基于非负矩阵分解的短文本特征扩展与分类

黄梦婷 张 灵 姜文超
(广东工业大学计算机学院 广州 510006)

摘 要 针对短文本特征稀疏的问题,提出了一种基于非负矩阵分解的特征扩展方法(NMFFE)。该方法只考虑数据自身,不借助外部资源进行短文本的特征扩展。首先,把文本及单词的内部关系考虑到文本和单词的关系矩阵分解中,通过双正则化非负矩阵三分解(DNMTF)方法获取词聚类指示矩阵;然后,对词聚类指示矩阵进行降维处理以获取特征空间;最后,根据单词之间的相关程度,从特征空间中选取特征并将其加入短文本中,从而解决短文本特征稀疏的问题,提高文本分类的准确率。实验数据表明,与 BOW 算法和 Char-CNN 算法中表现较优者相比,基于 NMFFE 算法的短文本分类的准确率分别在 Web snippets, Twitter sports 和 AGnews 数据集上提高了 25.77%,10.89%和 1.79%,这充分说明在分类准确率和算法鲁棒性方面,NMFFE 算法优于 BOW 算法和 Char-CNN 算法。

关键词 短文本分类,特征扩展,非负矩阵分解,特征空间,相关性

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/jsjkx.190400107

Short Text Feature Expansion and Classification Based on Non-negative Matrix Factorization

HUANG Meng-ting ZHANG Ling JIANG Wen-chao
(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract In this paper, a feature extension method based on non-negative matrix factorization (NMFFE) was proposed to overcome the sparse of short text feature. This method only considers the data itself and does not rely on external resources for feature extension. Firstly, the internal relationship of text and word is taken into account in the factorization of the relationship matrix between text and word , and word clustering instruction matrix is obtained by graph dual regularization non-negative matrix triple factorization (DNMTF) method. Then, word clustering instruction matrix is reduced in dimensionality to get the feature space. Finally, according to the degree of correlation between words, the feature in the feature space is added to the short text, thus solving the problem of feature sparse in short text and improving the accuracy of text classification. The experimental data show that compared with the better performance in BOW algorithm and Char-CNN algorithm, the accuracy of short text classification based on NMFFE algorithm is increased by 25.77%, 10.89% and 1.79% on the three datasets, which are Web snippets, Twitter sports and AGnews, respectively. The experimental data fully demonstrate that NMFFE algorithm is superior to BOW algorithm and Char-CNN algorithm in terms of classification accuracy and algorithm robustness.

Keywords Short text classification, Feature extension, Non-negative matrix factorization, Feature space, Correlation

1 引言

随着大数据时代的到来,互联网上的信息呈爆炸式增长,用户在各种媒体平台上生成的数据占据了网络信息资源的主体。短文本作为一种便捷的信息传输形式,具有更新速度快、易于扩散的特点,因此互联网中积累了海量的短文本数据。短文本存在的字数限制和编写不规范的特点,使得提取的特征具有稀疏性。因此,解决短文本特征稀疏的问题成为短文本研究领域的一个热点。

考虑到短文本特征的稀疏性和低频率,传统的词袋(Bag of Words, BOW)表示不再是分析短文本最合适的模型。处理稀疏性的一种解决方案是通过 Web 搜索、词汇数据库或机器翻译提供的语义信息来扩展短文本特征[1],这种方法被称为基于外部资源的方法。基于 Web 搜索的特征扩展技术[2]需要与具有高通信开销和高索引成本的搜索引擎进行交互,随着短文本数量的增加,数据分析的效率受到严重影响。为了解决这个问题,许多大型的知识库或语料库(如 Wikipedia 和 HowNet)被公开,用于进行显式概念分类[3-5]或通过隐含

到稿日期:2019-04-18 返修日期:2019-07-28 本文受广东省自然科学基金(2018A030313061),广东省科技计划(2017B030305003,2017B010124001),广东省产学研合作项目(2017B090901005)资助。

黄梦婷(1994—),女,硕士生,主要研究方向为数据挖掘与分析;**张 灵**(1968—),女,博士,教授,主要研究方向为智能化信息处理、自动化装备、人工智能和计算机视觉等;**姜文超**(1977—),男,博士,讲师,主要研究方向为云计算、高性能计算、分布式系统等,E-mail:june4567@21cn.com(通信作者)。

主题建模[6-7]来丰富短文本表示。然而,基于外部资源的特征扩展方法在特征扩展过程中极为耗时,且分类效果依赖于外部资源的完整性;另外,针对一些专业性很强或者语言比较特殊的短文本时,这些预定义的主题和分类可能不再适用。

另一种解决稀疏性问题的方法是使用隐藏在当前短文本上下文中的规则或统计信息来扩展特征,这种方法被称为基于自身资源的方法[5]。由于数据资源的限制,挖掘短文本中的额外信息成为了特征扩展的关键。Kim 等[8]提出了一种 LIS(Language Independent Semantic)内核,其通过提取文本的语法特征来计算短文本之间的相似性,无须使用语法标签和词法数据库。相比在词组层面进行语义或语法分析的文本分类,Zhang 等[9]提出了 Char-CNN 模型,其从字符层面进行文本分类,不考虑语义、语法信息和单词含义,因此可以跨语言使用。然而,这些研究都忽略了短文本中单词的相关性。在文本受限的情况下,单词之间的联系可成为额外信息,作为特征扩展的重要依据,有助于解决短文本特征稀疏的问题。

本文提出了一种基于自身资源特征扩展的短文本分类方法,该方法从短文本自身出发,考虑了数据的两种关系,即文本和单词之间的类型间关系和单词与单词(文本与文本)间的类型内部关系。基于这两种关系,通过非负矩阵分解得到词聚类指示矩阵,再通过降维处理获取特征空间,然后根据单词的相关性在特征空间中选择特征进行扩展,从而有效地解决了短文本特征的稀疏性问题。

2 相关工作

2.1 问题描述

给定短文本训练集 $T=\{t_1,\cdots,t_m\}$,对其进行预处理,从 T 中获取单词集 $W=\{w_1,\cdots,w_n\}$。其中,T 和 W 的样本数量分别为 m 和 n,关系矩阵 $\boldsymbol{R}_{n\times m}$ 描述单词和文本的类型间关系,关联矩阵 $\boldsymbol{A}_t$ 和 $\boldsymbol{A}_w$ 分别描述文本和单词的类型内部关系,$\boldsymbol{D}_t$ 和 $\boldsymbol{D}_w$ 分别为 T 和 W 的度矩阵,$\boldsymbol{L}_t$ 和 $\boldsymbol{L}_w$ 分别为 T 和 W 的拉普拉斯矩阵。聚类过程中,将 T 和 W 划分成 k 类,聚类指示矩阵 $\boldsymbol{F}_{n\times k}$ 描述单词集 W 的聚类结果,$\boldsymbol{F}_{ij}$ 表示 w_i 属于类 k_j 的可能性。同理可得短文本集 T 的聚类指示矩阵 $\boldsymbol{G}_{m\times k}$,由于训练集的短文本类别标签是已知的,因此矩阵 $\boldsymbol{G}$ 是已知的。本文将作为特征扩展依据的特征空间构造问题转换成文本和单词的联合聚类问题。

2.2 整体框架

基于非负矩阵分解的短文本特征扩展与分类的整体框架,主要包括特征空间构造、短文本特征扩展、特征空间更新和文本分类 4 部分,如图 1 所示。

特征空间作为特征扩展的依据,来自于短文本自身,描述的是单词所属类别的可能性。考虑到文本和单词之间的类型间关系和单词与单词(文本与文本)间的类型内部关系,分别构造关系矩阵和关联矩阵,通过加入流形正则约束对关系矩阵进行非负矩阵分解,从而得到词聚类指示矩阵,再对其进行降维处理(删除类别可能性分配比较均匀的特征),得到特征空间;短文本特征扩展是利用特征空间中的特征与文本特征的相关性对短文本进行扩展;特征空间的更新是通过同一文本中已知特征的聚类指示平均值来表示未知特征的聚类指示值,然后将新特征加入特征空间;文本分类完成分类器的训练及测试。

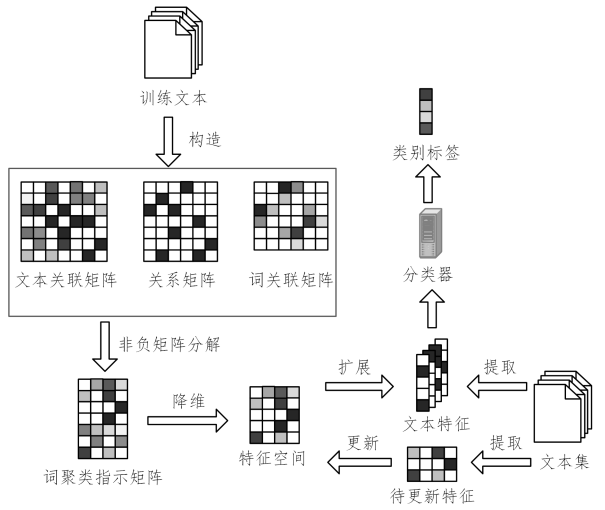


图 1 短文本分类的框架

Fig. 1 Framework of short text classification

3 基于非负矩阵分解的特征空间构造

3.1 非负矩阵分解框架

特征空间构造的主要内容就是对关系矩阵进行分解。首先,根据短文本训练集的预定义类别可以直接得到聚类指示矩阵 $\boldsymbol{G}$,将其作为关系矩阵 $\boldsymbol{R}$ 基于非负矩阵三分解的输入;然后,在关系矩阵分解的基础上加入流形正则化处理,最终分解得到词聚类指示矩阵 $\boldsymbol{F}$。

将关系矩阵 $\boldsymbol{R}$ 分解为 3 个矩阵 $\boldsymbol{F},\boldsymbol{S},\boldsymbol{G}$,使得 $\boldsymbol{R}\approx\boldsymbol{FSG}^{\mathrm{T}}$。其中,矩阵 $\boldsymbol{F}$ 和 $\boldsymbol{G}$ 分别为两类实体的聚类指示矩阵,矩阵 $\boldsymbol{S}$ 为具有一定自由度的平衡矩阵,当对 $\boldsymbol{R}$ 进行分解时,$\boldsymbol{S}$ 提供的自由度可以保证低维矩阵表示的准确性[10]。

3.2 关系矩阵和关联矩阵的构造

关系矩阵 $\boldsymbol{R}$ 的构造遵循文本和单词之间的自然关系。如果文本 t_j 中出现单词 w_i,则 $\boldsymbol{R}_{ij}=1$,否则 $\boldsymbol{R}_{ij}=0$。

关联矩阵 $\boldsymbol{A}_t$ 和 $\boldsymbol{A}_w$ 的构造需要借助文本和单词之间的统计信息来实现。两个样本 x_i 和 x_j 之间的关联强度的计算方法如式(1)所示:

$$\boldsymbol{A}_{ij}=\frac{B(x_i,x_j)}{\sum\limits_{x_a,x_b\in T(W)}B(x_a,x_b)} \tag{1}$$

其中,$B(x_i,x_j)$ 为短文本集 T(单词集 W)中的样本 x_i 和 x_j 共同出现的单词(文本)数。

3.3 流形中的关系矩阵分解

根据流形假设[11],如果两个样本 x_i 和 x_j 在几何结构中相近,那么这两个样本的现实意义也相近,在聚类中体现为两个样本的聚类标签相近。因此,本文采用 Shang 等[12]提出的双正则化非负矩阵三分解(DNMTF)方法,考虑了两类实体的类型间关系,同时还考虑了同类实体的类型内部关系。基于流形正则化的关系矩阵分解如式(2)所示:

$$\begin{aligned}&J_1=\|\boldsymbol{R}-\boldsymbol{FSG}^{\mathrm{T}}\|^2+\mu tr(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_w\boldsymbol{F})+\varphi tr(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}_t\boldsymbol{G})\\&\text{s.t. }\boldsymbol{F}\geqslant 0,\boldsymbol{S}\geqslant 0,\boldsymbol{G}\geqslant 0\end{aligned} \tag{2}$$

其中,μ 和 $\varphi>0$ 为正则化参数,用于平衡式(2)中第一项因

聚类重构产生的误差和第二、三项聚类标签的平滑度；拉普拉斯矩阵 $\boldsymbol{L}_w=\boldsymbol{D}_w-\boldsymbol{A}_w$，$\boldsymbol{L}_t=\boldsymbol{D}_t-\boldsymbol{A}_t$，度矩阵 $\boldsymbol{D}_w$ 和 $\boldsymbol{D}_t$ 为对角矩阵，$\boldsymbol{D}_{ii}^w=\sum_j \boldsymbol{A}_{ij}^w$，$D_{ii}^t=\sum_j \boldsymbol{A}_{ij}^t$。

对于式(2)的求解，采用选择性求解变量 $\boldsymbol{F}$ 和 $\boldsymbol{S}$ 的方式。根据短文本训练集的预定义类别，可以直接得到聚类指示矩阵 $\boldsymbol{G}$ 作为 J_1 的输入。式(2)中的目标函数可以改写成式(3)：

$$\begin{aligned}J_1 &= tr((\boldsymbol{R}-\boldsymbol{FSG}^{\mathrm{T}})(\boldsymbol{R}-\boldsymbol{FSG}^{\mathrm{T}})^{\mathrm{T}})+\mu tr(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_w\boldsymbol{F})+\\ &\quad \varphi tr(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}_t\boldsymbol{G})\\ &= tr(\boldsymbol{RR}^{\mathrm{T}})-2tr(\boldsymbol{RGS}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}})+tr(\boldsymbol{FSG}^{\mathrm{T}}\boldsymbol{GS}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}})+\\ &\quad \mu tr(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_w\boldsymbol{F})+\varphi tr(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}_t\boldsymbol{G})\end{aligned} \tag{3}$$

引入拉格朗日乘数 $\boldsymbol{\alpha}_{n\times k}$，$\boldsymbol{\beta}_{m\times k}$，$\boldsymbol{\gamma}_{k\times k}$，得到拉格朗日函数，如式(4)所示：

$$\begin{aligned}L= &\ tr(\boldsymbol{RR}^{\mathrm{T}})-2tr(\boldsymbol{RGS}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}})+tr(\boldsymbol{FSG}^{\mathrm{T}}\boldsymbol{GS}^{\mathrm{T}}\boldsymbol{F}^{T})+\\ &\mu tr(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_w\boldsymbol{F})+\varphi tr(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}_t\boldsymbol{G})+tr(\boldsymbol{\alpha F}^{\mathrm{T}})+tr(\boldsymbol{\beta G}^{T})+\\ &tr(\boldsymbol{\gamma S}^{\mathrm{T}})\end{aligned} \tag{4}$$

求解矩阵 $\boldsymbol{S}$ 的过程中，把矩阵 $\boldsymbol{F}$ 和 $\boldsymbol{G}$ 看作已知条件，设置 $\frac{\partial L}{\partial S}=0$，得：

$$\boldsymbol{\gamma}=2\boldsymbol{F}^{\mathrm{T}}\boldsymbol{RG}-2\boldsymbol{F}^{\mathrm{T}}\boldsymbol{FSG}^{\mathrm{T}}\boldsymbol{G} \tag{5}$$

根据 KKT 条件[13] $\boldsymbol{\gamma}_{ij}\boldsymbol{S}_{ij}=0$，得到：

$$[\boldsymbol{F}^{\mathrm{T}}\boldsymbol{RG}-\boldsymbol{F}^{\mathrm{T}}\boldsymbol{FSG}^{\mathrm{T}}\boldsymbol{G}]_{ij}\boldsymbol{S}_{ij}=0 \tag{6}$$

根据式(6)，矩阵 $\boldsymbol{S}$ 遵循以下更新规则：

$$\boldsymbol{S}_{ij}\leftarrow\boldsymbol{S}_{ij}\frac{[\boldsymbol{F}^{\mathrm{T}}\boldsymbol{RG}]_{ij}}{[\boldsymbol{F}^{\mathrm{T}}\boldsymbol{FSG}^{\mathrm{T}}\boldsymbol{G}]_{ij}} \tag{7}$$

求解矩阵 $\boldsymbol{F}$ 的过程中，把矩阵 $\boldsymbol{S}$ 和 $\boldsymbol{G}$ 看作已知条件，设置 $\frac{\partial L}{\partial F}=0$，得：

$$\boldsymbol{\alpha}=2\boldsymbol{RGS}^{\mathrm{T}}-2\boldsymbol{FSG}^{\mathrm{T}}\boldsymbol{GS}^{\mathrm{T}}-2\mu\boldsymbol{L}_w\boldsymbol{F} \tag{8}$$

将 $\boldsymbol{L}_w=\boldsymbol{D}_w-\boldsymbol{A}_w$ 代入式(8)，并结合 KKT 条件[13] $\boldsymbol{\alpha}_{ij}\boldsymbol{F}_{ij}=0$，得：

$$[\boldsymbol{RGS}^{\mathrm{T}}-\boldsymbol{FSG}^{\mathrm{T}}\boldsymbol{GS}^{\mathrm{T}}-\mu\boldsymbol{D}_w\boldsymbol{F}+\mu\boldsymbol{A}_w\boldsymbol{F}]_{ij}\boldsymbol{F}_{ij}=0 \tag{9}$$

根据式(9)，矩阵 $\boldsymbol{F}$ 遵循以下更新规则：

$$\boldsymbol{F}_{ij}\leftarrow\boldsymbol{F}_{ij}\frac{[\boldsymbol{RGS}^{\mathrm{T}}+\mu\boldsymbol{A}_w\boldsymbol{F}]_{ij}}{[\boldsymbol{FSG}^{\mathrm{T}}\boldsymbol{GS}^{\mathrm{T}}+\mu\boldsymbol{D}_w\boldsymbol{F}]_{ij}} \tag{10}$$

4　基于自身资源的特征扩展

4.1　特征扩展

利用前文得到的特征空间 $\boldsymbol{H}_{p\times k}$（含 p 个特征词）计算最相关特征，据此对短文本进行扩展。首先，对短文本提取 q（$p\gg q$）个特征 $f_i(i=1,\cdots,q)$，并在特征空间中找到一个子集 $\boldsymbol{H}_{q\times k}^*$ 包含且仅包含这些特征。然后，将 $\boldsymbol{H}_{q\times k}^*$ 与特征空间 $\boldsymbol{H}_{p\times k}$ 相乘，得到矩阵 $\boldsymbol{E}_{q\times p}$，如式(11)所示：

$$\boldsymbol{E}=\boldsymbol{H}^*\boldsymbol{H}^{\mathrm{T}} \tag{11}$$

其中，矩阵 $\boldsymbol{E}$ 描述的是 $f_i(i=1,\cdots,q)$ 与空间中所有特征的相关性。

为了便于选择扩展特征，再对矩阵 $\boldsymbol{E}$ 进行压缩，将每列数值累加后求均值，得到 p 维的向量 $\boldsymbol{e}$，如式(12)所示：

$$\boldsymbol{e}(j)=\frac{\sum_{i=1}^{q}\boldsymbol{E}_{ij}}{q},j=1,\cdots,p \tag{12}$$

向量 $\boldsymbol{e}$ 描述的是特征空间中的每个特征词与以 $f_i(i=1,\cdots,q)$ 为代表的短文本的相关性。除已有的文本特征外，根据 $\boldsymbol{e}$ 中的相关性大小，选择前 K 个特征对短文本进行扩展。

4.2　特征空间的更新

在对短文本特征进行扩展的过程中，存在一种可能：短文本经过预处理后，提取出来的特征有一部分不包含在特征空间中。此时，仅通过初始训练集构造的特征空间成为了特征扩展不充分的证据。因此，在对短文本进行特征扩展之前，首先检测文本特征中是否含有待更新特征。待更新特征须满足以下两个条件：

(1)该特征不存在于特征空间；

(2)该特征不是词聚类指示矩阵经过降维处理后被删除的特征。

此时，文本特征被分成两部分：1)q 个特征，能在特征空间中找到一个子集 $\boldsymbol{H}_{q\times k}^*$ 包含且仅包含这些特征；2)a 个待更新特征。由于这些特征出现在同一个短文本中，特征间也必然存在关联性，因此待更新特征的聚类指示矩阵 $\boldsymbol{H}_{q\times k}^{**}$ 是根据 $\boldsymbol{H}_{q\times k}^*$ 计算得到的。每个待更新特征的聚类指示计算方式如式(13)所示：

$$\boldsymbol{H}_i^{**}(j)=\frac{\sum_{g=1}^{q}\boldsymbol{H}_{gj}^*}{q},j=1\cdots k,i=1,\cdots,a \tag{13}$$

最后，把 $\boldsymbol{H}^{**}$ 并入 $\boldsymbol{H}$ 中，从而得到更新后的特征空间，并将其作为特征扩展的依据。

4.3　算法描述

短文本特征扩展算法 NMFFE 如算法 1 所示。

算法 1　短文本特征扩展算法 NMFFE

输入：短文本训练集 $T_1=\{t_1,\cdots,t_m\}$，短文本测试集 $T_2=\{t_1,\cdots,t_g\}$，单词集 $W=\{w_1,\cdots,w_n\}$，特征扩展个数 K，类别数 k

输出：T_2 的文本特征向量 $\mathbf{v}(t_i)$

```
for each t_i of T_1 and w_j of W
    构造关系矩阵 R，并且根据式(1)构造关联矩阵 A_t 和 A_w
endfor
根据式(7)计算 S
根据式(10)计算 F
通过降维处理得到 H
for each t_i={f_1,…,f_{q+a}} of T_2
    获取 t_i 的特征空间子集 H*
    if a≠0
        for each f_b(b=1,…,a)
            根据式(13)得 H**，更新 H
    endif
    初始化 v(t_i)
    根据式(11)计算 E
    根据式(12)计算 e
    for each d={1,…,K}
        选择 e 中值最大的特征 f_c
        if f_c ∉ t_i
            把特征 f_c 扩展至文本 t_i 中
            d++
        endif
    endfor
    for each f_d ∈ t_i
```

24. 特征 f_d 在 $\mathbf{v}(t_i)$ 中的对应位置设为 1
25. endfor
26. endfor
27. 返回 $\mathbf{v}(t_i)$

5 实验及分析

本文实验在 Eclipse,Python 和 Matlab 下实现,硬件平台的基本配置为运行内存 6GB,i5 型 CPU 处理器。

5.1 数据集

本文将在 3 个数据集上验证所提方法的有效性,实验中使用开源工具 libsvm[1)] 作为短文本分类器。第一个数据集为 Web snippets,由 Phan 等[14]从网页搜索中获取,是一个常用的短文本分类测试集。其中包含 8 个类别,有 10 060 个训练集和 2 280 个测试集,平均文本长度为 17.93 个词,具体信息如表 1 所列。

表 1 Web snippets 数据集

Table 1 Web snippets dataset

类别	训练集	测试集
Business	1 200	300
Computers	1 200	300
Culture-Arts-Entertainment	1 880	330
Education-Science	2 360	300
Engineering	220	150
Health	880	300
Politics-Society	1 200	300
Sports	1 120	300

第二个数据集是由 Hu 等[15]公开的 Twitter100k 数据集,文本由用户以非正式语言编写,并且受字数限制。但是,由于此数据集没有类标签,因此只从中抽取与体育相关的数据进行手动标记后作为实验数据,并按照体育项目分类,为实验验证做参考。为了使训练数据充足,剔除了文本数量少于 400 的体育项目,最终剩余 6 个项目,其中包含 3 000 个训练集和 630 个测试集,平均文本长度为 12.95 个词,具体信息如表 2 所列。

表 2 Twitter sports 数据集

Table 2 Twitter sports dataset

类别	训练集	测试集
棒球	500	100
篮球	500	100
足球	400	80
高尔夫	400	50
橄榄球	800	200
游泳	400	100

第三个数据集是由 Zhang 等[10]在 Web2 上获取的 AG 新闻数据,本文选择了其中数据最大的 4 个类构建数据集,包含 120 000 个训练集和 7 600 个测试集,平均文本长度为 38.82 个词,具体信息如表 3 所列。

表 3 AGnews 数据集

Table 3 AGnews dataset

类别	训练集	测试集
World	30 000	1 900
Sports	30 000	1 900
Business	30 000	1 900
Sci/Tech	30 000	1 900

1) https://github.com/cjlin1/libsvm

5.2 参数选择

对于式(2)中正则化参数 μ 和 φ 的选择,将采用常见的 Purity[16],NMI[17],ARI[18] 这 3 个评估指标作为度量标准。Purity 计算正确聚类的文档数占总文档数的比例,取值范围为 0～1;NMI 度量两个聚类结果的相近程度,取值范围为 −1～1;ARI 衡量聚类结果与真实情况的吻合程度,取值范围为 0～1。在关系矩阵分解的过程中,正则化参数设置为 $\mu=\varphi$,取值范围为 0～1。利用稀疏的测试数据,对不同 μ 值的 DNMTF 方法进行随机初始化的 50 次重复实验。图 2 给出了不同正则化参数 μ 的对比结果。

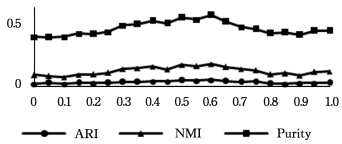
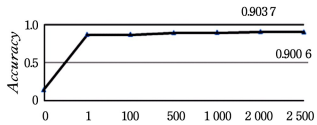


图 2 不同正则化参数 μ 的影响

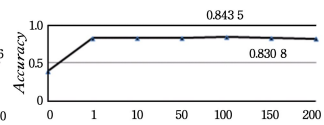
Fig. 2 Effect of different regularization parameter μ

综合 3 个评估指标,Purity,NMI,ARI 越大,聚类结果就越好,因此后续的实验中矩阵分解的正则化参数 μ 设置为 0.6。

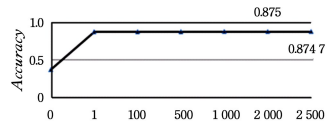
特征扩展个数 K 的选择直接影响分类结果,因此在 3 个数据集上分别选取不同的参数 K 进行对比实验,结果如图 3 所示。当只扩展一个特征时,分类结果的准确率接近于最优分类结果,这是因为在特征扩展阶段,根据式(12)在特征空间中找到与短文本相关性最强的特征,这个特征一定是某个类别中指示性最强的特征。原本稀疏的特征向量之间的相似性大大提高,从而对分类结果产生积极的影响。当扩展特征逐渐增加时,分类结果的准确率缓慢上升,在达到最优结果后,开始呈下降趋势。



(a) Web snippets 数据集



(b) Twitter sports 数据集



(c) AGnews 数据集

图 3 不同参数 K 在不同数据集上的实验结果

Fig. 3 Experimental results of different parameter K on different dataset

5.3 算法对比结果

为了验证 NMFFE 算法的效果,在不考虑语义信息的情况下,将 BOW 和 Char-CNN 即词袋法和字符级卷积神经网络方法作为对比方法,结果如表 4 所列,表中对应的最好结果分别加粗表示。在文献[10]中,BOW 算法和 Char-CNN 算法在 AGnews 数据集上的准确率分别为 88.81% 和 87.18%。由于实验环境和数据处理操作等存在差异,该结果与表 4 中的结果不一致,本文实验结果以表 4 所列为准。

表 4　各算法在不同数据集上的实验结果

Table 4　Experimental results of algorithms on different dataset

（单位：%）

算法	Web snipets	Twitter sports	AGnews
BOW	64.60	73.46	84.21
Char-CNN	62.5	52.34	85.71
NMFFE	**90.37**	**84.35**	**87.5**

从数据集规模层面进行分析，Char-CNN 算法在大规模数据集上表现良好，在小规模数据集上占据弱势。训练数据有限时，无法体现数据整体的分布，导致模型出现过拟合现象。

从信息完整性层面进行分析，AGnews 数据集文本较长，其充足的语料使得 3 种算法在文本分类上均表现良好，并且分类结果的准确率相差较小。Web snippets 数据集的测试集和训练集的相似性（关键词的共现情况）没有其他 2 个数据集高，使得基于词频统计的 BOW 算法在这个数据集上的分类结果较差。

从算法整体表现层面进行分析，NMFFE 算法在 3 个数据集上的分类结果优于其他两种算法，并且在处理不同规模的数据集时的鲁棒性都优于后者。BOW 算法和 Char-CNN 算法更适用于大规模数据集。

为了从多方面比较算法的性能，分别在 3 个数据集上对 3 种算法的运行时间进行对比，结果如图 4 所示。BOW 算法的运行时间短于其他两种算法，在大规模数据集上的速度优势更加明显，这主要是因为 BOW 算法的模型比较简单。NMFFE 算法在特征扩展过程中耗时最长，因为其中涉及到大量的矩阵运算。Char-CNN 算法的模型包含 6 个卷积层和 3 个全连接层，大量的计算使其运行时间最长。

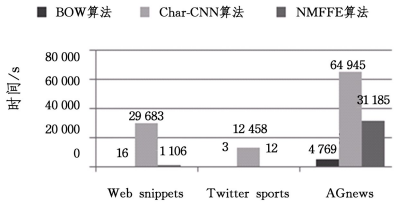


图 4　各算法运行时间的对比

Fig. 4　Comparison of running time of algorithms

结束语　针对短文本特征稀疏的问题，提出了一种基于非负矩阵分解的特征扩展方法。在不考虑外部资源的情况下，通过非负矩阵分解从文本数据集中获取词聚类指示矩阵，对其进行降维处理后，获取特征空间作为特征扩展的依据，以解决短文本特征稀疏的问题，进而提高短文本分类的准确性。实验结果表明，NMFFE 算法在短文本分类上的表现优于其他两种算法。

NMFFE 算法考虑了单词之间的相关性，下一步将在语义层面挖掘文本中更有价值的信息。

参 考 文 献

[1] TOMMASEL A，GODOY D. Short-text feature construction and selection in social media data：a survey[J]. Artificial Intelligence Review，2018，49(3)：301-338.

[2] BOLLEGALA D，MATSUO Y，ISHIZUKA M. A Web Search Engine-Based Approach to Measure Semantic Similarity between Words[J]. IEEE Transactions on Knowledge and Data Engineering，2011，23(7)：977-990.

[3] LI X，SU Y，MA H，et al. Combining Statistical Information and Semantic Similarity for Short Text Feature Extension[C]// International Conference on Intelligent Information Processing. Springer，2016：205-210.

[4] LI J，CAI Y，CAI Z，et al. Wikipedia Based Short Text Classification Method[M]// Database Systems for Advanced Applications. Berlin：Springer，2017：275-286.

[5] LI P，HE L，WANG H，et al. Learning From Short Text Streams With Topic Drifts[J]. IEEE Transactions on Cybernetics，2017，48(9)：1-15.

[6] VO D T，OCK C Y. Learning to classify short text from scientific documents using topic models with various types of knowledge[J]. Expert Systems with Applications，2015，42(3)：1684-1698.

[7] ZHANG H，ZHONG G. Improving short text classification by learning vector representations of both words and hidden topics [J]. Knowledge-Based Systems，2016，102(C)：76-86.

[8] KIM K，CHUNG B S，CHOI Y R，et al. Language independent semantic kernels for short-text classification[J]. Expert Systems with Applications，2014，41(2)：735-743.

[9] ZHANG X，ZHAO J，LECUN Y. Character-level convolutional networks for text classification[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. ACM，2015，1：649-657.

[10] DING C H Q，LI T，PENG W，et al. Orthogonal nonnegative matrix t-factorizations for clustering[C]// Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2006.

[11] GU Q，ZHOU J. Co-clustering on manifolds[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2009：359-368.

[12] SHANG F，JIAO L C，WANG F. Graph dual regularization non-negative matrix factorization for co-clustering[J]. Pattern Recognition，2012，45(6)：2237-2250.

[13] BOYD S，VANDENBERGHE L. Convex Optimization[M]. Cambridge：Cambridge University Press，2004.

[14] PHAN X H，NGUYEN L M，HORIGUCHI S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]// Proceeding of the 17th International Conference on World Wide Web. Beijing：ACM，2008：91-100.

[15] HU Y，ZHENG L，YANG Y，et al. Twitter100k：A Real-world Dataset for Weakly Supervised Cross-Media Retrieval[J]. IEEE Transactions on Multimedia，2018，20(4)：927-938.

[16] ZHAO Y，KARYPIS G. Criterion functions for document clustering[C]// Proceedings of the Thirteenth ACM Conference on Information and knowledge Management. ACM，2005：1-30.

[17] STREHL A，GHOSH J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research，2003，3(3)：583-617.

[18] HUBERT L，ARABIE P. Comparing Partitions[J]. Journal of Classification，1985，2(1)：193-218.