用辅助手段可达9级，为了保证科技期刊论文检索的准确性，要求论文标引深度要适中，以4～6级较为适宜。由于文本分类技术一般都采用单层次分类法，即把所有的类目都放到同一个层面上，不考虑类目之间的相互关系，而《中图法》采用树形结构，具有一定的广度和深度，文本分类模型不适合《中图法》这种具有复杂类目和不均匀深度的类目体系。为此，本文引入SUMO本体，借助其丰富的概念语义关系和清晰的层次结构，利

用斯坦福大学开发的本体编辑工具Protégé，将《中图法》转换成适合文本分类的科技期刊论文类别体系。由于标引深度的级别越高，类别之间的区分度就越低，为了保证分类的准确性，本文分类体系结构深度为4级。

2.2　科技期刊论文多标签分类模型构建

对科技期刊论文进行人工分类标引需要根据论文的学科内容、主题多寡、作者意旨等，按照一定的分类体系，科学、系统地表达论文的主题性质。若要准确地分类标引需经过以下步骤：首先，通过对论文标题、摘要、关键词、文内各标题以及全文的浏览，判明论文的学科主题特征。然后，根据论文中所涉及的不同主题，从《中图法》的类目表中分别寻找其所属的类别。论文的自动分类标引过程与人工分类标引相似，是将相关的论文内容转换成计算机可以识别的数据，由计算机进行“阅读学习”，构建相应的分类库。

一篇论文可能同时属于多个类别，即有多个标签，这种多标签分类有别于传统的单标签分类问题，在单标签分类中，标签之间互不相关，但是在多标签分类中，由于标签之间存在很大的关联性和共现性，使得单标签分类中的特征选择算法、分类算法等不能直接应用于多标签分类中。鉴于此，多标签分类问题是根据多标签数据和多标签分类的特点，对训练样本（已预知类别的样本）进行训练，寻找标签与内容之间以及标签之间的潜在关系，构建相应的分类模型，然后通过分类模型预测测试样本（未标记待分类样本）所属标签集合，并对分类结果进行评估。具体分类过程如图1所示。

（1）数据预处理。科技期刊论文包含题名、作者、机构、摘要、关键词、分类号、正文等信息，对于一篇论文，标题、摘要和关键词展示了它的主要信息，通过这些信息可以大致分析出文章的重点，故本文选取题名、摘要、关键词作为分类特征来源，即作为训练、测试样本集。采用中文分词工具ICTCLAS对样本集中的文档进行分词、去停用词，并从分词结果中剔除对分类没有贡献的词（高频词、稀有词），完成数据预处理，实现对特征词的粗降维，以减少分类噪音。

（2）特征选择算法的选取。通过数据预处理后产生的初始特征集维数非常高，特征集维数过高不仅会使分类算法计算量过大，而且会造成分类结果不够准确。因此，需要从初始特征集中删除稀疏特征，保留更有利于分类的特征。特征选择是进行特征降维最常用的方法，单标签分类中常用的特征选择算法有信息增益IG、互信息MI、Relief、ReliefF、F统计量法等，它们大都是针对单标签数据，很难直接应用于多标签数据。为此，有学者根据多标签数据的特征，对单标签特征选择算法进行改进，使它们适用于对多标签数据进行特征选择。如：由ReliefF算法改进的多标签ReliefF算法（ML-ReliefF），由F统计量算法改进的多标签F统计量算法（ML-F）等。

No Free Lunch理论表明：不同的算法通常各有其优劣，没有哪种算法绝对优于另一种算法[6]。所以，如果能把几种特征选择算法的优缺点结合起来，将会提高特征选择的稳定性和合理性。基于该思想，本文将ML-ReliefF算法和ML-F进行组合，提出组合多标签特征选择算法（CMLFS）。该算法首先为以上两种选择算法选取两个与原始训练集相同的训练样本集（SR，SF），然后采用投票的方法确定每个特征选择算法的权重，

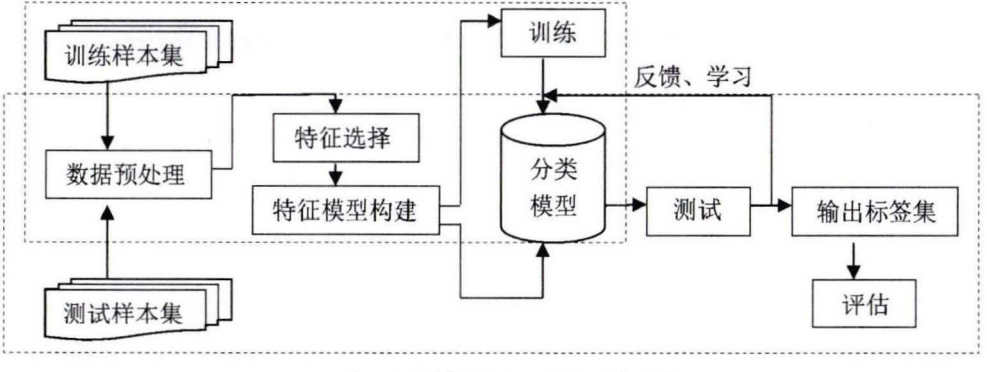

图1　科技期刊论文多标签分类模型

$W_{(R,F)}=1/L\sum_{i=1}^{L}W_i$,二者进行特征选择后,再对两种特征选择算法的结果进行投票,$W=W(R)+W(F)$,并依据其值大小,对特征权重进行排序。最后,根据需要从中选取合适的特征。

(3)特征模型构建。通过特征选择进行特征降维后得到的特征集是特征词的集合,计算机无法直接对特征词进行计算分析,需要将其表示成计算机能够识别的数据格式,即进行特征模型的构建。目前,表示文本特征模型的方法主要有向量空间模型(VSM)、布尔逻辑模型、概率推理模型等。本文采用向量空间模型(VSM)来表示各类别科技论文的特征。

在进行向量空间模型构建时,需要计算每个特征词对分类所起作用的大小即权重,通过权重将科技论文文档表示为向量形式。计算权重的方法主要有布尔权重、TF、IDF、TF-IDF等。其中,TF-IDF权重公式既考虑了高频词的作用也考虑了特征词的类别集中度因素,具有很好的效果。因此,本文采用TF-IDF方法来表示科技期刊论文中的特征词在向量空间中的权重。

(4)分类模型。分类模型主要是利用多标签分类学习算法对科技期刊论文进行自动分类模型的训练,多标签分类的目的就是从多标签训练集D中学习一个函数$h:x\rightarrow 2^y$来预测未知样本的类标签集。目前,多标签分类问题主要有两大类解决方法,即基于问题转化的方法(PT)和基于算法转化的方法(AA)[7],本文采用一种基于随机游走模型的多标签分类算法,称为多标签随机游走算法(MLRW)。

MLRW算法[8]是将训练集D中的每个训练数据$x\in X$映射为图中的一个点,如果两个训练数据x_i、x_j具有相同的类标签,则将这两个训练数据对应的顶点v_i、v_j相连,由此,将具有相同标签的训练数据所对应的顶点相连可得到随机游走图系列G。在遍历G中的每个图G_t的过程中,每次游走可得到图中某个顶点被访问到的概率分布向量s。用此向量作为下一次游走的输入,并反复迭代此过程,当满足一定条件时,这个概率分布会趋于收敛,收敛后得到训练数据x具有每个标签的稳定概率分布向量π。然后,将π与设定的阈值向量进行比较,进而确定每个标签的取舍。相关公式如下:

$$s=(1-\alpha)\times P^T\times s_0+\alpha\times d,0<\alpha<1 \quad (1)$$

$$\pi=(1-\alpha)\times P^T\times \pi+\alpha\times d,0<\alpha<1 \quad (2)$$

其中,α为发生跳转时跳转到图中每个顶点的概率分布向量,P为随机游走图G上的权重矩阵W的邻接矩阵,边的权值即为训练数据对应顶点在d维空间中的距离,本文采用欧式距离作为距离函数。

(5)分类评估。分类模型完成训练和测试之后,需要选择合适的评价指标评估分类算法的优劣。由于多标签分类的特殊性,其评价方法不同于单标签分类,本文选取常用的多标签性能评估标准:汉明损失(Hamming Loss)、One-Error、排序损失(Ranking Loss)、平均精度(Average Precision),相关公式如下:

$$Hamming\ Loss=\frac{1}{P}\sum_{i=1}^{P}\left|h(x_i)\Delta Y_i\right|$$,其中,Δ是对称差算子 (3)

$$One\text{-}Error=\frac{1}{P}\sum_{i=1}^{P}\left[\left[\arg\max_{y\in Y} b_{y}(x_i,y)\in Y_i\right]\right] \quad (4)$$

$$Ranking\ Loss=\frac{1}{P}\sum\frac{1}{\left|Y_i\right|\left|\overline{Y_i}\right|}\left|\left\{(y_1,y_2)\,|b(x_i,y_1)\leqslant b(x_i,y_2),(y_1,y_2)\in Y_i\times\overline{Y_i}\right\}\right| \quad (5)$$

$$Average\ Precision=\frac{1}{P}\sum_{i=1}^{P}\frac{1}{\left|Y_i\right|}\sum_{y_1\in Y_i}\frac{\left|\left\{y_2\in Y_i\,|\,Rank_b(x_i,y_2)\leqslant Rank_b(x_i,y_1)\right\}\right|}{Rank_b(x_i,y_1)} \quad (6)$$

其中,Hamming Loss评价单个标签的分类误差,即实例标签对错误分类的次数;One-Error评价排序靠前的标签不应该在实际分类中的次数;Ranking Loss评价所有标签的排序出错程度;Average Percision评价预测出的标签准确精度。前3个评价标准的评估值越小越好,而最后一个评价标准的评估值越大表现越好。

3 科技期刊论文多标签分类实验及结果分析

3.1 实验设置

(1)实验数据。本文从中国知网下载所需样本集,样本集主要选自期刊论文的G、R、T 3个大类,从中筛选出可进行多标签分类标引的文章,每个大类分别选取1 000、2 000、3 000条数据。对于训练集和测试集的划分,比较权威的建议是训练集为70%,测试集为30%[9]。为了客观地验证实验效果,本文引入十倍交叉验证的方法来进行实验。

(2)实验方法。实验中参与比较的算法有特征选

择算法CMLFS、ML-ReliefF、ML-F,分类算法MLRW、MLKNN、BPMLL。其中,ML-ReliefF算法[10]是将标签对特征区分性能的影响即贡献值,加入到ReliefF算法的特征权值更新公式中,改进特征权值更新公式。ML-F算法是将贡献值加入到F统计量法的均值和方差公式中,改进F值的计算公式。MLKNN算法[11]是采用K近邻技术处理多标签分类数据,BPMLL算法[12]是利用BP神经网络技术解决多标签分类问题,这两种算法都具有较好的分类效果,常用于算法实验对比。为了保证每种算法都能表现优良,本文参考原著设置它们的算法参数,MLRW算法中α的取值为0.15[13],MLKNN算法中K的取值为10,BPMLL算法中隐含神经元的个数设为特征总数的20%,最大训练步数设为100。

(3)实验环境。本文采用基于Weka平台开发的多标签学习Java库Mulan[14],Mulan包含了多种多标签分类算法及评价框架,是开源的,已发布于GNU GpL licence。

3.2 实验结果分析

3.2.1 特征选择算法比较

实验采用多标签随机游走算法(MLRW)学习和训练分类器,对CMLFS、ML-ReliefF、ML-F 3种特征选择算法的降维效果进行比较。实验样本集选用G大类的1 000条数据,为了验证算法的稳定性,分别选取前20%特征和前80%特征进行实验对比,结果如表1所示。从表1可以看出,采用CMLFS算法进行特征选择所得到的Hamming Loss、One-Error、Ranking Loss 3个评估值都小于ML-ReliefF、ML-F算法的上述3个评估值,并且CMLFS的Average Precision值明显高于ML-ReliefF、ML-F算法的此项值。由此看来,通过对不同特征选择算法进行组合,可以有效利用其他算法的优点,消除某一算法的缺点,产生更佳的特征选择效果,分类的准确性更高。

通过不同特征的选取,CMLFS算法的4个评估值变化幅度较小,ML-ReliefF、ML-F算法的4个评估值变化幅度大些,这说明CMLFS算法的性能更稳定。

ML-ReliefF、ML-F算法都是通过加入贡献值对原算法进行改进,每种原算法的特征选择机理以及对特征重要性度量的方法都有很大的差异,并且贡献值的选取对特征选择效果有很大的影响。而组合投票后的CMLFS算法相当于从两个方面对特征进行综合评价,能够很好地融合原算法之间的差异,消减它们的缺陷,同时又结合了它们之间的优点,最终提高了算法的整体性能。

3.2.2 分类算法比较

为了验证MLRW分类算法对科技期刊论文分类标引的有效性,将MLRW分类算法与MLKNN算法、BPMLL算法进行实验对比,实验采用CMLFS特征选择算法。由于不同类别的样本集属于不同的领域,它们分别包含各自的特征种类和标签种类,在一定程度上可以验证分类算法在不同条件下的性能,所以,本实验样本集选用科技论文的G、R、T 3个大类的数据。

(1)不同分类算法分类效果如表2所示,从表2可以看出:第一,Hamming Loss值。在T大类上,MLRW算法值最低,3种算法相比较,MLKNN算法表现最好,这是由于MLRW算法的随机游走性带来的样本偏差不确

表1 不同特征选择算法分类效果

评估标准	选取前20%特征			选取前80%特征		
	CMLFS	ML-ReliefF	ML-F	CMLFS	ML-ReliefF	ML-F
Hamming Loss	0.165	0.175	0.181	0.158	0.186	0.178
One-Error	0.182	0.201	0.189	0.178	0.185	0.215
Ranking Loss	0.076	0.095	0.086	0.082	0.112	0.098
Average Precision	0.889	0.575	0.628	0.821	0.625	0.718

表2 不同分类算法分类效果

评估标准	MLRW			MLKNN			BPMLL		
	G类	R类	T类	G类	R类	T类	G类	R类	T类
Hamming Loss	0.178	0.129	0.085	0.183	0.125	0.148	0.156	0.112	0.121
One-Error	0.267	0.268	0.165	0.284	0.263	0.179	0.279	0.242	0.185
Ranking Loss	0.312	0.354	0.229	0.334	0.376	0.235	0.346	0.383	0.247
Average Precision	0.738	0.899	0.989	0.738	0.805	0.854	0.709	0.793	0.861

定,造成了单个标签分类误差的增大。第二,One-Error值。在R大类上,MLRW算法比MLKNN算法值略高,在G、T大类上,MLRW算法值最低,在G、R大类上,MLKNN算法比BPMLL算法表现优,但在T大类上,BPMLL算法比MLKNN算法表现又好些,尽管3种算法的性能排序有些变化,但总体上讲,MLRW算法还是优于其他两种算法。第三,Ranking Loss值。MLRW算法表现最优,MLKNN算法次之,BPMLL最后。第四,Average Precision值,在C大类,MLRW算法与MLKNN算法值相同,在R、T大类上,MLRW算法值最高,MLKNN算法与BPMLL算法差异不大,由此看来,MLRW算法为最优。

通过表2还可以看出,随着数据量的增大,Hamming Loss值 、One-Error值、Ranking Loss值总体呈减少趋势,Average Precision值随之增大,这说明每种算法的分类性能受样本量的影响,样本量越大,分类效果越好。并且,随着样本量的增加MLRW算法分类性能变化愈加明显,由此说明基于随机游走所建立的结构化模型,其结构化风险比经验风险要低得多,在此基础上,MLRW算法的分类性能更佳。

(2)不同分类算法分类时间如表3所示。从表3中可以看出,MLRW算法所用时间最短,但随着数据量的减少,3种算法的差距在不断减少,说明MLKNN算法与BPMLL算法更适用于中小规模的多标签分类任务。

表3 不同分类算法分类时间

分类算法	分类使用时间/s		
	G	R	T
MLRW	25	34	68
MLKNN	38	62	121
BPMLL	33	50	115

综上所述,无论是分类性能还是样本数量的影响以及训练时间,MLRW算法都要优于MLKNN、BPMLL算法,因此,采用MLRW算法对科技期刊论文进行多标签分类标引能够取得比较满意的分类效果。

4 结语

目前,科技期刊论文主要采用《中图法》进行手工分类,针对人工分类不够规范、不够准确,尤其是对科技期刊论文多标签分类所存在的问题,借助文本分类技术来实现科技期刊论文的多标签自动分类具有重要意义。

针对多标签数据的特点,将ML-ReliefF和ML-F算法进行组合,用组合多标签特征选择算法(CMLFS)对多标签数据进行特征选择,实现对科技期刊论文样本集的特征降维。根据多标签分类的特点,采用先进的多标签随机游走算法(MLRW)构建分类模型,该模型将多标签数据映射成多标签随机游走图,然后采用随机游走模型遍历每个图,得到每个顶点被访问到的概率分布,并将这个点概率分布转化成每个标签的概率分布。实验结果表明,上述多标签特征选择算法和分类算法具有一定的可靠性和稳定性,能够简化科技期刊论文多标签分类的过程,分类效果比较理想。但是,本文的研究仍然存在不足,如由于人工标引不够规范、不够准确,文中所选取的数据集中分类号的标注存在一定问题,对分类结果会产生一定的影响;在多标签分类方面,未充分考虑科技期刊论文样本集中标签的数量和分布对分类的影响,今后将针对这些问题进一步改进现有算法,以达到更好的分类效果。

参考文献

[1] 郑伟,王朝坤,刘璋,等.一种基于随机游走模型的多标签分类算法[J].计算机学报,2010(8):1419-1426.

[2] 吕小勇,石洪波.基于粗糙集的多标签文本分类算法[J].广西师范大学学报(自然科学版),2009(3):150-153.

[3] 王昊,叶鹏,邓三鸿.机器学习在中文期刊论文自动分类研究中的应用[J].现代图书情报技术,2014(3):80-87.

[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[J].情报理论与实践,2011(4):115-118.

[5] 杨琳.基于本体与结构权重的中文科技论文多标签分类研究[D].长春:东北师范大学,2012.

[6] WOLPERT D H, MACREADY W G. No free lunch theorems for optimization[J].IEEE Transactions on Evolutionary Computation,1997,1(1):67-82.

[7] TSOUMAKAS G.Multi-label classification[J].Inter-

national Journal of Data Warehousing & Mining, 2007, 3 (3):12-16.

[8] SCHAPIRE R E, SINGER Y. BoosTexter: A boosting-based system for text categorization [J]. Machine Learning, 2000, 39(3): 135-168.

[9] MITCHELL T M. 机器学习[M]. 北京: 机械工业出版社, 2003: 70.

[10] KONONENKO I. Estimating attributes: Analysis and extensions of RELIEF [C]//Proceedings of the 1994 European Conference on Machine Learning, LNCS 784. Berlin: Springer, 1994: 171-182.

[11] ZHANG Minling, ZHOU Zhihua. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognitioan, 2007, 40(7): 2038-2048.

[12] TSOUMAKAS G, VLABAVAS I. Random k-Labelsets: An Ensemble Method for Multilabel Classification [C]. Berlin Heidelberg: Springer, 2007.

[13] ZHANG L, WU J, ZHUANG Y, et al. Review oriented metadata enrichment: A case study[C]// Proceedings of JCDL. New York: ACM, 2009: 173-182.

[14] STREICH A P, BUHMANN J M. Classification of multi labeled data: A generative approach [C]//Machine Learning and Knowledge Discovery in Database: Europen Conference, September 15-19, 2008, Antwerp, Belgium. Berlin: Springer, 2008: 390-405.

(责任编辑:崔 静)

作者简介:马 芳,女,1975年生,烟台工程职业技术学院图书馆副研究馆员;黄翠玉,女,1963年生,烟台工程职业技术学院图书馆副研究馆员。

Research on Multi-label Classification of Papers in Chinese Sci-tech Periodicals

MA Fang, HUANG Cuiyu

ABSTRACT: Traditional manual classification is not standardized and accurate enough. With the continuous improvement of the digitalization of periodicals, automatic text classification technology can greatly improve the accuracy of classification and relieve the pressure of manual classification. This paper uses the *Chinese Library Classification* to establish the classification system of sci-tech periodicals papers, uses the combined multi-label feature selection algorithm (CMLFS) to select features from multi-label data, and uses the advanced multi-label random walk algorithm (MLRW) to train and test the sample set of sci-tech periodicals papers. The results show that multi-label automatic classification of Chinese sci-tech periodical papers can simplify the process of multi-label classification of sci-tech periodical papers, improve the classification efficiency and achieve satisfactory classification results.

KEY WORDS: Chinese sci-tech periodical; paper; multi-label classification; feature selection algorithm; random walk algorithm