基于隐性语义索引的多标签文本分类集成方法

龚 静[1]，黄欣阳[2]

（1. 湖南环境生物职业技术学院 信息技术系，湖南 衡阳 421001；
2. 南华大学 计算机学院，湖南 衡阳 421001）

摘 要：针对多标签文本分类的概念歧义和底层语意结构问题，提出一种集成分类方法，将随机森林（RF）算法和隐性语义索引（LSI）有机结合在一起。通过词汇的随机分割增加集成的多样性，获得低维隐性语义空间的不同正交投影，在低维空间的正交投影基础上执行 LSI。随机森林可以有效解决二进制分类问题，隐性语义揭示了文本的底层语义结构，两者结合可代表群体的多样性和个体准确性。Yahoo 数据集上的实验结果验证了该方法的有效性，其在汉明损失、覆盖度、首位误差和平均精度方面优于其它方法。

关键词：文本分类；随机森林；多标签；正交投影；隐性语义索引

中图法分类号：TP391 **文献标识号**：A **文章编号**：1000-7024（2017）09-2556-06

doi：10.16208/j.issn1000-7024.2017.09.047

Multiple label text classification integration method based on latent semantic indexing

GONG Jing[1]，HUANG Xin-yang[2]

（1. Department of Information Technology，Hunan Polytechnic of Environment and Biology，Hengyang 421001，China；
2. College of Computer Science，University of South China，Hengyang 421001，China）

Abstract：Aiming at the concept of ambiguity and the underlying semantic structure for multiple label text classification，an integration classification method was presented，in which random forest（RF）algorithm and the latent semantic index（LSI）were combined. The diversity of integration was increased by the random segmentation of words，and the orthogonal projection of the low dimensional latent semantic space was obtained. Based on the orthogonal projection of the low dimensional space，LSI was implemented. Random forests can effectively solve the problem of binary classification，which reveals the underlying semantic structure of texts. And the combination of the two can represent the diversity of the population and individual accuracy. The effectiveness of the proposed method is verified by the experimental results on Yahoo data sets. It is better than several other methods in Hamming loss，coverage，first error and average accuracy.

Key words：text classification；random forest；multiple label；orthogonal projection；latent semantic index

0 引 言

文本分类[1-4]的目标是基于概念内容的相似性，将文本分配到一个或多个预定义标签（类别）中。由于数据集规模特别大，如何有效处理概念歧义和底层语意结构是研究的关键。

大部分多标签分类方法是利用不同标签间的关联性，以提高学习器的鲁棒性和通用能力[5]。已有研究大致可分为 3 大类：自适应方法、问题转化方法和集成方法。

自适应方法一般是将特定的学习算法（如决策树、支持向量机和粗糙集等）延伸，以直接处理多标签数据。如 Zhang 等[6]在模糊粗糙集模型的基础上，扩展用于处理多标签文本数据；Yu 等[7]提出了一种基于变精度模糊理论集的多标签分类方法，考虑了标签的相关性方面，以及特征空间与标签空间之间的映射不准确性。然而，Yu 等[7]指出该方法的结果取决于数据性质，对高维数据表现不佳。

问题转换方法是将多标签任务转换为一个或多个单标签任务，利用传统算法解决单标签问题。常见的有二进制

收稿日期：2016-11-18；修订日期：2017-01-03

作者简介：龚静（1972-），女，湖南岳阳人，硕士，副教授，研究方向为数据挖掘和分类算法等；黄欣阳（1971-），男，湖南祁阳人，硕士，副教授，研究方向为数据挖掘、信息安全等。E-mail：profgj70@126.com

关联（binary relevance，BR）方法。BR 将多标签问题转换为不同的二进制分类问题。Wang 等[8]训练每个二进制分类器去学习以往的预测标签对应的标签，捕捉标签的关联性，并在预测阶段细化分类。Zhou 等[9]运用转换思想实现对多标签分类方法的改进，以满足信息加权和阈值调整等特点。

集成方法是在前两类方法的基础上发展而来，是一种融合升级版。如 Tsoumakas 等[10]使用一种简单的投票方案确定最终分类集，将标签关联性纳入考量。不同于此，Groves 等[11]不使用标签投票方法，在标记步骤中使用特定数据的阈值，通过交叉验证程序进行校准。针对标签关系的相互独立、主次权重有别、数量要求不一的特点，Li[12]提出一种主次标签的多标签分类方法。

本文也是一种集成方法，选择随机森林[13]和隐性语义索引[14]（latent semantic indexing，LSI），其主要有两方面收益。一是随机森林可以很好地解决二进制分类问题；二是 LSI 揭示了文本底层隐性语义结构，提高了分类的精确度。本文的主要工作总结如下：①在低维空间的不同正交投影基础上，执行隐性语义索引；②提出的分类方法同时代表群体内的多样性和个体精确性。

1 隐形语义索引（LSI）

随着数据规模的爆炸式增长，以及说明性的特征越来越多，数据集的维度越来越高，在这样的数据集中，很多传统的信息检索和模式识别方法在计算效率和准确性方面受到了严重的限制。因此，由语义索引（或特征投影）带来的维数降低对文本的分析和处理至关重要。其中，LSI 可以有效克服词汇匹配问题，该方法假设在数据中存在一个底层隐性语义结构，在统计上捕捉文本术语中隐含的关联性结构，通过导出概念索引，在文本集合中检索。

LSI 是完全无监督式的，其本质是检测文本中最具代表性的特征，而非最具差异性的特征[14]。为了应用隐性语义索引，文本在向量空间模型中使用“文本逐个术语”矩阵 $x_{n,d}$ 表示（n 是文本的数量，d 是词语数量），通过执行 SVD 分解从高维度的输入空间到低维度的隐性空间中找到一个线性映射，使得大多数结构在数据中可以得到解释。

SVD 分解是将矩阵 x 分解到 3 个矩阵，即

$$x_{n,d}=\boldsymbol{U}_{n,d}\times\boldsymbol{S}_{d,d}\times\boldsymbol{V}_{d,d}^{\mathrm{T}} \tag{1}$$

式中：$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 是分别包含 x 的左奇异向量和右奇异向量的正交矩阵，$\boldsymbol{S}$ 是包含 x 的奇异值对角矩阵。SVD 中的奇异值有助于决定何种变量的信息性最强，以及哪种变量是无用的。当 $\boldsymbol{S}$ 中的奇异值以降序排列，对应着 k 个最大奇异值（$k\leqslant\min(n,d)$）的前 k 位奇异值向量被用于构建一个较低的 k 维空间。事实上，降维矩阵 $\boldsymbol{V}_{d,k}\times\boldsymbol{S}_{k,k}^{-1}$ 的乘积可以视为一种转换矩阵，即较高维特征的矩阵 $x_{n,d}$ 投影到一个较低维的矩阵，具体如下

$$x'_{n,k}=x_{n,d}\times\boldsymbol{V}_{d,k}\times\boldsymbol{S}_{k,k}^{-1} \tag{2}$$

在多种模型中，LSI 在文本分析和信息检索中得到广泛应用[7]，但主要是应用在单一分类问题上。由于 LSI 是一个完全无监督式的转换方法，考虑到多标签问题，需要多标签隐性语义索引，以确保数据中的信息能同时捕获标签间的关联性。

2 提出的多标签文本分类方法

2.1 算法描述

提出的多标签文本分类方法的主要思想包括在不同特征集和不同随机分割中应用特征提取算法，以集成多标签分类器属性。为此，本文选择 LSI 方法，具体算法如算法 1 所示。下面对这个框架作进一步阐述。

算法 1：$fun(x,\ y,\ F,\ d,\ L,\ N,\ K,\ k,\ S)$

变量：

训练数据中的文本（x），其中 $x_i=(x_{i1},\ \cdots,\ x_{id})$；

训练文本中的标签（y），其中 $y_i=(y_{i1},\ \cdots,\ y_{id})$；

输入空间（$F=\{f_1,\ \cdots,\ f_d\}$），描述性术语数量（d），标签数量（L），代表群体大小（N），特征子集数量（K），LSI 的维度降阶参数（k），基准多标签学习算法 BoosTexter 迭代次数（S）。

（1） $\mathcal{H}=\theta$ **do**

（2） for $i=1:N$

（3） 将输入空间 F 分割为 K 个子集 $F_{i,j}$（for $j=1\cdots K$）

（4） for $j=1:K$ **do**

（5） $x_{i,j}=x$ 中的自举样本，大小为 x 映射到 $F_{i,j}$ 的文本数量的 75%

（6）在 $x_{i,j}$ 中应用隐性语义索引，以得到矩阵 $T_{i,j}$ 中 k 维隐性空间转换系数

（7）**end for**

（8）在旋转矩阵 $\boldsymbol{R}_i$ 中排列 $T_{i,j}$ 转换矩阵，$j=1\cdots K$

（9）通过对 $\boldsymbol{R}_i$ 中的列重排构建 R_i^a，以匹配 F 中的特征阶。

（10）学习个体多标签分类器：

$h_i=BoosTexter(x\times R_i^a,\ y,\ S)$

（11） $H=H\bigcup h_i$

（12） **end for**

（13） **return** 生成的基准多标签分类器的代表群体 $\mathcal{H}$ 通过平均组合方法生成测试文本的标记概率。

设 $\mathcal{L}=\{\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_L\}$ 为一个标签空间，x 为一个由 n 个文本组成的数据集，每个文本的形式为 $(x_i,\ y_i)$，其中 $(x_{i1},\ \cdots,\ x_{id})$ 是 d 描述性术语（词或特征）的一个向量，$y_i\in\mathcal{L}$ 是与 x_i 关联的标签集（表述为二进制特征向量 $(y_{i1},\ \cdots,\ y_{i,L})\in\{0,\ 1\}$）。多标签分类器的集成代表群体用 $\mathcal{H}=\{h_1,\ h_2,\ \cdots,\ h_N\}$ 表示，用 F 表示术语集。和大多数集成方法一样，本文需要预先设定代表群体 N 的

大小。为构建分类器成员 h_i，本文采取以下步骤：

（1）和经典的随机森林算法一样，本文首先将 F 随机分割到 K 子集中（K 是算法的一个参数）。为促进多样性，这些子集可能不相交。为简化起见，假定 K 是 D 的一个因子，因此每个特征子集中都包含 $m=d/K$；

（2）用 $F_{i,j}$ 表示多标签分类器 h_i 的训练集第 j 个特征子集。对于每个这样的子集，本文均从中取一个样本，该样本大小为文本数量的 75%。运行隐性语义索引，仅使用 $F_{i,j}$ 中的 m 特征及 X 中选择的文本。储存转换矩阵的系数 $\boldsymbol{T}_{i,j}=\boldsymbol{V}_{d,k}\times\boldsymbol{S}_{k,k}^{-1}$，表示为 $T_{i,j}^{(1)}$，…，$T_{i,j}^{(K)}$，每个系数大小为 $m\times1$（K 是算法的一个参数，表示 LSI 中为降阶步骤所选择的低维隐性空间大小）。将 LSI 运行在文本中的某个样本上，而不是应用于整个文本集，其目的在于避免在同样的术语集中，选择不同分类器时出现相同的隐性空间；

（3）在一个稀疏正交矩阵 $\boldsymbol{R}_i$（$\boldsymbol{R}_i$ 不再是一个旋转矩阵）中，以系数来组织得到的向量，具体如下

$$\boldsymbol{R}_i=\begin{pmatrix} T_{i,1}^{(1)},\cdots,T_{i,1}^{(k)}, & & \cdots & \\ [0] & T_{i,2}^{(1)},\cdots,T_{i,2}^{(k)}, & \cdots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \cdots & T_{i,K}^{(1)},\cdots,T_{i,K}^{(k)} \end{pmatrix} \tag{3}$$

此对角矩阵的维数将是 $d\times\sum_j k$。为了降维，计算 $\sum_i k$ 以得到多标签分类器 h_i，本文首先重排 $\boldsymbol{R}_i$ 中的列，以使其对应原始术语。将重排的矩阵表示为 R_i^a，h_i 的低维特征矩阵集由 $x\times R_i^a$ 给出。

（4）基准多标签分类器 h_i 通过新数据空间的学习得到。该部分使用 BoosTexter 算法[15]。具体如下：已知一个文本 x，以 $P_i(x,\lambda_j)$ 表示标签 λ_j 和 x 相关联的概率，该概率通过多标签分类器 h_i 赋值。x 属于标签 H 的最终概率 $P_i(x,\lambda_j)$，由集成代表群体通过平均组合的方式给出，该方法使用集成代表 $P_i(x,\lambda_j)$（$i=1,2,\cdots,N$）生成标签概率

$$P_i(x,\lambda_j)=\frac{1}{N}\sum_{i=1}^{N}P_i(x,\lambda_j) \tag{4}$$

2.2 算法分析

在随机森林方法中，主成分分析（principal component analysis，PCA）是一种常见的特征空间转换方法。本文使用 LSI 而非 PCA，主要有两个理由：首先，PCA 本来是解决文本间的相关性或术语间的相关性而 LSI 则将两者一起分析[14]。其次，PCA 计算相关性矩阵的成本较高，而且大多数时候，文本数据的计算量是病态的。

值得一提，许多术语有不同的含义，这取决于所考虑的文件语境。LSI 主要是为了映射出术语之间的关联，以帮助解读文本的意义。因此，可在某种程度上解决同义词问题。但是，LSI 不能很好地处理多义词问题（即一个词语拥有多个含义），因为词汇中的每个术语会被考虑成输入空间的一个单点。而多标签随机森林会通过区分词汇表分割块的“语义”空间解决多义词问题。另外，本文选择 BoosTexter 作为基准学习器，是考虑到 BoosTexter 在文本分类中优良的预测表现，精确的基准学习器非常必要。

3 实验结果与分析

3.1 数据集

针对文本分类，本文使用了 9 个不同的标签，均来自 Yahoo 数据集，实验过程中使用了一个比较流行的数据挖掘工具 Weka，且这些数据集被分为了训练集和测试集，以尽可能使比较保持一致。基本统计数据见表 1。对于每个数据集，给出了特征数量（词汇大小）和标签数量；同时在测试集和训练集给出文本数量（# texts）；文本属于多个类别的百分比（# PM）以及每个文本的平均标签数（# AL）。

3.2 评价标准

本文使用以下流行的多标签文本评价标准：汉明损失、覆盖度、首位误差和平均精度。

表 1 多标签文本分类的数据集描述

数据集	标签数量	特征数量	训练集			测试集		
			# texts	# PM	# AL	# texts	# PM	# AL
艺术	32	490	1800	45.23%	1.61	2400	48.12%	1.72
商业	29	458	1800	54.33%	1.54	2300	50.17%	1.43
计算机	25	451	1800	43.11%	1.41	2700	40.55%	1.46
教育	27	502	1690	50.23%	1.39	3000	51.80%	1.37
娱乐	23	530	2000	27.34%	1.46	2100	27.13%	1.56
健康	31	609	2000	54.33%	2.90	2000	50.72%	2.87
医药	35	1439	1500	43.11%	2.31	1400	40.15%	2.24
消遣	32	890	1700	50.23%	1.41	1800	49.52%	1.19
参考	36	410	1700	27.34%	1.56	1900	29.10%	1.59

测试集表示为 $T_e=\{(x_1, y_1), (x_2, y_2), \cdots, (x_p, y_p)\}$，其中 $y_i \in \mathcal{L}$ 为真标签集，$\mathcal{L}=\{\lambda_1, \lambda_2, \cdots, \lambda_L\}$ 是全标签集。对于一个给定实例 x_i，其从多标签分类器 h 中的预测标签集表示为 $h(x_i)$，标签 λ_i 的估计排序表示为 $r_i(\lambda_j)$，相关性最高的标签排序最高，而相关性最低的标签则排序最低（L）。评价度量的讨论和数学公式如下：

汉明损失评价一个多标签分类任务中的精准性，计算错误预测标签占总标签数的百分比

$$Hl(h)=\frac{1}{p}\sum_{i=1}^{p}\frac{|y_i \Delta h(x_i)|}{L} \tag{5}$$

式中：Δ 是两个集之间的对称差分。

覆盖度评估标签列表中的平均距离，以覆盖所有适合的标签

$$Co(h)=\frac{1}{p}\sum_{i=1}^{p}\max_{\lambda_a \in y_i} r_i(\lambda_a)-1 \tag{6}$$

首位误差评估实例中排在首位的标签不在适合的标签集的次数

$$Oe(h)=\frac{1}{p}\sum_{i=1}^{p}\delta(\arg\min_{\lambda_i \in \mathcal{L}} r_i(\lambda_a)) \tag{7}$$

式中：$\delta(\lambda_a)=\begin{cases}1, \text{如果 } \lambda_a \in y_i \\ 0. \text{其它情况}\end{cases}$。

平均精度评估标签的排序高于一个特定标签 $\lambda_a \in y_i$ 而实际上在 y_i 的平均比率

$$Ap(h)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{|y_i|}\sum_{\lambda_a \in y_i}\frac{|\lambda_b \in y_i; r_i(\lambda_b) \leqslant r_i(\lambda_a)|}{r_i(\lambda_a)} \tag{8}$$

在汉明损失、覆盖度和首位误差中，分值越小，表现越好。平均精度的值越大，结果越好。

3.3 结果分析

比较的方法有两种集成方法（文献［11］和文献［12］）以及一种问题转化方法（文献［9］）。本文集成策略 N 和 BoosTexter 算法迭代次数设为 10，以使得集成分类器的数量为 100。且本文没有锁定 K 值，但每个子集特征数量定为 100，LSI 中的降维参数 k 设为 10。

表 2～表 5 给出了基准数据集上的比较结果。加粗为最好结果。在汉明损失中，本文方法在大多数情况下都获得了更好的分类效果，其中，文献［11］与本文方法的表现最为接近，只在艺术和娱乐数据集上略微超过本文方法，但总体本文方法更优。

覆盖度的结果见表 3，由表 3 可知，本文方法在 9 个数据集中排名第一。因此，在覆盖度方面，本文方法更优。表 4 的首位误差的结果显示：集成方法如文献［11］和文献［12］比问题转化方法文献［9］表现更好，包括本文方法在内的 3 种集成方法的表现大致相当。这也说明了集成方法对首位关联标签的排序是十分有效性的，这主要是因为问题转化方法，总是将多标签问题转化为单标签问题，造成了首位排序的误差较大。

表 2 多标签分类方法的汉明损失比较

数据集	文献[9]	文献[11]	文献[12]	本文方法
艺术	0.0613	**0.0504**	0.0547	0.0529
商业	0.0401	0.0349	0.0357	**0.0337**
计算机	0.0499	0.0389	0.0387	**0.0376**
教育	0.0598	0.0588	0.0601	**0.0552**
娱乐	0.0321	**0.0221**	0.0239	0.0267
健康	0.0312	0.0256	0.0241	**0.0219**
医药	0.0396	0.0342	0.0369	**0.0339**
消遣	0.0469	0.0435	0.0434	**0.0425**
参考	0.0367	0.0358	0.0398	**0.0334**

表 3 多标签分类方法的覆盖度比较

数据集	文献[9]	文献[11]	文献[12]	本文方法
艺术	7.363	4.980	5.090	**4.812**
商业	5.674	4.874	4.911	**4.093**
计算机	6.071	5.014	5.274	**4.374**
教育	8.042	6.032	6.373	**5.686**
娱乐	2.995	2.963	3.116	**2.873**
健康	3.936	3.226	3.214	**2.736**
医药	4.027	2.051	2.065	**1.629**
消遣	6.870	4.970	5.017	**4.851**
参考	4.341	2.981	3.051	**2.788**

表 4 多标签分类方法的首位误差比较

数据集	文献[9]	文献[11]	文献[12]	本文方法
艺术	0.598	0.477	0.444	**0.422**
商业	0.604	0.236	0.178	**0.127**
计算机	0.630	**0.340**	0.428	0.456
教育	0.601	**0.201**	0.277	0.237
娱乐	0.507	0.388	**0.287**	0.301
健康	0.499	**0.284**	0.297	0.290
医药	0.402	0.188	0.171	**0.156**
消遣	0.501	0.198	**0.182**	0.204
参考	0.487	0.215	0.241	**0.173**

表 5 是平均精度结果，可以看出，本文方法的表现明显优于其它方法，领先幅度较大。总体而言，本文方法能构建一个有益于文本分类的多标签集成分类器。对标签的排序更加令人满意，这得益于其更准确的标签概率预测。与其它 3 种方法比较，本文的概率策略可以得到一个机器学习集成分类框架，能明显提高分类效果，特别是在基于概率的排序度量中（如覆盖度，平均精度）。因此在统计上

具有更好的优势。

表 5 多标签分类方法的平均精度比较

数据集	文献[9]	文献[11]	文献[12]	本文方法
艺术	0.689	0.780	0.752	**0.813**
商业	0.801	0.807	0.821	**0.851**
计算机	0.523	0.621	0.687	**0.794**
教育	0.804	**0.873**	0.812	0.860
娱乐	0.814	0.875	0.824	**0.914**
健康	0.780	0.810	0.835	**0.901**
医药	0.812	0.819	0.802	**0.877**
消遣	0.753	0.793	0.801	**0.855**
参考	0.683	0.731	0.752	**0.837**

为进一步评价特征数量对分类结果的影响。这里研究了 4 个标准的平均值与特征数关系，选取特征数最高为 450。具体结果如图 1～图 4 所示。可以看出，基本上，所有方法是随着特征数的增加，评价结果越好，各方法在幅度上有所不同。图 1 是汉明损失结果，可以看出，随着特征数的增加，本文方法以及两种集成方法均优于问题转换方法。3 种集成方法的汉明损失的减小幅度差不多。覆盖度与特征数关系如图 2 所示。本文方法明显优于其它方法，验证了表 3 的结果。首位误差与特征数关系如图 3 所示。可以看出，随着特征数的增加，3 种集成方法的首位误差的减少幅度大致相当，这也说明了，集成方法在首位误差方面相近，而比问题转换方法更优。平均精度结果如图 4 所示，可以看出，随着特征数的增加，平均精度都有所增加，而本文方法平均精度的增加幅度明显高于其它 3 个方法，即更高的特征数会有更多收益。那么特征数有没有达到饱和的时候呢？并没有讨论，因为考虑到运行效率问题，太多的特征数会明显降低运行速度。

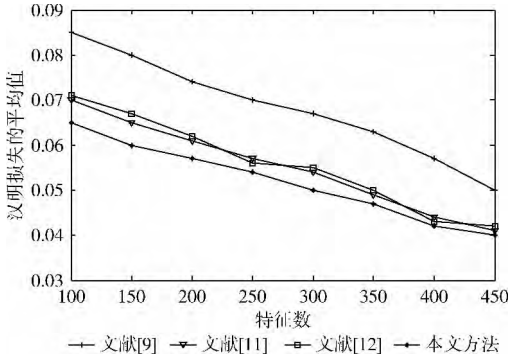

图 1 特征数与汉明损失的平均值关系

总体而言，本文的多标签集成分类器对文本分类更加令人满意，这得益于其更加准确的标签概率预计。与其它集成方法比较，提出方法的概率投票策略可以获得一个机器学习的集成分类框架，集成分类框架基于 3 个关键理念：①在低维概念空间的不同正交投影上使用 LSI；②随机分割词汇；③使用强大的 BoosTexter 进行文本分类。通过词汇的随机分割增加了集成的多样性，而文本中未发现的底层隐性语义结构提高了代表群体中代表成员的准确性。因此，这些理念的结合可以同时促进代表群体中的多样性和个体准确性，更好地理解文本数据的语义。另外值得一提，基于 LSI 词汇随机分割的投影和对 BoosTexter 的训练可独立操作，互不影响。

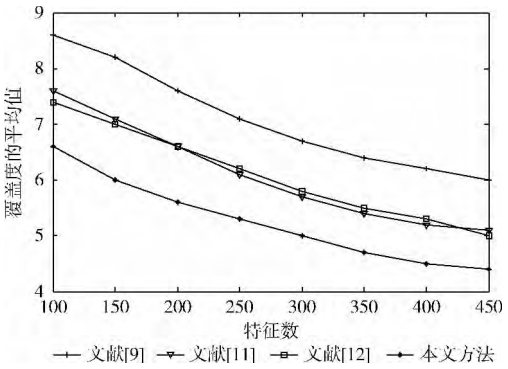

图 2 特征数与覆盖率的平均值关系

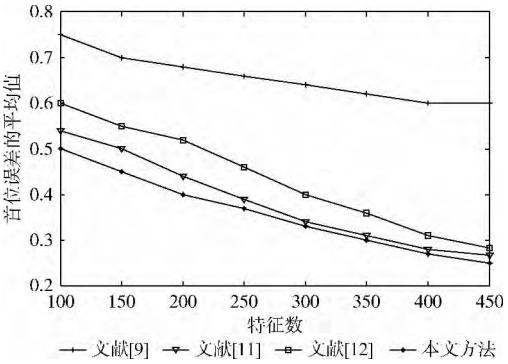

图 3 特征数与首位误差的平均值关系

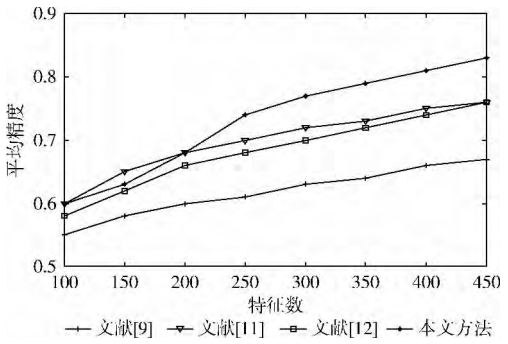

图 4 特征数与平均精度关系

3.4 运行效率比较

一个算法的执行时间取决于诸多因素，如处理器结构、

内存大小、操作系统、编程语言、编译器选项、编程技术和代码优化技术等。因此，将多个使用不同环境的文本分类方法进行比较并不合适，因此，这里只能大概比较各方法的运行速度。方法在 Intel Core i3 双核 CPU@ 2.29GHz RAM 8G 的台式机上运行，在 matlab2011b 平台上，采用 matlab 和 C 混合编程。从表 6 可以看出，作为一种简单的转化类方法文献［9］的运行速度非常快，这是因为处理多个单标签的时间是比较短的，而且多个单标签任务可以并行处理，而集成方法需要处理更多的标签和特征信息，要求更多的运行时间。

表 6　各方法的分类速度比较/ms

数据大小/KB	本文方法	文献[9]	文献[11]	文献[12]
10	420	280	490	510
600	3100	1900	5500	5100
6000	56 500	38 600	67 000	69 800

4　结束语

本文将随机森林和隐性语义索引结合起来，主要基于低维概念空间的不同正交投影和随机分割词汇等。促进在代表群体的多样性和个体准确性，通过文本中未发现的底层隐性语义结构促进代表群体中成员的准确性。由此提高文本分类能力。实验结果表明，提出的方法能明显提高平均精度，覆盖度和首位误差。

本方法的缺点是代表群体的大小有限制，基准多标签学习器运行的迭代次数，LSI 需要固定数值。未来着眼于处理更大数据集，例如在 Hadoop Map/Reduce 框架下。

参考文献：

［1］LUO Xianfeng，ZHU Shenglin，CHEN Zejian，et al. Improved KNN text categorization algorithm based on K-Medoids algorithm［J］. Computer Engineering and Design，2014，35（11）：3864-3867（in Chinese）.［罗贤锋，祝胜林，陈泽健，等. 基于 K-Medoids 聚类的改进 KNN 文本分类算法［J］. 计算机工程与设计，2014，35（11）：3864-3867.］

［2］Torii M，Yin L，Nguyen T，et al. An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics［J］. International Journal of Medical Informatics，2011，80（1）：56-66.

［3］Gomide Jana，Veloso A，Meira W，et al. Dengue surveillance based on a computational model of spatio-temporal locality of twitter［C］//International Web Science Conference. ACM，2011：1-8.

［4］CHEN Xin，SUN Jianjun. Exploration on the feasibility of application of design thinking in subject navigation［J］. Information Studies：Theory & Application，2015，38（1）：93-97（in Chinese）.［陈欣，孙建军. 关于设计思维在学科导航建设中应用可行性的探索性思考［J］. 情报理论与实践，2015，38（1）：93-97.］

［5］WANG Zhen. Multiple label classification algorithm based on learning label correlation［D］. Hefei：University of Science and Technology of China，2015（in Chinese）.［王臻. 基于学习标签相关性的多标签分类算法［D］. 合肥：中国科学技术大学，2015.］

［6］ZHANG Jing，LI Deyu，WANG Suge，et al. Multi-label text classification based on robust fuzzy rough set model［J］. Computer Science，2015，42（7）：270-275（in Chinese）.［张晶，李德玉，王素格，等. 基于稳健模糊粗糙集模型的多标记文本分类［J］. 计算机科学，2015，42（7）：270-275.］

［7］Yu Ying，Pedrycz Witold，Miao Duoqian. Multi-label classification by exploiting label correlations［J］. Expert Systems with Applications，2014，41（6）：2989-3004.

［8］WANG Jun，PENG Jiaxu，LIU Ou. A classification approach for less popular webpages based on latent semantic analysis and rough set model［J］. Expert Systems with Applications，2015，42（1）：642-648.

［9］ZHOU Hao，LI Xiang，LIU Gongshen. Adaptive algorithm for multi-label classification based on related information weighting［J］. Computer Applications and Software，2015，32（1）：239-243（in Chinese）.［周浩，李翔，刘功申. 相关信息加权的自适应多标签分类算法［J］. 计算机应用与软件，2015，32（1）：239-243.］

［10］Tsoumakas G，Katakis I，Vlahavas I P. Random klabelsets for multilabel classification［J］. IEEE Transactions on Knowledge and Data Engineering，2011，23（7）：1079-1089.

［11］Groves W. Multi-label classification methods for multi-target regression［J］. Computer Science，2014，43（12）：3209-3220.

［12］LI Xiao. Multiple label text classification method with primary and secondary labels［D］. Beijing：Beijing Institute of Technology，2015（in Chinese）.［李晓. 一种具有主次标签的多标签文本分类方法［D］. 北京：北京理工大学，2015.］

［13］HAN Min，LIU Ben. An improved classification algorithm of random forest［J］. Journal of Electronics and Information，2013，35（12）：2896-2900（in Chinese）.［韩敏，刘贲. 一种改进的随机森林分类算法［J］. 电子与信息学报，2013，35（12）：2896-2900.］

［14］XU Ge，WANG Houfeng. The development of topic models in natural language processing［J］. Chinese Journal of Computers，2011，34（8）：1423-1436（in Chinese）.［徐戈，王厚峰. 自然语言处理中主题模型的发展［J］. 计算机学报，2011，34（8）：1423-1436.］

［15］Silva P N D，Gonçalves E C，Plastino A，et al. Distinct chains for different instances：An effective strategy for multi-label classifier chains［M］//Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg，2014：453-468.