文章编号:1671-4598(2018)05-0237-05 DOI:10.16526/j.cnki.11—4762/tp.2018.05.059 中图分类号:TP3 文献标识码:A

基于数据挖掘中文书目自动分类算法

纪 纲,王海东,陈小飞
(海军航空大学 舰面航空保障与场站管理系,山东 青岛 266041)

摘要:提出一种改进的数据挖掘算法;首先采用ICTCLAS系统进行文本预处理,以词频特征构建词条向量;然后融合词频特征和词频—逆向文件频率特征,构建训练样本集的特征矩阵;接着对该矩阵进行奇异值分解变换,得到语义空间,用于对文本特征向量进行语义空间变换,得到语义向量;最后构建联合支持向量机分类器,实现中文书目所对应的语义向量的自动分类;最后做了大量的仿真实验,实验结果表明,文章方法的分类准确率高于现有方法。

关键词:数据挖掘;中文书目分类;文本挖掘;支持向量机

Automatic Classification Algorithm of Chinese Bibliography Based on Data Mining

Ji Gang, Wang Haidong, Chen Xiaofei
(Department of Carrier Aviation Security and Station Management, Naval Aviation University, Qingdao 266041, China)

Abstract: an improved algorithm for data mining is proposed. The first use of ICTCLAS system for text preprocessing, construct the term vector in frequency characteristics; then the fusion frequency characteristics and frequency—inverse document frequency features, construct the characteristic matrix of the training sample set; then the matrix singular value decomposition, get the semantic space for semantic space transform of text feature vector, semantic vector; the construction of combined support vector machine classifier, automatic classification of semantic vector corresponding to the Chinese bibliography. At last, a lot of simulation experiments have been done, and the experimental results show that the classification accuracy of this method is higher than that of the existing methods.

Keywords: data mining; chinese bibliography classification; text mining; support vector machine

0 引言

文本数据[1]挖掘(Text Mining)是数据挖掘的主要分支之一,是从海量文本数据中抽取有价值的信息和知识的计算机处理技术,在图书分类检索、企业情报分析、搜索引擎等领域都有广泛应用[2]。文本数据挖掘方法主要包括文本分类、文本聚类、信息抽取、摘要和压缩等。其中,文本分类是文本数据挖掘的主要研究方向。文本分类依据文本之间的差异性特征实现不同类别文本的分类,一般包括文本预处理、统计和特征抽取、分类器设计等步骤。首先在预处理阶段将原始语料转换为结构化数据,然后统计词频等特征,采用诸如信息增益、互信息等特征提取方法提取文本描述特征,接着采用诸如支持向量机、神经网络等机器学习方法构建特征的分类器,实现特征的分类[3-9]。在现代图书馆管理领域,目前逐渐开始使用文本数据挖掘技术来实现图书的管理,如采用机器学习架构实现中文书目的自动分类。该技术主要包括文本预处理、特征提取和机器学习三个部分,目前已经有一些成熟的方法[10-14]。如文献[12]利用ICTCLAS分词系统对书名和摘要信息进行中文分词,为标题和摘要的特征词赋予不同的权重,采用词频—逆向文件频率提取特征,采用支持向量机进行特征分类。文献[13]同样采用ICTCLAS分词系统对书名和摘要信息进行中文分词,为每个书目构建书目+关键词的二元关联矩阵,分别采用支持向量机和BP神经网络进行特征分类。文献[14]采用概率主题模型表示书目信息,克服因文本短小而产生的特征稀疏问题;依据书目信息体例结构和类目区分能力等先验知识构建复合加权特征,结合概率主题模型实现中文书目信息分类。这些方法在中文书目自动分类领域都有有益的效果,然而分类准确率还有待进一步提高。

收稿日期:2018-02-23; **修回日期**:2018-03-27。

作者简介:纪 纲(1985-),男,江苏镇江人,讲师,硕士,主要从事计算机科学与技术(仿真),机械工程及其自动化,人工智能方向的研究。
王海东(1974-),男,山东安丘人,副教授,硕士,主要从事航母舰面航空保障方向的研究。

1 中文书目自动分类框架

在机器学习阶段,首先需要对中文图书的书目数据进行分析,抽取中文书目内容特征和中图法类目信息;然后对中文书目内容特征进行预处理,得到中文书目内容所包含的词条信息,将非结构化的文本信息转换为结构化的词条信息;接着依据词条信息提取能够描述不同类别中文书目内容的特征向量;最后,结合数据库中各个中文书目所对应的特征向量以及中图法类目信息组建训练数据集,选择合适的机器学习算法进行学习和训练,构建中文书目类目分类器。

在类目分析阶段,对于待分类的中文书目,首先抽取中文书目内容特征,然后进行预处理,得到词条信息;接着提取特征向量;最后将特征向量送进中文书目类目分类器,得到中文书目分类结果。

可见,基于机器学习的中文书目自动分类系统架构涉及的关键技术主要有三个部分:文本预处理、特征提取和机器学习,简要描述如下。

1.1 文本预处理

该部分主要任务是将非结构化的文本数据转换为结构化的

词条信息。对于中文书目分类而言，目前大多是采用中国科学院计算机研究所开发的 ICTCLAS 分词系统来进行文本预处理工作。该系统对中文书目目录的各个著录项的文本进行分词操作，这样将中文书目目录信息转换为词条信息的集合；然后，将词条集合中的冗余词条（如停用词、部分高频词和低频词等）删除。这样，对于任意一条中文书目，可以依据是否包含词条来构建一个词条向量，表示为：

$$q = [o_1, o_2, \cdots, o_n]^T \tag{1}$$

其中：n 表示词条的数量。元素 $o_i; i = 1,2,\cdots,n$ 表示第 i 个词条在中文书目内容中是否出现，出现则值为 1，否则为 0，也即：

$$o_i = \begin{cases} 1, & \text{第 } i \text{ 个词条在该图书内容中出现} \\ 0, & \text{其他} \end{cases} \tag{2}$$

这样，非结构化的中文书目文本数据转换为结构化向量数据。

1.2 特征提取

该部分主要任务是从中文书目对应的词条向量中抽取具有区分能力的特征。常用的文本特征提取方法有：词频（Word Frequency）、文档频次（Document Frequency）、词频－逆向文件频率（Term Frequency－Inverse Document Frequency，TF－IDF）、互信息（Mutual Information）、期望交叉熵（Expected Cross Entropy）、信息增益（Information Gain）、文本证据权（The Weight of Evidence for Text）。不同特征提取方法对不同的文本数据的表达能力不同，需要依据数据的分布来选择最合适的特征提取方法。在中文书目分类领域，词频特征和词频－逆向文件频率特征应用较多[12]。

1.3 机器学习

中文书目数据对应的特征向量需要经过机器学习方法构建的分类器来进行分类。目前，机器学习方法很多，如 Adaboost、决策树（Decision Tree）、随机森林（Random Forest）、人工神经网络（Nerve Net）、支持向量机（Support Vector Machine，SVM）、朴素贝叶斯（Naive Bayes）、深度网络（Deep Net）等。下面简要介绍中文书目分类领域常用的决策树、人工神经网络和支持向量机方法。

1.3.1 决策树

决策树以信息增益为训练依据，对训练样本集中的特征向量进行学习，构建由内部节点和节点组成的二叉树或多叉树结构。其中，每一个节点都包含一个逻辑判断函数，可以对输入该节点的特征进行判决，为其选择合理的分录路径。

1.3.2 人工神经网络

人工神经网络通过模拟人脑思维设计学习框架，以错误率为训练依据对网络中的权重和偏移量参数进行调整，寻找错误率最低时的网络参数来构建，可以对大规模样板数据充分学习，从而实现对未知数据的分类和预测。

1.3.3 支持向量机

支持向量机是建立在统计学习理论和结构风险最小原理基础上的一种机器学习方法，主要优点是可以实现小样本集的学习，泛化能力强，其决策函数仅由少数的支持向量确定，而不是样本空间的维数，这样不仅可以避免“维数灾难”，而且计算复杂度小，是目前应用范围较广、具有较好识别能力的机器学习方法。

2 改进的支持向量机中文书目自动分类

本文仍采用上述的基于机器学习的中文书目自动分类系统架构。与之相比，本文主要在文本特征提取部分进行改进，主要改进在于，将现有方法中常用的词频特征和词频－逆向文件频率特征进行融合，提高特征区分能力。并采用奇异值分解方法将特征矩阵变换到语义空间，增强特征的稳健性，最终提高中文书目分类的准确率。另外，在机器学习部分，针对中文书目分类的多元性，在现有二元 SVM 分类器的基础上设计联合 SVM 分类器，实现多类中文书目的自动分类。下面首先介绍本文方法涉及的基本理论，然后介绍本文方法的实现方法。

2.1 基本理论

本文方法涉及的基本理论主要有两个：奇异值分解和支持向量机，简要介绍如下。

2.1.1 奇异值分解

在线性代数中，奇异值分解（Singular Value Decomposition，SVD）是一种非常重要的矩阵分解，可以看作是正规矩阵酉对角化的推广。其数学公式为：

$$\boldsymbol{X} = \boldsymbol{LSR}^T \tag{3}$$

其中：$\boldsymbol{L}$ 和 $\boldsymbol{R}$ 分别表示左奇异向量矩阵和右奇异向量矩阵，$\boldsymbol{S}$ 表示奇异值的对角矩阵。$\boldsymbol{S}$ 的对角元素按从大到小的顺序进行排列。其中，奇异值越大，说明对应向量越重要。

奇异值分解与潜在语义索引（Latent Semantic Indexing）关系密切，对于词条和语料的关联矩阵，如果进行一次 SVD 分解，那么可以实现相似词条和语料的分类，同时得到词条和语料之间的相关性。因此，SVD 也可称为语义空间变换。通过语义空间变换，将高维的文本数据转换为较低维度的隐含语义空间。

2.1.2 支持向量机

SVM 的主要设计思想是寻找一个最优的分类超平面，使得分为不同类别的数据点之间的间隔最大。令 $\{x_1, x_2, \cdots, x_n\}$ 表示样本数据集，则 SVM 分类超平面可以表示为：

$$w^T x - b = 0 \tag{4}$$

其中：w 表示分类超平面的法向量，b 表示偏移量，x 表示分类超平面上的点。

寻找在两个类别的数据集上与分类超平面平行的两个超平面，表示为：

$$\begin{cases} w^T x - b = 1 \\ w^T x - b = -1 \end{cases} \tag{5}$$

两个平行超平面之间的距离为 $\frac{2}{\|w\|^2}$。如果训练样本是线性可分的，样本数据集中的所有样本点都需要在上述两个平行超平面之外，也即样本点 $x_i; i = 1,2,\cdots,n$ 满足 $w^T x_i - b \geqslant 1$ 或者 $w^T x_i - b \leqslant -1$。那么，在训练分类器是，通过最小化 $\|w\|^2$，可以得到最优的分类超平面。公式为：

$$\min \quad \varphi(\omega) = \frac{1}{2}\|\omega\|^2 \tag{6}$$

$$s.t. \quad y_i(\omega^T x_i + b) \geqslant 1 \quad i = 1,2,3,\cdots,n \tag{7}$$

其中：y_i 表示样本数据 x_i 的类别标签。当 x_i 为正样本时，$y_i = 1$；否则，$y_i = -1$。

通过最优化求解，可以得到最优的参数 w 和 b。这样，对于新输入的数据 x，计算 $w^T x - b$ 的值，如果该值大于 0，则判定该数据为正样本，否则判定为负样本。

SVM 对于小样本数据的处理性能好，泛化能力强。

2.2 实现方法

本文方法的实现主要包括三个环节：文本预处理、语义空间变化和语义特征向量提取、联合支持向量机分类。详细介绍如下。

2.2.1 文本预处理

本文仍采用 ICTCLAS 分词系统来进行文本预处理。与文献［12］不同的是，本文在进行文本数据结构化转换时，更注重词条出现频率信息而不是词条是否存在信息，这样利于更充分描述文本数据。具体地，对于任意一条中文书目 d，记录每一个词条出现的频率，可以得到一个向量 $f=[f_{1,d},f_{2,d},\cdots,f_{n,d}]^T$。其中，元素 $f_{i,d}$；$i=1,2,\cdots,n$ 表示第 i 个词条在中文书目 d 中出现的次数。这样，非结构化的中文书目文本数据转换为结构化向量数据。

在机器学习阶段，整个训练样本集中的所有中文书目文本数据可以转换为一个维数为 $n\times m$ 的矩阵 $\boldsymbol{F}$，其中，m 表示中文书目的数量。矩阵 F 可以表示为：

$$\boldsymbol{F}=\begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,m} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n,1} & f_{n,2} & \cdots & f_{n,m} \end{bmatrix} \tag{8}$$

其中，矩阵中任意元素 $f_{i,j}$；$i=1,2,\cdots,n$；$j=1,2,\cdots,m$；表示第 i 个词条在中文书目 j 中出现的次数。

2.2.2 语义空间变换与语义特征向量提取

一般地，词条与语料库之间存在隐含语义关系，本文通过挖掘两者之间隐含的语义空间，来描述词条与语料库之间的联系。本文采用常用的 TF－IDF 方法进行文本数据的转换。该方法在数据挖掘和信息检索领域应用广泛，其主要设计思想是：某一个词条在某文档中出现的频率越高，而在语料库的其他文档中出现的频率越低，则该词条对于该文档而言的重要程度越高。给定语料库 D，词条 t 和中文书目 d，$d\in D$。则中文书目 d 的权重可以表示为：

$$t_{t,d}=f_{t,d}\times\log(|D|/f_{t,D}) \tag{9}$$

其中：$f_{t,d}$ 表示词条 t 出现在中文书目 d 中出现的次数，$|D|$ 表示语料库中中文书目的数量，$f_{t,D}$ 表示语料库 D 中出现词条 t 的中文书目数量。

这样，对于任意一条中文书目，采用 TF－IDF 方法可以得到一个特征向量 $t=[t_{1,d},t_{2,d},\cdots,t_{n,d}]^T$。

在机器学习阶段，整个训练样本集中的所有中文书目文本数据可以采用 TF－IDF 方法转换为一个维数为 $n\times m$ 的矩阵 $\boldsymbol{T}$，表示为：

$$\boldsymbol{T}=\begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,m} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n,1} & t_{n,2} & \cdots & t_{n,m} \end{bmatrix} \tag{10}$$

也即，用每一个词条对语料库中每一个文档的权重来构建矩阵 $\boldsymbol{T}$，将非结构化的文本数据转换为结构化数据。

然而，当词条出现频次过大时，TF－IDF 方法得到的权重会下降，影响特征区分能力。为此，本文融合词频和 TF－IDF 特征，构建的特征矩阵可以表示为：

$$X=\lambda F+(1-\lambda)\boldsymbol{T} \tag{11}$$

其中：λ 表示加权权重。

类似地，特征向量之间的融合公式为：

$$q=\lambda f+(1-\lambda)t \tag{12}$$

为了特征矩阵的冗余，尽可能地反映词条与文档之间的原始关系，本文采用 SVD 方法对特征矩阵 $\boldsymbol{X}$ 进行分解，如公式（3）所示。奇异值的对角矩阵 $\boldsymbol{S}$ 的对角元素按从大到小的顺序进行排列。奇异值越大，说明对应的词条向量越重要，词条与文本的关联性越强。可见，采用 SVD 分解之后的三个矩阵能反映词条与语料库之间语义联系。因此，本文将上述变换过程称之为语义空间变换。考虑到奇异值下降速度非常快，前 10％的奇异值的和通常可以达到全部奇异值之和的 99％以上了。因此，本文采用前 k 个奇异值来近似描述矩阵。简化后的矩阵记为：

$$\boldsymbol{X}_k=\boldsymbol{L}_k\boldsymbol{S}_k\boldsymbol{R}_k{}^T \tag{13}$$

其中：与 $\boldsymbol{S}$ 相比，矩阵 $\boldsymbol{S}_k$ 中只保留对角元素的前 k 个奇异值，其他位置的奇异值置为 0。与 $\boldsymbol{L}$ 和 $\boldsymbol{R}$ 相比，矩阵 $\boldsymbol{L}_k$ 和 $\boldsymbol{R}_k$ 中只保留前 k 行向量，其他行的元素都置为 0。

这样，可以通过语义空间变换，将高维的文本数据转换为较低维度的隐含语义空间。具体地，对于任意一个中文书目所对应的特征向量 q，可以通过语义空间的变换将其转换为语义空间中相同维度的语义向量 q_k，表示为：

$$q_k=S_k{}^{-1}L_k{}^Tq \tag{14}$$

本文将语义向量作为文档的特征向量，据此进行文档的分类。

2.2.3 联合支持向量机分类

基于机器学习的书目分类方法通常需要构建分类器来完成文档所对应特征向量的分类任务。考虑到支持向量机泛化能力强，计算复杂度样本空间维数关联小的特点，本文选择支持向量机方法进行特征向量的学习与分类。

由前面介绍可见，*SVM* 分类器是一个二元分类器，分类结果只有正样本和负样本两类。对于书目而言，类别数肯定不止两类。为了实现多类书目数据的分类，本文设计联合 *SVM* 分类器，为每一个书目类别构建一个 *SVM* 分类器，通过各个 *SVM* 分类器的投票来得到最终的分类结果。在训练每一个书目的 *SVM* 分类器时，将训练数据集中该书目的数据看作正样本，而将其他书目的数据看作负样本，来训练 *SVM* 分类器。假设书目类别总数为 C，那么可以得到 C 个 *SVM* 分类器，记为：

$$SVM_i=\{w_i,b_i\mid i=1,2,3,\cdots,C\} \tag{15}$$

在分类时，对于输入数据 x，可以计算 C 个分类得分，记为：

$$s_i=w_i{}^Tx+b_i \tag{16}$$

本文选择分类得分最大的类别作为数据 x 的分类类别，表示为：

$$i*=\underset{i=1,2,3,\cdots,C}{\arg}(s_i) \tag{17}$$

在本文中，用于 *SVM* 训练和测试的数据为每一个文档所对应的语义向量 q_k。

3 实验与结果分析

本文通过中文书目的自动分类实验来验证本文所述的基于语义空间变换的中文书目数据挖掘方法的有效性。首先，我们从学校中文书目馆随机抽取了 5 个大类的中文书目作为实验数据集，包括 D 类书目 3 364 条，F 类书目 5 482 条，I 类书目

3 638条，K 类书目 2 874 条，T 类书目 4 877 条，共计 20 235 条中文书目信息。一般地，中文书目信息包括书号、价格、书名、分卷号、分卷名、作者、版本项、出版地、出版社、出版时间、页码、开本、内容摘要、读者对象、分类号等字段信息。本文与文献［12］一样，选取书名和内容摘要这两个字段作为实验的测试语料，因为这两个字段能有效反映中文书目的主题。考虑到基于机器学习的中文书目自动分类方法一般包括机器学习和类目分析两个阶段，这里将中文书目数据集分为两个子集，一个为训练数据子集，另一个为测试数据子集。其中，训练数据子集是从每一类书目中随机抽取一半书目条目构成的，剩下的一半放入测试数据子集。下面首先介绍本文方法的实验情况，然后再与现有中文书目分类方法进行性能对比，验证本文方法的优势。

3.1 本文方法实验分析

本文方法的训练步骤如下。

*Step*1：文本预处理，构建矩阵 $\boldsymbol{F}$；

Step2：TF－IDF 特征提取，构建矩阵 $\boldsymbol{T}$；

Step3：特征融合，构建特征向量 q 和矩阵 $\boldsymbol{X}$；

Step4：语义空间变换，得到矩阵 $\boldsymbol{L}_k$、$\boldsymbol{R}_k$、$\boldsymbol{S}_k$ 和 $\boldsymbol{X}_k$；

Step5：语义向量生成，得到语义向量 q_k；

Step6：机器学习，对不同类别的语义向量进行训练，为每一类中文书目构建一个 SVM 分类器。

本文方法的测试步骤是：

Step1：文本预处理，得到向量 f；

Step2：TF－IDF 特征提取，得到向量 t；

Step3：特征融合，得到特征向量 q；

Step4：语义向量生成，得到语义向量 q_k；

Step5：特征分类，得到对每一个类别的分类得分；

Step6：选择分类得分最大的类别作为分类结果。

本文方法涉及两个参数，分别是特征融合阶段的权重参数 λ 和 SVD 分解阶段的参数 k。下面通过实验来选择最优的参数。

图 1 给出了参数 λ 取值不同时本文方法的分类准确率分布情况（此时 SVD 阶段不进行约简）。当 $\lambda=0$ 时表示仅使用 TF－IDF 特征，当 $\lambda=1$ 时表示仅使用词频特征。由图 1 可见，当参数 λ 取值为 0.3 时中文书目的分类准确率最大。这说明，TF－IDF 特征的分类效果优于词频特征，融合 TF－IDF 特征和词频特征的分类效果优于单独采用一种特征的分类效果。

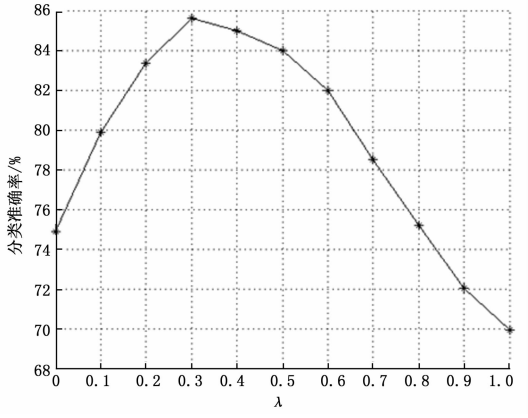

图 1 参数 λ 取值不同时分类准确率分布曲线

图 2 给出了参数 k 取值不同时本文方法的分类准确率分布情况。可见，前期随着 k 的增加，分类准确率提升。当 $k=80$ 时分类准确率增加不再明显，当 $k=120$ 时分类准确率反而下降。这说明，词条与文档之间的关联关系主要体现在前 80 个奇异值上，后面的奇异值所含噪声偏多，不利于分类。

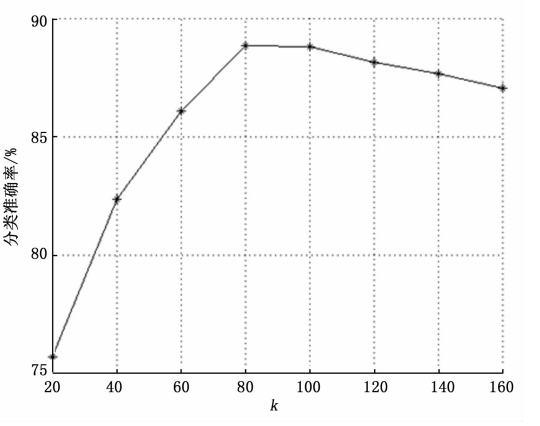

图 2 参数 k 取值不同时分类准确率分布曲线

3.2 不同方法对比分析

下面将本文方法与文献［12－14］所述的三种中文书目分类方法进行实验对比，具体结果见表 1。其中，文献［12］所述方法中特征选择混合特征，特征权重参数为 0.5。文献［13］所述方法中分类器选用其实验性能更优的 SVM 分类器。本文方法的实验参数为：$\lambda=0.3$、$k=80$。四种方法所用的实验环境相同，计算机平台性能参数为：Intel I7 CPU、DDR3 16 G 内存。软件开发环境为 Matlab 2012。机器学习模块使用 MATLAB 自带的开发包。分词系统都采用 ICTCLAS 分词系统。

表 1 不同方法分类准确率对比(单位：%)

方法	分类准确率					
	D 类	F 类	I 类	K 类	T 类	平均
文献[12]	83.12	86.89	87.86	88.14	84.37	86.08
文献[13]	80.64	82.98	83.64	85.17	81.33	82.75
文献[14]	83.09	84.45	86.39	88.04	84.58	85.31
本文	86.86	89.37	88.98	90.74	88.26	88.84

下面对实验结果进行具体的分析。本文方法与文献［12］所述方法都使用了词频和 TD－IDF 特征，不过本文方法没有区分特征在标题或者摘要中的差异，而是通过两类特征的加权融合以及语义空间变换来生成文本表示特征。这样可以去除冗余，增强特征的稳健性，提高分类准确率。由表 1 可见，本文方法在 D、F、I、K 和 T 五类书目的分类准确率都高于文献［12］方法，且平均分类准确率高于文献［12］方法 2.76%。与文献［13］方法相比，本文方法也使用了 SVM 分类器。然而在特征提取阶段，文献［13］中单独使用 TD－IDF 特征，而本文方法在此基础上融合了词频特征，特征区分能力增强。另外，本文方法在分类时构建联合 SVM 分类器，这也优于文献［13］方法使用的级联 SVM 分类器。因为使用级联分类器时如果某一层分类错误，那么分类结果就是错误的。而联合 SVM 分类器相当于每一个分类器都对分类结果进行投票，选择投票分数最高的类别作为最终的分类结果，这明显优于选择某一层分类结果。因此本文方法在五类书目上的分类准确率也

都高于文献［13］方法，且平均分类准确率高于文献［13］方法 6.09%。文献［14］所述方法与本文方法和文献［12－13］所述方法差异都较大，该方法的主要特点是构建复合特征，但在特征构建时使用了一些先验知识，导致特征的主观性较强，对数据的鲁棒性差。因此，在本文的测试数据下，该方法的分类准确率不高，在某些领域可能分类准确度较高，在五类书目上的分类准确率都低于本文方法，且平均分类准确率低于本文方法 3.53%。总的来说，本文方法对五类中文书目的分类准确度都高于其他三种方法，平均分类准确率高于其他方法 2.76%以上。

4　结束语

本文提出了一种基于语义空间变换的数据挖掘方法，主要设计思想是：融合词频和 TF－IDF 两种特征描述文本数据，结合奇异值分解实现语义空间变换，生成用于文本表示的语义向量，设计联合 SVM 分类器实现语义向量的学习与分类。通过进行中文书目自动分类实验，验证了本文方法能够提高中文书目分类的准确率。类似地，本文方法还可以用于其他文本分类与检索领域，有益于挖掘文本数据信息。

参考文献：

[1] Wu D，Olson D L. A TOPSIS Data Mining Demonstration and Application to Credit Scoring［J］. International Journal of Data Warehousing & Mining，2017，2（3）：16－26.

[2] Nassirtoussi A K，Aghabozorgi S，Wah T Y，et al. Text mining for market prediction：A systematic review［J］. Expert Systems with Applications，2014，41（16）：7653－7670.

[3] Mostafa M M. More than words：Social networks' text mining for consumer brand sentiments［J］. Expert Systems with Applications，2013，40（10）：4241－4251.

[4] He W，Zha S，Li L. Social media competitive analysis and text mining：A case study in the pizza industry［J］. International Journal of Information Management，2013，33（3）：464－472.

[5] Huh J，Yetisgen－Yildiz M，Pratt W. Text classification for assisting moderators in online health communities［J］. Journal of Biomedical Informatics，2013，46（6）：998－1005.

[6] Lin Y S，Jiang J Y，Lee S J. A Similarity Measure for Text Classification and Clustering［J］. IEEE Transactions on Knowledge & Data Engineering，2014，26（7）：1575－1590.

[7] D'Aspremont A. Predicting abnormal returns from news using text classification［J］. Quantitative Finance，2015，15（6）：999－1012.

[8] Sarker A，Gonzalez G. Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi－corpus Training［J］. Journal of Biomedical Informatics，2015，53：196－207.

[9] Uysal A K，Gunal S. The impact of preprocessing on text classification［J］. Information Processing & Management，2014，50（1）：104－112.

[10] Murtagh F，Kurtz M J. The Classification Society' s Bibliography Over Four Decades：History and Content Analysis［J］. Journal of Classification，2016，33（1）：6－29.

[11] Weldon S P. Organizing knowledge in the Isis bibliography from Sarton to the early twenty－first century.［J］. Isis；an international review devoted to the history of science and its cultural influences，2013，104（3）：540－550.

（上接第 236 页）

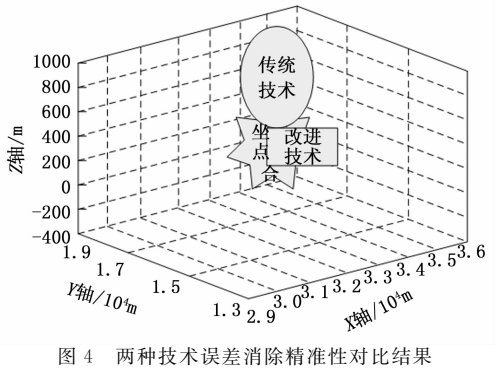

图 4　两种技术误差消除精准性对比结果

3.3　实验结论

根据实验内容，可得出实验结论：传统技术定位误差消除技术是建立在无人机自定位基础上实现的，对导航模块要求非常高，而改进技术是通过对轻小型无人机遥感范围内的显著目标进行标识与判断，并在遥感地图上进行标记，将误差消除，可减少灵敏程度上的定位误差，并从上述内容可得出结果，传统技术坐标点远远偏离正确坐标位置，而改进技术坐标点大部分与正确坐标一致，由此说明，改进技术误差消除精准度高。

4　结束语

轻小型无人机进行目标定位时，可适应复杂地形，构建以无人机为目标指示模型，可降低轻小型无人机自身产生的随机误差所带来的定位误差，进而提高目标定位指示精准度。与传统技术相比，改进技术主要结合了遥感地图通过目标匹配对参照物进行标记，无人机作为中继，消除飞行过程中其他不确定因素干扰所造成的目标定位误差。由实验结果可知，该技术误差消除精准度高，可为复杂地形目标定位提供有效技术支持。

参考文献：

[1] 贾配洋，彭晓东，沈菲菲，等. 基于 Apriltags 改进算法的无人机移动目标识别与跟踪［J］. 电子设计工程，2017，25（17）：31－35.

[2] 孙中宇，陈燕乔，杨　龙，等. 轻小型无人机低空遥感及其在生态学中的应用进展［J］. 应用生态学报，2017，28（2）：528－536.

[3] 黄海峰，林海玉，吕奕铭，等. 基于小型无人机遥感的单体地质灾害应急调查方法与实践［J］. 工程地质学报，2017，25（2）：447－454.

[4] 贾鹏宇，冯　江，于立宝，等. 小型无人机在农情监测中的应用研究［J］. 农机化研究，2015，20（4）：261－264.

[5] 王义坤，亓洪兴，韩贵丞，等. 轻小型面阵摆扫热红外成像系统研究［J］. 激光与红外，2015，45（10）：1216－1220.

[6] 黄海峰，易　武，张国栋，等. 引入小型无人机遥感的滑坡应急治理勘查设计方法［J］. 防灾减灾工程学报，2017，11（1）：99－104.

[7] 葛明锋，亓洪兴，王义坤，等. 基于轻小型无人直升机平台的高光谱遥感成像系统［J］. 红外与激光工程，2015，44（11）：3402－3407.

[8] 朱惠民，王航宇，孙世岩. 基于蒙特卡罗的单无人机侦察平台误差修正方法研究［J］. 科学技术与工程，2017，17（15）：255－259.

[9] 徐龙威，刘　晖，刘玉洁，等. 一种顾及 GNSS 系统间偏差的伪距单点定位方法［J］. 大地测量与地球动力学，2016，36（9）：813－816.

[10] 王亭亭，蔡志浩，王英勋. 小型无人机立体视觉目标追踪定位方法［J］. 电光与控制，2016，21（5）：6－10.