DOI: 10.19364/j.1674-9405.2018.06.006

基于主题模型的水利信息分类方案设计

诸葛庆子[1]，张审问[2]，蔡朝晖[1]，徐　华[2]，周　琦[1]

（1. 武汉大学计算机学院，湖北　武汉　430072；
2. 甘肃省水利厅信息中心，甘肃　兰州　730000）

摘　要：水利信息分类是水利科学数据共享标准化最为重要的一项工作，因此对水利领域大量数据信息的分类十分有必要。针对水利文本数据非结构化的特点，设计一个基于主题模型的水利文本信息分类方案，通过结合 LDA 主题模型和 GloVe 词向量模型的优点，提出一种新的主题模型。利用 AdaBoost 算法改进 KNN 分类器，在迭代中对分类器的错误进行适应性调整，最终得到分类器的集合。实验结果表明，使用 AdaBoost 提升 KNN 对于水利文本分类效果良好，分类效果远好于常见的朴素贝叶斯和决策树，和原来的 KNN 分类器相比，微观准确率提高 1.1 个百分点，宏观准确率提高了 4.1 个百分点，说明在水利文本分类中使用 AdaBoost 算法可提升 KNN 分类器的有效性。

关键词：主题模型；水利文本信息；文本分类；方案；LDA；GloVe

中图分类号：TV211　　**文献标识码**：A　　**文章编号**：1674-9405(2018)06-0027-08

0　引言

水利信息分类是进行水利信息交换和实现信息资源共享的重要前提，是水利科学数据共享标准化的一项最为重要的工作。目前，针对水利信息资源集成度低、有效利用率不高的情况，应对信息资源进行统一管理，因此，建立水利领域大量信息的分类十分必要。根据不同的业务需求和管理要求，从不同的角度出发，形成不同的水利信息分类体系。综观现有各分类体系，主要面临如下 2 个问题：1）现有的水利信息分类不能完全满足水利科学数据共享分类的要求；2）原有分类体系如何与共享分类体系对应。

建立分类体系，主要是对水利信息中大量文本数据的分类。文本数据分类的难点是特征的高维度和稀疏性，给分类算法带来以下 2 个问题：1）训练和分类时间上需要很大的开销；2）过多的特征往往会导致维数灾难问题。不同特征选择的方法对于不同场景下的文本分类有着不同的效果，在水利领域的文本分类中采用适当的特征选择方法有着重要意义 [1]。针对水利领域的非结构化文本数据的特点，设计一个基于主题模型的水利文本信息的分类方案，按照水利信息的科学属性进行分类。主题模型是一种自动化的无监督模型，在模式识别和自然语言处理等领域被使用，是能够在离散数据集中发现浅层主题信息的一种统计概率模型 [2]。直观来讲，如果 1 篇水利文档包含多个中心思想，那么一些表达这些主题思想的特定词语就会出现得比较频繁，就可以利用这些信息，建立一个多层的图模型，将语料、文档、主题、词等层面的信息，以及他们之间和内部的关联等信息融合起来，这些信息对水利文本分类、聚类、摘要、过滤等都非常有价值。

1　水利文本预处理

1.1　文本分词

使用基于概率统计语言模型的分词方法，利用 1 个包含 2 万多条词语的词典（包含哈工大、百度停用词表，搜狗词库，以及搜集整理水利中的专有名词），将词语放入一个 Trie 树中，利用 Trie 树高

收稿日期：2018-09-29

作者简介：诸葛庆子（1994-），女，湖北仙桃人，硕士研究生，研究方向：水利信息化。E-mail: 284827089@qq.com

效扫描词图并生成句子中所有可能成词情况，从 0 到 $n-1$（n 为句子的长度），每个开始位置作为词典的键，键值对的值为 value，里面存放了可能的词语结束位置，并将这些成词情况构成有向无环图，根据动态规划查找最大概率路径的方法，水利文本处理对句子采用从右往左的方式计算反向计算最大概率，由于汉语句子常常将重心落在后面，因此反向计算比正向计算的正确率更高。最后由最大概率路径获得分词结果，这种方法能够解决歧义词问题。

对于文本中较多的未登录词（各类专有名词、缩写词、新增词汇等），采用 HMM 模型和 Viterbi 算法。未登录词指词典中并未出现过的词，中文词汇有 B，E，M，S 等 4 个状态，B 是开始，E 是结束，M 是中间，S 是 singgle，使用 HMM 找到一个最佳的 BEMS 序列，使用 Viterbi 算法得到最佳的隐藏状态序列。在人工标注语料的情况下，使用 HMM 模型和 Viterbi 算法也能够单独对句子进行分词处理。

1.2 去除停用词

经过分词之后，水利文本中还有大量高频，但对于分类无意义的形容词和副词，还有一些出现频率不高的特殊符号和英文字符，这些词通常本身没有明确的意思，只有被放在一个完整的句子中才会有一定的作用。在文本分类中广泛使用停用词会轻易导致对有效信息的噪声干扰，也会影响文本分类器对于文章类别的判断，而通过文本特征选择不一定能被完全剔除。在特征加权和提取之前将大量无意义的中英文符号等噪声滤去非常重要，停用词的滤去也会帮助有效提高关键词的密度，减少词的数量和特征选择时的计算复杂度。

哈工大及百度停用词表总结整理了日常生活中许多的语气词、副词、形容词等，另外在水利文本中经常会出现例如“长江”“湖南”“湖北”这样一些名词，这些高频名词对于水利文本类别的判断同样没有任何意义，将哈工大和百度停用词表和这些单词整合起来，可作为本研究使用的停用词表，以滤去文本中的噪声。在对每篇文本去停用词时，需要将文本中的单词扫描 1 遍，对每个单词都在停用词表中查询，若存在则该单词去除。

2 基于词频的卡方检验提取特征

卡方检验（CHI）[3] 是一种常见的文本特征选择方法，卡方检验首先假设 2 个变量之间是相互独立的关系，然后对实际值和理论值的偏差进行计算，实际值可以通过观察得到，理论值是指两者确实独立情况下的预计值。当两者之间的偏差程度足够小，认为测量不够精确导致偶然误差的发生，因此可以认为两者之间相互独立，接受原假设；而偏差大到一定程度时，认为偏差不可能是偶然发生，即两者之间相互关联，这时否定原假设，选择接受备择假设。

在水利文本中使用卡方检验提取特征时，对于 M 篇水利文本，其中有 N 篇关于水利工程，考察特征词“水库”与类别“水利工程”之间的相关性，一共有以下 4 个观察值可以使用：

1）包含“水库”，而且类别是“水利工程”的文本数，命名为 A；

2）包含“水库”，而且类别不是“水利工程”的文本数，命名为 B；

3）不包含“水库”，但类别是“水利工程”的文本数，命名为 C；

4）既不包含“水库”，而且类别也不是“水利工程”的文本数，命名为 D。

卡方检验观察值统计表如表 1 所示。

表 1 卡方检验观察值统计表

特征项	类别 C_j	$\overline{C_j}$	总计
t_i	A	B	$A+B$
$\overline{t_i}$	C	D	$C+D$
总计	$A+C$	$B+D$	N

特征 t_i 出现在类别 C_j 中文本数的期望值 $E_{i,j}$ 为

$$E_{i,j}=(A+C)\ \frac{A+B}{N}, \tag{1}$$

则偏差项 $D_{\mathrm{ev}}(t_i, C_j)$ 为

$$D_{\mathrm{ev}}(t_i, C_j)=\frac{(A-E_{i,j})^2}{E_{i,j}}。 \tag{2}$$

特征项与类别相关的卡方检验值为 $x^2(t_i, C_j)$，则卡方检验的公式为

$$\begin{aligned} x^2(t_i, C_j)&=D_{\mathrm{ev}}(t_i, C_j)+D_{\mathrm{ev}}(\overline{t_i}, C_j)\\ &\quad+D_{\mathrm{ev}}(t_i, \overline{C_j})+D_{\mathrm{ev}}(\overline{t_i}, \overline{C_j}) \\ &=\frac{N(AD-BC)^2}{(A+C)(A+B)(B+D)(C+D)}。 \end{aligned} \tag{3}$$

考虑水利文本数量不平衡的特点和卡方检验中存在的低频缺陷问题，对卡方检验进行 3 个方面的改进。

2.1　特征项与类别的正负相关性

在卡方检验基本公式中 N 是该语料库所有文本的数量，是个常数，在对同一个类别中特征项计算卡方值时可以被忽略。其中的（$A+C$）代表了某一个类别的所有文本数，（$B+D$）代表了其他所有类别的文本数，同样作为常数可以被忽略。在进行卡方检验时，根据数学原理，若认为该词与分类类别相关性大，则认为文章中出现该词时很有可能是属于这个类别，而没有该词时很有可能不属于这个类别。参考公式（3），可以得出当 $AD-BC>0$ 时，即 $A\div C-B\div D>0$，一般是因为这个词在该类别文本中出现概率较高，而在别的类别文本中出现概率较低，认为该特征与该类别成正相关，即该特征可以代表这个类别；而 $AD-BC<0$ 时，一般是这个单词在这个类别文本中出现概率较低，而在别的类别文本中出现概率较高，认为该特征与该类别成负相关，该特征不能很好地代表这个类别。

为提高从语料库中提取和分类类别正相关的特征的能力，应判断特征项与分类类别的相关性，取其正相关性，去其负相关性，即当特征项与分类类别呈负相关时取值为 0。判断公式如下：

$$x^2(t_i, C_j)^+=\begin{cases}\dfrac{(AD-BC)^2}{(A+B)(C+D)}, & AD-BC>0\\ 0, & AD-BC<0\end{cases}。(4)$$

2.2　类间词频对数差

若 1 个单词能够在这个类别的大部分文本中都出现，且在每篇文本中出现的频率都较高时，可以认为这个词与这个类别分类关联性强。为了区别这个词在本类别中出现的频率和别的类别中出现的频率差，显示该词与本类别的关联，引入类间词频对数差因子 F，定义如下：

$$F(t_i, C_j)=\lg\left(\overline{w_{(t_i, C_j)}}\right)-\lg\left(\overline{w_{(t_i, \overline{C_j})}}\right), \quad (5)$$

式中：$w_{(t_i, C_j)}$ 为特征 t_i 出现在类别 C_j 中的总次数；$w_{(t_i, \overline{C_j})}$ 为特征出现在不属于类别 C_j 的文章中总次数，类别 C_j 中一共有文章数 n_j，定义语料库中文本总数为 N；$\overline{w_{(t_i, C_j)}}$ 是指在特征 t_i 出现在类别 C_j 的每篇文本中的平均词频。$\overline{w_{(t_i, \overline{C_j})}}$ 是指特征 t_i 出现了除类别 C_j 之外其他所有类别中每篇文本的平均词频。

类间词频对数差因子 F 考虑到单词在该类别每篇文本中出现的平均次数和在其他类别中出现的平均次数的差异，当一个特征项出现在这个类别每篇文本中的平均次数越多，而出现在别的类别中的平均次数越少，那么这个词和这个分类类别的相关性较大，即是这个类别的强分类特征；反之，这个词语与该分类类别的相关性较差，是这个类别的弱分类特征。如果直接将该词频因子加入卡方检验，对于特征选择的结果扰动会过大，进而会影响特征提取的效果，因此取单词在该类别中和别的类别中出现平均次数的对数差。

计算每个特征项和类别之间的 F 值，并乘以基于正负相关性考虑的卡方检验值，可得到改进后基于词频的卡方检验值，特征选择依然按计算所得值从高到低对特征进行排序，取排名靠前的特征，认为是对水利文本分类相关性较强的特征。

2.3　局部特征选择

当卡方检验用于全局特征选择时，既可以取它与每个类别计算卡方检验的最大值，也可以按它分布在每个类别的概率权重乘以卡方检验值计算最后结果。但这 2 种做法都无法保证在小类中选择特征项的数量，为了提高对小类的识别能力，引入一种局部特征选择的思想。

局部特征选择是将基于特征与类别正负相关性的卡方检验值与类间词频对数差因子相乘，对水利文本数据中的水利工程、水资源、水雨情、水土保持、自然环境和防汛抗旱 6 个类别，分别统计每个类别中特征项与该分类类别的相关性，并在每个类别中按照相关性从高到低排列。一般而言，对于局部的特征选择方法，在每个类别中提取的特征数与该类别的篇数成正比，但对于不平衡文本，采用此方法，获取到的代表小类的特征数较少。因此为了提高对小类别的识别能力，增加从小类别文本中获取到的特征项，在每个类别中取排名前 n 的特征。

3　结合 LDA 和 GloVe 模型的水利文本表示

3.1　水利文本主题建模

传统文本分类是将文本表示为向量空间模型，向量空间模型具有特征维度高和稀疏的特点，但不能表示文本中的语义。隐含狄利克雷分布（LDA）模型能够在一定程度上改善向量空间模型的缺点。LDA 模型认为一篇文本以一定概率分布在若干主题上，一个主题以一定概率分布在若干词语上，这 2 个多项分布的求解采用吉布斯采样，获得需要求解的概率分布的样本值从而反过来确定概率分布的样本

数。LDA 主题模型如图 1 所示。

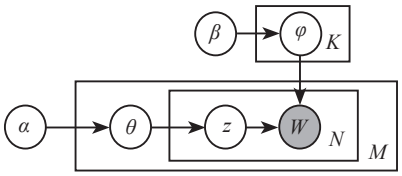

图 1 LDA 主题模型

图 1 中，阴影部分圆圈是可被观察的变量，非阴影部分圆圈是潜在变量，不能直接通过观测得到；M 为语料库中所有文章的数量；K 为语料中包含主题的个数（被手动设置）；W 为语料库中所有的词的个数；α 为超参数，每篇文档的主题分布中先验分布狄利克雷分布的参数；β 为超参数，是每个主题的词分布中先验分布狄利克雷分布的参数；θ 为一个 $M\times K$ 的矩阵，表示文本主题之间的关系；$\vec{\theta}_m$ 为第 m 篇文章的主题分布向量；φ 为一个 $K\times N$ 的矩阵，表示主题与词的关系；$\vec{\varphi}_k$ 是第 k 篇文章的词分布；z 表示一个主题。

一篇文章的生成步骤如下：首先生成这篇文章中的词所对应的主题，然后再生成词，即不考虑词位置的先后顺序，在主题被生成的情况下任意 2 个词的生成是可以交换的。这样就得到语料生成的联合分布概率 P，公式如下：

$$P(\vec{w},\vec{z}\mid\vec{\alpha},\vec{\beta})=P(\vec{w}\mid\vec{z},\vec{\beta})P(\vec{z}\mid\vec{\alpha})。\quad(6)$$

LDA 模型主要用来提取文本主题，所需求解的参数为主题变量 z 的后验分布 $P(z\mid w)$ 的参数，求解的概率计算表达式如下：

$$P(\theta,\vec{z}\mid\vec{w},\alpha,\beta)=\frac{P(\theta,\vec{z},\vec{w}\mid\alpha,\beta)}{P(\vec{w}\mid\alpha,\beta)}。\quad(7)$$

在实际建模中，由于训练过程过于复杂，采用吉布斯采样求解参数可以简化求解步骤。吉布斯采样先要进行随机初始化，对语料库中每篇水利文本中的每个特征词 w，采用随机的方式将其分配给一个主题，主题编号赋值为 z。对语料库中每篇文本重新扫描 1 遍，利用吉布斯采样公式重新采样语料库中每个特征词的主题，并不断在语料库中更新，然后在吉布斯采样收敛之前不断重复上述过程，最后统计主题和特征词的共现频率，得到主题词分布矩阵，即为 LDA 的模型，如表 2 所示。

3.2 利用 GloVe 生成文本和主题向量

词向量模型 GloVe 统计词共现矩阵，利用词共现矩阵中的非零元素对词向量进行训练。统计词共现矩阵过程中，第 i 行第 j 列的值为词 w_i 和 w_j 在整个语料库中共同出现次数 x_{ij} 的对数，x_i 表示词 w_i 上下文所有词出现次数综合，求得词 w_j 在词 w_i 上下文中出现的概率 P_{ij} 为

$$P_{ij}=P(j\mid i)=\frac{x_{ij}}{x_i}。\quad(8)$$

可以从 2 个词的共现概率中表示 2 个词的相关性。给定任意一个词通过计算 $F(w_i-w_j,w_k)=\frac{P_{ik}}{P_{jk}}$，判断词 w_k 和 w_i 及 w_j 的相关性，如果比值 $\frac{P_{ik}}{P_{jk}}>1$，说明词 w_k 和 w_i 的相关性更大，如果 $\frac{P_{ik}}{P_{jk}}<1$，则词 w_k 和词 w_j 的相关性更大。根据词共现概率的比值进行转化最终得到要最小化的代价函数 J：

$$J=\sum_{ik}f(x_{ik})(w_i^Tw_k+b_i+b_k-\lg x_{ik})^2,\quad(9)$$

式中：w_i^T 为矩阵 w_i 的转置；b_i 和 b_k 是为了解决对称性问题而引入的 2 个偏差项。

一般认为权重函数应符合以下 3 个特点：1）$f(0)=0$，如果 2 个词没有共同出现过，权重为 0；2）$f(x)$ 必须是非减函数，即随着词共现频率的增大，权重增大或者不变；3）对于较大的 x，$f(x)$ 不能取太大值。最后得到的权重函数如下：

$$f(x)=\begin{cases}\left(\frac{x}{x_{\max}}\right)^{\alpha}, & x<x_{\max}\\ 1, & x\geqslant x_{\max}\end{cases},\quad(10)$$

根据实验经验，选择 $x_{\max}=100$，$\alpha=\frac{3}{4}$。最后根据代价公式（9），使用双线性对数回归收敛代价函数，得到词向量，最后将每个单词映射到一个多维的向量空间上。例如，用水的词向量如下：-0.608784，0.037898，0.125220，-0.461315，

表 2 主题词分布

主题 15	主题 44
（达标率，0.041 407 80）	（洪水，0.037 099 58）
（水质，0.038 051 57）	（警戒水位，0.032 604 23）
（用水，0.032 760 15）	（暴雨，0.031 264 48）
（诸河，0.021 670 09）	（历史，0.012 289 92）
（用水量，0.019 524 42）	（上半年，0.010 970 76）
（取水，0.013 668 67）	（江苏，0.010 301 90）
（地表水，0.013 055 44）	（局部地区，0.009 723 80）
（营养，0.010 948 18）	（入汛，0.008 875 45）
（降水量，0.011 023 47）	（幅度，0.008 702 85）
（费，0.010 920 07）	（石臼，0.008 219 56）

−0.028186，0.340113，0.867771，−0.254154，−0.168743，−0.464037，1.086523，−0.021376，−0.510881，−0.076837，0.701027，…。

采用简单的方法，将水利文本每篇文章中所有词的向量累加，考虑文本有长有短，排除文章长度对生成文档向量的影响，对累加后的向量除以文章中词的个数求平均值，公式如下：

$$V(d)=\frac{\sum_{t\in d} w_t}{N(d)}, \tag{11}$$

式中：w_t 代表特征 t 的词向量；$N(d)$ 是文本 d 中词的总数；$V(d)$ 为文本向量。采用这种方法对每篇水利文本生成文本向量，使文本的维度从向量空间模型的上万维降低到数百维。

对于主题向量的生成，从表 2 可以看出主题分布在每个单词上的概率不同，在 LDA 生成的主题模型 k 个主题中，主题 15 分布在所有单词上的概率之和为 0.213 0，主题 44 分布在所有单词上的概率之和为 0.170 1。为了排除主题分布在单词上概率总和对生成主题向量的影响，对于第 k 个主题，这个主题中含有 n 个词，第 i 个词的词向量为 w_{ki}，主题 k 分布在该单词上的概率为 ϕ_{ki}，则生成主题 k 的向量可以表示为

$$V(t_k)=\frac{\sum_{i=1}^{n}\phi_{ki}w_{ki}}{\sum_{i=1}^{n}\phi_{ki}}。 \tag{12}$$

3.3　文本表示模型

对于一篇水利文本 i，生成的 m 维文本向量为 $d_i=\{a_{i1}, a_{i2}, ..., a_{im}\}$，$a_i$ 为每一维度的文本向量；对于主题 k，主题向量为 $t_k=\{b_{k1}, b_{k2}, ..., b_{km}\}$，$b_k$ 为每一维度上的主题向量。本研究使用余弦相似度度量 2 个向量之间的距离。

余弦相似度是机器学习中通过衡量 2 个向量间的夹角衡量 2 个向量的相似程度的方法，两向量之间的余弦值可以通过欧几里得点积和量级公式推导，欧几里得公式为 $a\times b=\|a\|\times\|b\|\cos\theta$，鉴于 2 个向量属性，余弦相似度 S 被表达为

$$S=\cos(\theta)=\frac{A\cdot B}{\|A\|\|B\|}=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}。 \tag{13}$$

产生的相似性范围为 −1～1，相似性为 −1，意味着 2 个向量所指的方向截然相反；1 表示它们所指方向是相同的；当相似性取 0 时，2 个向量之间相互独立。

最后利用文本与主题间的距离表示文本，模型如图 2 所示。

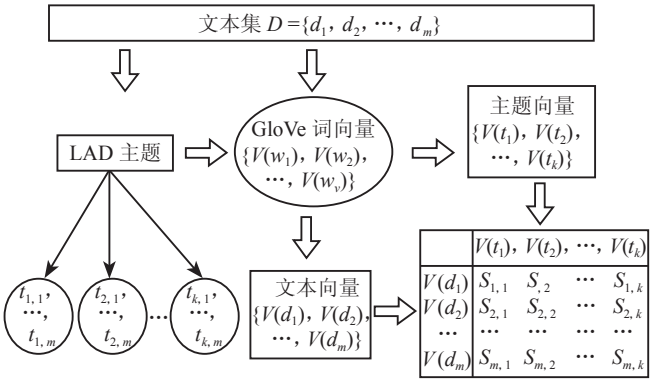

图 2　文本表示模型

4　有监督学习的水利信息分类

采用结合 LDA 和 GloVe 的文本表示模型表示水利文本后，将文本与主题的距离矩阵作为输入，使用分类器进行分类。KNN（邻近算法）是一种文本分类中常见的分类器，分类效果较好，同时也是一种弱分类器。对于弱分类器分类性能的提升，AdaBoost 算法是一种有效的方法。AdaBoost 算法从初始训练样本集得到基分类器，然后对训练样本进行调整，增加错分样本的权重，使用改变后的样本学习下一个分类器，重复学习得到 T 个分类器，并对这 T 个分类器的分类结果加权求均值得到最终分类结果。AdaBoost 训练分类器如图 3 所示。

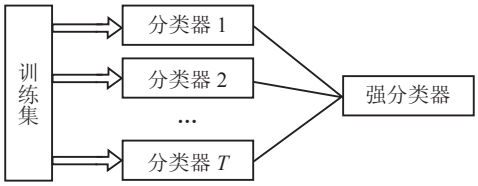

图 3　AdaBoost 训练分类器

利用 AdaBoost 算法将 KNN 作为基分类器生成 1 个分类器集合，得到 1 种线性组合的分类器模型，采用前向分布算法，首先确定初始分类器 $f_0(x)=0$，然后每一步都通过经验风险极小化确定下一个 KNN 分类器的参数。分类器训练算法如图 4 所示。

分类器训练算法经过 T 次的迭代，每一次迭代中根据当前的权重分布对样本 x 定义一个分布 P，在这个分布下的样本使用 KNN 算法得到 1 个分类

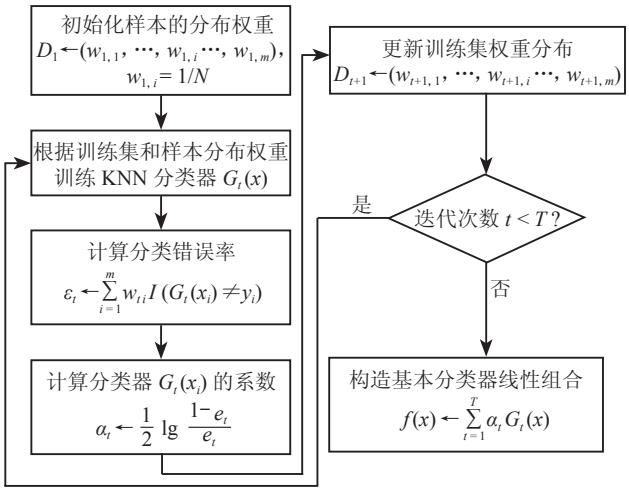

图 4 AdaBoost 提升 KNN 算法

器。通过每次迭代中更新权重，减小分类器分类效果较好的数据权重，增大分类器分类效果较差的数据权重，最终得到的分类器是多个 KNN 分类器的线性组合，分类结果取每个 KNN 分类结果的加权平均值。

5 实验结果分析

水利文本数据集采用从水利行业相关部门收集到的 2 056 篇水利文本，文本按照水利行业的特点被划分为自然环境、水利工程、水雨情、水资源、水土保持和防汛抗旱 6 类，各文本数量如图 5 所示。最大类与最小类的数量比大约达到 16∶1，数据具有不平衡性。

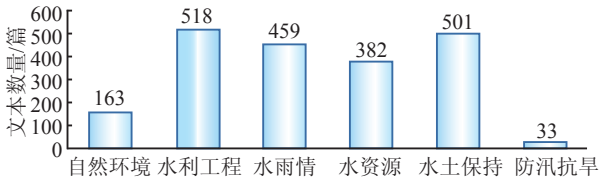

图 5 水利文本数据集

常用的特征选择方法有信息增益（IG）、卡方检验（CHI）、互信息（MI）3 种，分别使用 3 种常见的分类器朴素贝叶斯（NB）、决策树和 KNN 对特征选择后的水利文本进行分类，并采用宏观 F_1、准确率、召回率，以及微观 F_1、准确率、召回率对分类结果做全面的评估。具体分析如下：

1）使用 CHI 进行特征选择，取不同特征维度进行水利文本分类，实验结果如图 6 和 7 所示。

在使用卡方检验进行特征选择时，可以看到

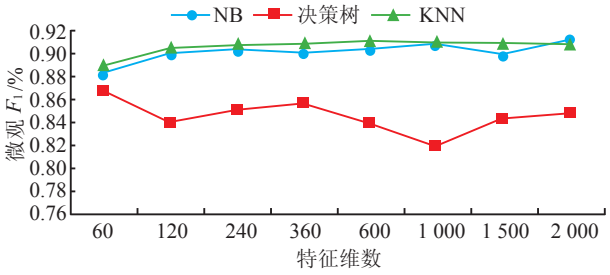

图 6 CHI 取不同维度的分类微观 F_1

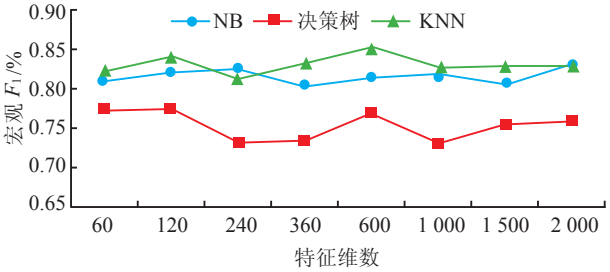

图 7 CHI 取不同维度的分类宏观 F_1

KNN 分类器在水利文本中表现良好。其中当特征维度为 600 维时，分类效果最好，此时微观 F_1 为 0.913，宏观 F_1 为 0.842，微观准确率为 0.914，微观召回率为 0.912，宏观准确率为 0.794，宏观召回率为 0.914。使用卡方检验的特征选择方法在维度较低的情况下可以获得较好的分类效果。

2）针对不平衡文本，使用基于词频改进的卡方检验（W_CHI）的实验结果如图 8 和 9 所示。

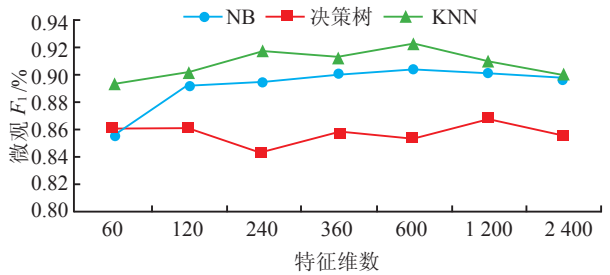

图 8 W_CHI 取不同维度的分类微观 F_1

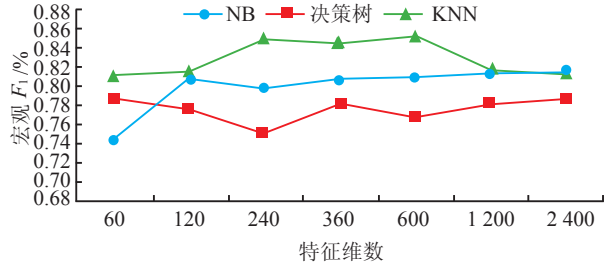

图 9 W_CHI 取不同维度的分类宏观 F_1

从实验结果可以看到，改进的特征选择方法 W_CHI 在特征维度为 240～600 的区间内，KNN 分

类器取得较好的分类效果，在维度为 600 时，取得最好分类效果，微观 F_1 为 0.924，微观准确率达到 0.924，微观召回率 0.924，宏观 F_1 为 0.854，宏观准确率 0.788，宏观召回率 0.935。

同样对 IG 和 MI 进行特征选择，取不同特征维度对水利文本进行分类，从实验结果得知，在特征维度为 1 200 维时，使用 IG 进行特征选择分类效果最好，此时，微观 F_1 为 0.911，宏观 F_1 是 0.816，微观准确率为 0.912，微观召回率为 0.916，宏观准确率为 0.786，宏观召回率为 0.839；而使用 MI 进行特征选择时，维度越高，分类效果越好。当特征维度低于 1 200 维时，使用 3 种分类器的情况下分类效果都远差于 IG 和 CHI。

为证实提出的特征选择方法的有效性，将特征维度取 1 200 维的 IG、取 600 维的 CHI、取 3 000 维的 MI，和特征维度为 600 维的 W_CHI 进行对比，分类器选择分类效果最好的 KNN，实验结果如图 10 和 11 所示。

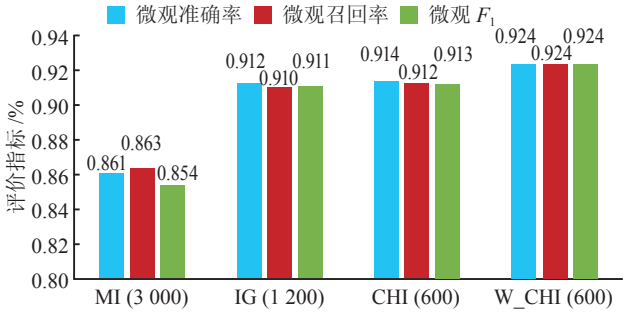

图 10　4 种特征选择的微观指标比较

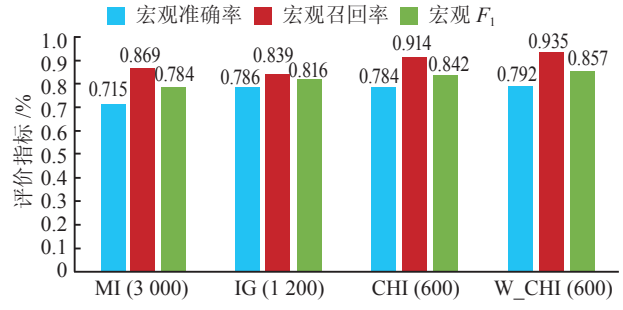

图 11　4 种特征选择的宏观指标比较

从图 10 和 11 中可以看到，CHI 对比 IG 有着微弱的优势，而 W_CHI 在多项指标中，相对于原来的 IG 和 CHI 均有着绝对的优势，在微观准确率和 F_1 上均提升了接近 1 个百分点。在宏观准确率上也有所提升。W_CHI 对于水利文本分类的效果最好，增加了每篇文本中高频、有效特征被选择的几率，能够改善 CHI 的“低频缺陷”。

使用 W_CHI 用于特征选择，然后使用结合 LDA 和 GloVe 的文本表示模型表示水利文本，最后使用 Adaboost 提升 KNN 的方法进行分类，与单独使用 KNN 作为分类器进行对比，实验结果如图 12 和 13 所示。

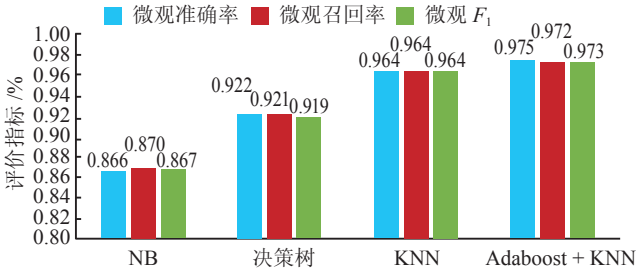

图 12　4 种分类器的实验结果比较

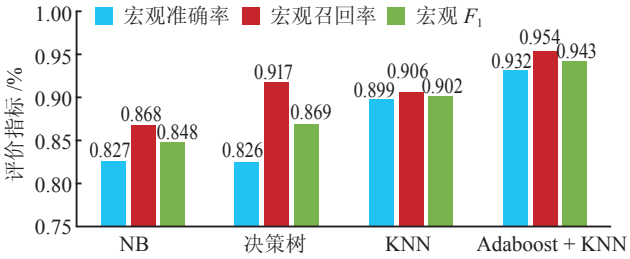

图 13　4 种分类器的实验结果比较

结果显示：使用 AdaBoost 提升 KNN 对于水利文本分类效果良好，分类效果远远好于常见分类器NB 和决策树，和原来的 KNN 分类器相比微观准确率提高了 1.1%，宏观准确率提高了 3.3%，微观 F_1 提高了 0.9%，宏观准确率提高了 4.1%。使用 AdaBoost 算法在宏观角度对水利文本分类的提升尤其明显，证明了在水利文本分类中使用 AdaBoost 算法提升 KNN 分类器的有效性。

6　结语

围绕着水利文本分类中的特征选择、文本表示和分类器优化 3 个方面进行研究，讨论出一种适合于水利文本的分类方法，结合改进卡方检验的特征提取、LDA 和 GloVe 的文本表示及 AdaBoost 提升 KNN 的分类方法，相对于传统的分类方法，大大提高了水利文本的分类效果，在小类别的分类效果上也有了很大的提升，对水利领域大量数据信息的分类具有重要的意义。但是其中也有不足和需要完善的地方，如简单地对文本或主题中所有单词的向量值进行累加求平均值，这种方法生成的文本和主题向

量都会存在一定的误差，如何利用 GloVe 获得更好的文本和主题向量，也是有待进一步研究的方向。

参考文献：

[1] YAO H, LIU C, ZHANG P, et al. A feature selection method based on synonym merging in text classification system[J]. Eurasip Journal on Wireless Communications & Networking, 2017, 2017 (1): 166.

[2] 张志飞，苗夺谦，高灿. 基于 LDA 主题模型的短文本分类方法[J]. 计算机应用，2013，33 (6): 1587−1590.

[3] 汪友生，樊存佳，王雨婷. 一种基于卡方统计的自适应特征选择方法：中国，CN105512311A[P]. 2016−04−20.

[4] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM Computing Surveys (CSUR), 2002, 34 (1): 1−47.

[5] JIN P, ZHANG Y, CHEN X Y, et al. Bag-of-embeddings for text classification[C]// International Joint Conference on Artificial Intelligence. NewYork, AAAI Press, 2016: 2824−2830.

[6] 段宏湘，张秋余，张墨逸. 基于归一化互信息的 FCBF 特征选择算法[J]. 华中科技大学学报（自然科学版），2017，45 (1): 52−56.

[7] 徐燕，李锦涛，王斌，等. 不均衡数据集上文本分类的特征选择研究[J]. 计算机研究与发展，2007，44 (增刊 2): 58−62.

[8] 谢娜娜，房斌，吴磊. 不均衡数据集上文本分类方法研究[J]. 计算机工程与应用，2013，49 (20): 118−121.

[9] LI L J, CHE Y Y, ZHANG H L, et al. KNN text categorization algorithm based on LSA reduce dimensionality[C]//Information Technology and Artificial Intelligence Conference. Chongqing: IEEE, 2011: 72−75.

[10] WANG Z B, MA L, ZHANG Y Q. A Hybrid document feature extraction method using latent dirichlet allocation and word2Vec[C]// IEEE First International Conference on Data Science in Cyberspace. Changsha: IEEE Computer Society, 2016: 98−103.

[11] CHU L L, GAO H, CHANG W B. A new feature weighting method based on probability distribution in imbalanced text classification[C]// International Conference on Fuzzy Systems & Knowledge Discovery. Yantai: IEEE, 2010: 2335−2339.

[12] TRICU L Q, TRAN T N, TRAN M K, et al. Document sensitivity classification for data leakage prevention with twitter-based document embedding and query expansion [C]// International Conference on Computational Intelligence and Security. HongKong: IEEE Computer Society, 2017: 537−542.

[13] PENNINGTON J, SOCHER R, MANNING C. GloVe: Global vectors for word representation[C]// Conference on Empirical Methods in Natural Language Processing. Doha: The Association for computational linguistics, 2014: 1532−1543.

Design of water conservancy information classification scheme based on theme model

ZHUGE Qingzi [1], ZHANG Shenwen [2], CAI Zhaohui [1], XU Hua [2], ZHOU Qi [1]

(1. School of Computer Science, Wuhan University, Wuhan 430072, China;

2. Information Center, Water Resources Department of Gansu Province, Lanzhou 730000, China)

Abstract: The classification of water conservancy information is the most important work of data sharing standardization in water conservancy science. So it is necessary to classify a large amount of data information in water conservancy fields. Aiming at the unstructured characteristics of water-based text data, a topic-based model of water-based text information classification scheme is designed. By combining the advantages of LDA theme model and GloVe word vector, a new topic model is proposed. The AdaBoost algorithm is used to improve the KNN classifier. It adaptively adjusts the error of the classifier in the iteration, and finally obtains the set of classifiers. The experimental results show that using AdaBoost to improve KNN has a good effect on classification of water conservancy texts, and the classification effect is much better than the common naive Bayes and decision trees. Compared with the original KNN classifier, the microscopic accuracy is improved by 1.1%, and the macro accuracy rate is improved by increasing 4.1 percentage points.It explain that the AdaBoost algorithm is used to improve the validity of the KNN classifier in the classification of hydraulic texts.

Key words: topic model; hydraulic text information; text classification; design; LDA; GloVe