# 熵权约束稀疏表示的短文本分类算法

脱 婷[1],马慧芳[1,2,3],李志欣[3],赵卫中[4]
(1. 西北师范大学计算机科学与工程学院,甘肃兰州 730070; 2. 桂林电子科技大学广西可信软件重点实验室,广西桂林 541004;
3. 广西师范大学广西多源信息挖掘与安全重点实验室,广西桂林 541004; 4. 华中师范大学计算机学院,湖北武汉 430079)

**摘 要:** 针对短文本特征稀疏性问题,提出一种熵权约束稀疏表示的短文本分类方法. 考虑到初始字典维数较高,首先,利用 Word2vec 工具将字典中的词表示成词向量形式,然后根据加权向量平均值对原始字典进行降维. 其次,利用一种快速特征子集选择算法去除字典中不相关和冗余短文本,得到过滤后的字典. 再次,基于稀疏表示理论在过滤后的字典上,为目标函数设计一种熵权约束的稀疏表示方法,引入拉格朗日乘数法求得目标函数的最优值,从而得到每个类的子空间. 最后,在学习到的子空间下通过计算待分类短文本与每个类中短文本的距离,并根据三种分类规则对短文本进行分类. 在真实数据集上的大量实验结果表明,本文提出的方法能够有效缓解短文本特征稀疏问题且优于现有短文本分类方法.

**关键词:** 短文本分类; 词向量; 熵; 稀疏表示
**中图分类号:** TP393.092 **文献标识码:** A **文章编号:** 0372-2112 (2020)11-2131-07
**电子学报 URL:** http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.11.006

# Effectively Classifying Short Texts by Entropy Weighted Constraints Sparse Representation

TUO Ting[1], MA Hui-fang[1,2,3], LI Zhi-xin[3], ZHAO Wei-zhong[4]
(1. *College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu* 730070, *China*;
2. *Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi* 541004, *China*;
3. *Guangxi Key Lab of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin, Guangxi* 541004, *China*;
4. *School of Computer Central China Normal University, Wuhan, Hubei* 430079, *China*)

**Abstract:** Aiming at the problem of short text feature sparsity, a short text sparse representation classification method based on entropy weighted constraint is proposed. Considering that the initial dictionary dimension is high, firstly, the word in the dictionary is represented as a word vector form via using the Word2vec tool, and then the original dictionary is reduced according to the average weighted vectors. Secondly, a fast feature subset selection algorithm is adopted to remove the irrelevant and redundant short texts in the dictionary, and the filtered dictionary can then be obtained. Thirdly, based on the sparse representation theory, an improved entropy-weighted sparse representation method is designed for the objective function, and the Lagrange multiplier method is introduced to obtain the optimal value of the objective function, thus the subspace of each class is obtained. Finally, the distance between the short text to be classified and the short text in each class is calculated under the subspace, and the short text is classified according to three classification rules. A large number of experimental results on real data sets show that the proposed method can effectively alleviate the short text feature sparse problem and exhibits better performance than the existing short text classification methods.

**Key words:** short text classification; word embedding; entropy; sparse representation

收稿日期:2018-07-30;修回日期:2020-07-01;责任编辑:张龑翔
基金项目:国家自然科学基金(No. 61762078, No. 61363058, No. 61663004, No. 61966004, No. 61762079);广西可信软件重点实验室研究课(No. kx202003);广西多源信息挖掘与安全重点实验室开放基金(No. MIMS18-08);西北师范大学 2019 年度青年教师科研能力提升计划(No. NWNU-LKQN2019-2)

## 1 引言

随着互联网的快速发展,如新闻或文章标题,产品评论以及微博等短文本信息开始充斥着人们的日常生活.对这些短文本进行有效的分类是网页搜索、商品推荐、情感分析等众多领域中的一项关键技术.然而由于短文本长度短、样本特征稀疏[1]以及实时性强等特点,使得传统的文本分类算法不能很好适用于短文本.

目前,解决短文本分类问题主要是通过外部知识丰富短文本的特征空间或利用附加信息.前者可以分为两类:基于网页搜索的方法和基于分类/主题的方法;后者主要使用短文本中存在的链接信息.基于网页搜索的方法利用相关语境丰富了短文本,短文本被视为查询并提交到搜索引擎,返回的结果被集成到训练和测试短文本中.已有研究[2]表明,基于网页搜索的方法可以明显提高分类性能.然而,这些方法的性能很大程度上取决于所涉及的搜索引擎.基于分类/主题的方法使用显式分类法或隐式主题.例如,张等人[3]使用 LDA(Latent Dirichlet Allocation)模型和各种外部知识库丰富短文本.在文献[4]中标题视为超边,标题中不同词项视为超点来构建超图,并对超图中超边与超点同时加权.近年来,随着神经网络的兴起,Mikolov 等人[5]提出词向量的概念,即每个词是一个稠密、低维的实数向量,其中每一维表示词语的一个潜在特征.在此基础上,Boom 等人[6]提出一种基于语义词嵌入和频率信息的方法对短文本进行低维表示.本质上,这些方法都是通过扩展短文本或利用附加信息来规避稀疏问题,使传统文本分类方法适用于短文本.

与上述观点不同,Gao 等人[7]提出一种结构稀疏表示的短文本分类器,但却忽略了在高维数据中,被聚类的对象不是针对整个空间且这些子空间的重要性不同的事实[8],例如在分类短文本时,不同主题的短文本应该在某些潜在的子空间下比较相似,而不是在整个空间中.

基于以上考虑,本文提出一种熵权约束稀疏表示的短文本分类方法(effectively classifying short texts by Entropy Weighted Constraints Sparse Representation, EWC-SR).首先字典中的词表示成词向量形式,根据加权向量平均值对原始字典进行降维.其次,利用一种快速特征子集选择算法对字典进行过滤.再次,基于稀疏表示,为目标函数设计一种熵权约束稀疏表示方法,引入拉格朗日乘数法求得目标函数最优值,从而得到每个类的子空间.最后,在子空间下根据三种分类规则对短文本进行分类.在真实数据集上的大量实验结果表明,本文的方法能够有效缓解短文本特征稀疏问题且优于现有短文本分类方法.

## 2 字典构建

利用稀疏表示的思想对短文本进行分类时,第一步就是需要构建字典.然而,有研究[7]已经证明稀疏表示可能会受到字典中数据关联的严重影响,特别是当字典维度较大时.为了克服这个问题,本文首先对原始的字典进行有效的降维,然后对降维后的字典进行去不相关和去冗余等操作进行过滤,最后利用过滤后的字典对短文本进行稀疏表示分类.

### 2.1 短文本向量化

词向量是自然语言处理中一组语言建模和特征学习技术的总称,词汇中的单词或短语被映射到实数向量上.给定 $n$ 个短文本 $D=\{d_1,d_2,\cdots,d_n\}$ 被划为 $c$ 个类 $C=\{1,2,\cdots,l,\cdots,c\}$,词项集 $T=\{t_1,t_2,\cdots,t_m\}$,$n$ 个短文本被转化成 $m$ 维的特征向量,作为训练样本用来构成原始字典 $\boldsymbol{A}^{m\times n}$.为解决原始字典 $\boldsymbol{A}$ 高维稀疏问题,本文利用 Word2vec 工具将字典中的词表示为其词向量形式 $\boldsymbol{Word2vec}(\boldsymbol{t}_i)$,其中词向量维度为 $k$,这里将 $k$ 固定为 300.目前 Word2vec 工具包含了两种训练模型,本文使用 CBOW 模型.然后,对字典中任意短文本 $d_i$ 利用该文本中词向量乘以其权重进行加权求平均得到短文本表示 $\boldsymbol{V}(d_i)$,从而对字典中所有短文本进行表示.短文本表示公式如下:

$$\boldsymbol{V}(d_i)=\frac{1}{|d_i|}\sum_{j=1}^{|d_i|}w_j\times\boldsymbol{Word2vec}(\boldsymbol{t}_j) \tag{1}$$

其中,$\boldsymbol{V}(d_i)$ 是短文本 $d_i$ 的向量表示,$|d_i|$ 是短文本 $\boldsymbol{d}_i$ 中的词的个数,$w_j$ 是 $d_i$ 中第 $j$ 个词的 $idf$ 权重,$\boldsymbol{Word2vec}(\boldsymbol{t}_j)$ 是 $d_i$ 中第 $j$ 个词的词向量.

由于 $\boldsymbol{Word2vec}(\boldsymbol{t}_i)$ 的维度为 $k$,对 $\boldsymbol{A}$ 中的所有短文本基于式(1)进行表示后,每个短文本被表示为 $1\times k$ 维向量,降维后的字典 $\boldsymbol{A}'=[\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_n]^{\mathrm{T}}\in\mathbf{R}^{k\times n}$.

### 2.2 字典过滤

字典中同类训练样本间存在相关性和冗余性,此外,不同类别的样本也存在相关性,因此有必要对字典进行过滤.字典过滤主要包括不相关短文本去除和冗余短文本消除两部分.给定第 $l$ 类中短文本数为 $n_l$,对于去不相关短文本,首先利用 Word2vec 工具得到每个类别的特征向量 $\boldsymbol{Word2vec}(l)$,然后计算 $l$ 类中所有短文本向量 $\boldsymbol{d}_i$ 与类别向量之间的相似度 $S(\boldsymbol{d}_i,l)$,相似度计算公式如下:

$$S(\boldsymbol{d}_i,l)=\exp(-\|\boldsymbol{V}(d_i),\boldsymbol{Word2vec}(l)\|_2),d_i\notin l \tag{2}$$

最终将每个类中与该类别向量间相似度值小于阈值 $\theta$(经过大量实验证明,当 $\theta$ 取 0.2 时,最终得到的短文本分类效果最好)的短文本认为是该类的不相关短文本,删除这些短文本,得到该类的相关短文本 $l'=\{d_1,d_2,$

$\cdots, d_{n_{l'}}\}$.

对于每个类中冗余短文本的消除涉及三个步骤:

(1)利用该类中所有相关短文本构建无向完全图 $G=(V,E)$,其中 $V=\{d_i \mid d_i \in l' \wedge i \in [1,n_{l'}]\}$, $E=\{(d_i,d_j) \mid d_i,d_j \in l' \wedge i,j \in [1,n_{l'}] \wedge i \neq j\}$,边的权重为相关短文本间相似度值.

(2)从 $G$ 中构建最小生成树 MST(Minimum Spanning Tree);本文利用 Prim 算法构造图 $G$ 的最小生成树 MST.

(3)划分 MST,选择最具代表性短文本.

完成最小生成树后,移除权重值小于端点与其类别相似度值所在的边,得到划分后的森林. 森林中每棵树为一个簇,文献[9]中已证明每个簇都是冗余的,因此,从每个簇中选出与该类别向量相似度最高的短文本作为最终的子集,即得到去冗余后字典 $\boldsymbol{A}^* \subset \boldsymbol{A}'$.

## 3　熵权约束稀疏表示的短文本分类

首先,给出基于熵权约束的稀疏表示的目标函数用以学习每个类的字空间;接着,在子空间上利用三种分类规则对短文本进行分类.

### 3.1　熵权约束的稀疏表示

根据稀疏表示思想,给定 $n$ 个短文本用来构成字典 $\boldsymbol{A}$,对于待分类短文本 $y$,试图从 $\boldsymbol{A}$ 上通过求解方程 $y=\boldsymbol{\beta A}=\boldsymbol{\beta}_1\boldsymbol{A}_1+\boldsymbol{\beta}_2\boldsymbol{A}_2+\ldots+\boldsymbol{\beta}_n\boldsymbol{A}_n$ 重建 $y$,其中 $\boldsymbol{\beta}=[\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_n] \in \mathbf{R}^n$ 是系数向量. 本节首先利用训练样本学出每个类的子空间,然后计算将待分类短文本与该类所在子空间下短文本的距离,从而根据分类规则对短文本进行分类. 由于每个类中短文本已知,不是动态变化的,本文的方法不需要迭代计算计算子空间. 以 $l$ 类为例,子空间学习的目标函数如下:

$$F(\boldsymbol{\beta}_l)=\sum_{i=1}^{n_l}\sum_{j=i+1}^{n_l}\sum_{t=1}^{k}\beta_{lt}(x_{it}-x_{jt})^2+\gamma\sum_{t=1}^{k}\beta_{lt}\log_2\beta_{lt}$$
$$\sum_{t=1}^{k}\beta_{lt}=1 \tag{3}$$

其中,$\boldsymbol{\beta}_l=[\beta_{l1},\beta_{l2},\cdots,\beta_{lk}]^{\mathrm{T}} \in \mathbf{R}^n$ 是系数向量. $n_{l'}$ 是 $l$ 类中短本文数,$x_{it}$ 是第 $i$ 个短文本的第 $t$ 维,$k$ 是短文本向量的维数,$\gamma$ 是一个正调节因子,用于控制子空间聚类激励强度. 式(3)的第一部分是类内距离,第二项是负权值熵. 最小化目标函数就是既要最小化类内距离,又要最小化负熵权值. 换句话说,最小化类内距离意味着要使类内短文本彼此相似,距离最小;最小化负熵权值意味需要刺激更多的维度,从而避免用稀疏数据通过几个维度识别短文本的问题.

通过 $F$ 最小化,求得:

$$\beta_{lt}=\frac{\exp\left(\frac{-D_{lt}}{\gamma}\right)}{\sum_{t=1}^{k}\exp\left(\frac{-D_{lt}}{\gamma}\right)} \tag{4}$$

其中,$D_{lt}=\sum_{i=1}^{n_{l'}}\sum_{j=i+1}^{n_{l'}}(x_{it}-x_{jt})^2$.

证明过程详见附录.

### 3.2　分类规则

通过前面的计算已经得到每个类的子空间,本节在学习到的子空间上计算待分类短文本与每个类的距离,并定义三种分类规则对短文本进行分类.

**规则 1(最小距离)**　给定短文本 $Y$,将其表示成其对应的向量形式 $\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_k]$,基于最小距离的短文本 $Y$ 的标签应该:

$$label(Y)=\min_l\left\{l \;\middle|\; \sum_{i=1}^{n_{l'}}\sum_{t=1}^{k}\beta_{lt}(x_{it}-y_t)^2,\ l=1,2,\cdots,c\right\} \tag{5}$$

**规则 2(平均距离)**　给定短文本 $Y$,基于平均距离短文本 $Y$ 的标签应该是:

$$label(Y)=\min_l\left\{l \;\middle|\; \frac{1}{n_l}\times\sum_{i=1}^{n_{l'}}\sum_{t=1}^{k}\beta_{lt}(x_{it}-y_t)^2,\ l=1,2,\cdots,c\right\} \tag{6}$$

**规则 3(中心距离)**　首先计算每个类的类中心,然后在该类的子空间上计算短文本 $Y$ 与类中心的距离从而对短文本 $Y$ 进行分类.

假定 $\boldsymbol{Z}_l$ 是 $l$ 类的类中心向量,其中 $\boldsymbol{Z}_l=[z_{l1},z_{l2},\cdots,z_{lt},\cdots,z_{lk}]$,$z_{lt}$ 的计算公式如下:

$$z_{lt}=\frac{1}{n_{l'}}\times\sum_{i=1}^{n_{l'}}x_{it} \tag{7}$$

因此,基于中心距离的短文本 $Y$ 的标签是:

$$label(Y)=\min_l\left\{l \;\middle|\; \sum_{t=1}^{k}\beta_{lt}(z_{lt}-y_t)^2,\ l=1,2,\cdots,c\right\} \tag{8}$$

由式(1)可知字典构建的时间复杂度为 $O(mn)$. 在字典过滤阶段,其复杂度已被证明当字典数量较大时远小于 $O((n_l)^2)$,本文需要遍历 $c$ 个类,所以时间复杂度小于 $O(c\times(n_l)^2)$. 由式(3)可知,对于 $c$ 个类子空间计算的时间复杂度为 $O(c\times k\times n_{l'}^2)$. 在分类过程中,根据三种分类规则可以得到时间复杂度为 $O(c\times k\times n_{l'})$. 因此,本文方法的时间复杂度为 $O(mn+c\times(n_l)^2+(c\times k\times n_{l'}^2)+O(c\times k\times n_{l'}))$,其中 $c$ 和 $k$ 为定值且取值较小,故本文算法的运行效率较高.

## 4　实验结果与分析

为验证本文方法的分类结果,本节首先对实验数

据和评价指标进行描述,其次设计两组实验对所提出的方法进行验证,并对实验结果进行分析.

## 4.1 数据描述及评价指标

### 4.1.1 数据描述

本节在五个数据集上对本文方法的分类结果进行验证. 20 新闻组语料库①中每个类选取 1000 篇新闻标题. 中文新闻数据集②包含 10 个类. 同样,只使用新闻标题,并进行中文分词. JSC 语料库③是由 1002 条合法邮件和 322 条垃圾邮件构成. IMDB 大型电影评论数据集④是由电影评论及其积极或消极的情绪组成的. 最后一个数据集是 Twitter⑤ 情感分析语料库. 实验所有训练集与测试集采用五折交叉验证进行选取. 数据统计结果如表 1 所示,每个文档的平均关键字数分别是 5.4,10,19.5,7.56 和 17.35,因此这些文本实际上都是短文本.

**表 1 五种数据集的统计数据**

| 数据集 | 类别数 | 文本总数 | 关键词 | 平均关键词/文本 |
|---|---|---|---|---|
| 20 新闻组 | 20 | 20000 | 8866 | 5.4 |
| 中文新闻 | 10 | 2814 | 7437 | 10 |
| Twitter | 2 | 157826 | 20135 | 19.5 |
| JSC 语料 | 2 | 1324 | 6375 | 7.56 |
| IMDB | 2 | 50000 | 11035 | 17.35 |

### 4.1.2 评价指标

常用的短文本分类性能有不同的评价指标,例如,准确率 Accuracy、精确率 P、召回率 R 和 F1-measure 值. F1-measure 越与 1 靠近说明 P 和 R 的平衡性越好,分类性能越好. 在实验中,首先计算每个类别的准确度、精确率及召回率值,然后对所有类别的宏观平均值进行评估. 本文仅介绍准确率 Accuracy 和 F1-measure 的实验结果(简称 F1).

## 4.2 实验结果与分析

本节在五种数据集上设计了实验进行验证. 首先,对实验中的参数 γ 进行分析;其次,将本文的 EWC-SR 方法与 SSR-DF(effectively classifying short texts by Structured Sparse Representation with Dictionary Filtering)方法进行分析比较;最后,实验比较了本文方法与现有的短文本分类方法,通过准确率 Accuracy 和 F1 值验证本文所提出算法的有效性.

### 4.2.1 参数分析

在第三节中,由式(2)可以知道 γ 是一个正调节因子,用于控制子空间聚类激励强度. 本小节将通过大量实验进行验证,以确定最优的实验参数.

综合图 1 和图 2 可以得出,在所有数据集上 γ 在[0.3,7]范围内获得了较高的 Accuracy 值和 F1 值,表明分类结果对 Accuracy 值和 F1 值的变化不敏感,同时也证明该算法具有良好的鲁棒性. 因此实验中将参数 γ 设为 0.5.

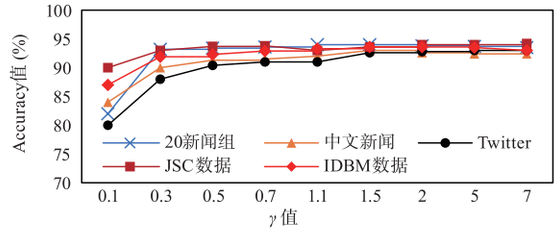



图1 五种数据集上子空间聚类的Accuracy值

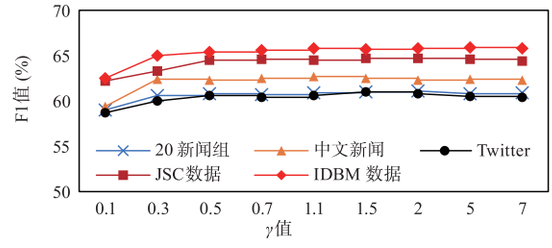



图2 五种数据集上子空间聚类的F1值

### 4.2.2 与 SSR-DF 方法性能比较

本节将 EWC-SR 方法与 SSR-DF 方法进行比较的主要原因有:

(1)两种方法均是基于稀疏表示理论,因此与 SSR-DF 方法进行比较是合理.

(2)两种方法均对稀疏表示的字典进行改进,且设计不同的分类规则.

本文方法与 SSR-DF 类似,但是本文方法引入了熵权约束对目标函数进行改进,可以避免用稀疏数据通过几个维度识别短文本的问题,且不需要迭代计算每个类的子空间. 实验结果如表 2 所示.

如表 3 所示,五种数据集上本文方法在 Accuracy、F1 以及运行时间上明显优于 SSR-DF 方法,尤其是在运行时间上有很大改进. 其主要原因是本文 EWC-SR 方法考虑到短文本应该在某个特定的子空间比较相似,且 EWC-SR 方法在计算每个类子空间时不需要迭代计算,大大节省时间开销. 此外,虽然在 JSC 数据集上 EWC-SR 方法的 Accuracy 和 F1 值略微低于 SSR-DF 方法,可能是受到了数据集本身的限制,JSC 数据集是一个情感分类数据集,即使在这种情况下 EWC-SR 方法的 F1 值也同样达到了 65.51%,且运行时间远远低于 SSR-DF 方法.

① http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html
② http://www.people.com.cn/
③ http://www.esp.uem.es/jmgomez/smsspamcorpus/
④ http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz
⑤ http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz

表 2　字典过滤的影响

| 数据集 | | SSR-DF | | EWC-SR | |
|---|---|---|---|---|---|
| | | 字典过滤 | 未过滤 | 字典过滤 | 未过滤 |
| 20 新闻组 | Accuracy | 94.24 | 93.64 | **95.03** | 94.87 |
| | F1 | 54.76 | 51.32 | **58.68** | 56.39 |
| | 时间(s) | 22409 | 67628 | **5742** | 11064 |
| 中文新闻 | Accuracy | 93.87 | 92.88 | **94.53** | 93.96 |
| | F1 | 61.77 | 58.94 | **62.87** | 61.77 |
| | 时间(s) | 1826 | 8623 | **658** | 1013 |
| Twitter | Accuracy | 92.37 | 92.04 | **93.14** | 93.06 |
| | F1 | 55.32 | 52.13 | **56.43** | 55.47 |
| | 时间(s) | 57716 | 120864 | **16304** | 29863 |
| JSC | Accuracy | **93.56** | 93.17 | 93.42 | 93.26 |
| | F1 | **65.93** | 63.07 | 65.51 | 63.31 |
| | 时间(s) | 769 | 3890 | **392** | 937 |
| IMDB | Accuracy | 93.76 | 93.16 | **94.43** | 94.18 |
| | F1 | 68.33 | 62.59 | **72.17** | 71.53 |
| | 时间(s) | 40619 | 17980 | **8176** | 16407 |

### 4.2.3　与现存方法性能比较

为验证本文方法的分类性能,将本文方法与经典的短文本分类方法 SVM(Support Vector Machine)和 KNN(k-Nearest Neighbor),三种稀疏表示分类方法 SRC(text classification using hierarchical Sparse Representation Classifiers)[10] 和 SR-SVM(text classification using combined Sparse Representation classifiers and Support Vector Machines)[11] 以及 SSR-DF 对比. 对于 SVM 使用 libSVM 来实现,KNN 中,利用 Python 中的 s-klearn 包解决参数设置问题. SRC 方法中用加权分解主成分分析来区分相似的类. SR-SVM 方法使用与训练文本对应词频向量表示的主成分来创建字典. 实验选取 20 新闻组及中文新闻数据集进行比较.

观察图3和图4,发现两种数据集上EWC-SR方法的 Accuracy 和 F1 值均优于其他五种方法. 尽管在两种数据集上所有方法的 Accuracy 值差别不是很大,但就 F1 值而言,SVM 和 KNN 方法的 F1 值均低于 50%. 其中,SVM 方法的性能并不理想,这是因为在维数稀疏空间中,距离测度不能很好地区分具有不同非重叠非零维数的向量,这使得 SVM 无法完美地选择支持向量. SRC 和 SRC-SVM 方法虽然利用了稀疏表示的思想,但还是利用原始的词频. SSR-DF 方法分类性能得到很大提升,但仍然低于 60%,原因是这两种方法虽然利用了稀疏表示的思想,却忽略了潜在的子空间. 本文 EWC-SR 方法的 F1 值均达到 60% 以上,且本文方法在计算子空间时不需要进行迭代,在运行时间上远远低于 SSR-DF 方法,因此本文方法能更好的对短文本进行分类.

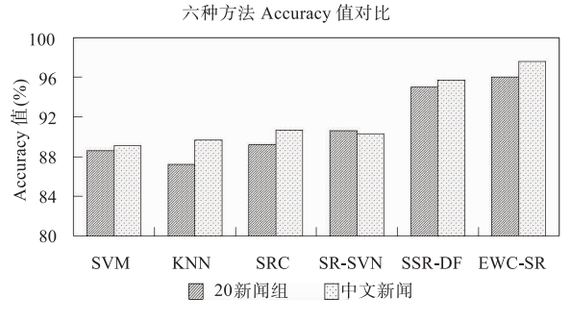



图3　两种数据集上六种方法Accuracy值对比

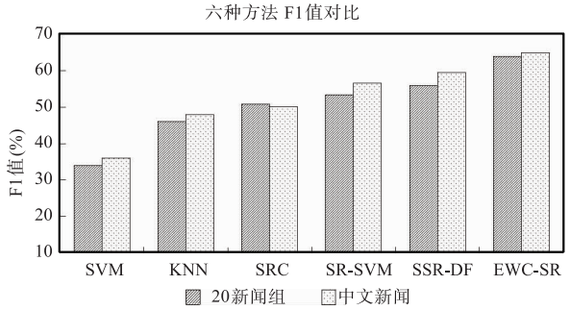



图4　两种数据集上六种方法F1值对比

表 3　分类规则的影响

| 数据 | 评价指标 | SSR-DF | | | EWC-SR | | |
|---|---|---|---|---|---|---|---|
| | | 规则 1 | 规则 2 | 规则 3 | 规则 1 | 规则 2 | 规则 3 |
| 20 新闻组 | Accuracy | 93.24 | **93.96** | 93.77 | **95.03** | 94.31 | 94.96 |
| | F1 | 51.76 | **54.76** | 53.88 | **60.68** | 59.61 | 60.47 |
| 中文新闻 | Accuracy | 92.87 | **93.99** | 92.82 | 94.53 | 94.62 | **94.67** |
| | F1 | 54.75 | 54.50 | **55.77** | 65.13 | 63.83 | **65.87** |
| Twitter | Accuracy | 86.87 | 89.83 | **92.76** | 94.14 | **94.28** | 94.13 |
| | F1 | 56.22 | 56.81 | **57.62** | 66.35 | **66.39** | 66.34 |
| JSC | Accuracy | **91.68** | 91.34 | 90.68 | 94.01 | 94.17 | **94.42** |
| | F1 | **53.73** | 52.43 | 51.89 | **67.87** | 64.87 | 67.51 |
| IMDB | Accuracy | 90.76 | 90.68 | **91.69** | **94.47** | 94.38 | 94.12 |
| | F1 | 54.33 | 53.33 | **57.86** | **72.17** | 69.71 | 70.97 |

## 5 结束语

与通过扩展短文本或利用附加信息来规避短文本稀疏性问题的分类方法不同,本文提出的熵权约束稀疏表示的短文本分类方法考虑到短文本分类时应该在潜在的子空间上相似的事实,为目标函数设计熵权约束的稀疏表示方法,从而求得每个类的子空间,在子空间上设计三种分类规则对短文本进行分类. 最终实验结果表明,本文的方法能够有效缓解短文本特征稀疏问题且优于现有短文本分类方法.

## 附录

在式(3)中,极小化 $F$ 形成了一类约束非线性优化问题,其解是未知的,因此本文引入拉格朗日乘数法得到以下无约束极小化问题:

$$\min F_l(\boldsymbol{\beta}_l,\delta_l) = \sum_{i=1}^{n_r}\sum_{j=i+1}^{n_r}\sum_{t=1}^{k}\beta_{lt}(x_{it}-x_{jt})^2 + \gamma\sum_{t=1}^{k}\beta_{lt}\log_2\beta_{lt} - \delta_l\left(\sum_{t=1}^{k}\beta_{lt}-1\right)$$

通过将 $F_l(\boldsymbol{\beta}_l,\boldsymbol{\delta}_l)$ 相对于 $\beta_{lt}$ 和 $\delta_l$ 的梯度设置为零,得到:

$$\frac{\partial F_1}{\partial\delta_l} = \left(\sum_{t=1}^{k}\beta_{lt}-1\right) = 0$$

$$\frac{\partial F_l}{\partial\beta_{l\ t}} = \sum_{i=1}^{n_r}\sum_{j=i+1}^{n_r}(x_{it}-x_{jt})^2 + \gamma(1+\log_2\beta_{lt}) - \delta_l = 0$$

令 $D_{lt} = \sum_{i=1}^{n_l}\sum_{j=i+1}^{n_l}(x_{it}-x_{jt})^2$,可得:

$$\beta_{lt} = \exp\left(\frac{-D_{lt}-\gamma+\delta_l}{\gamma}\right) = \exp\left(\frac{\delta_l-\gamma}{\gamma}\right)\cdot\exp\left(\frac{-D_{lt}}{\gamma}\right)$$

已知 $\sum_{t=1}^{k}\beta_{lt} = 1$,可得:

$$\sum_{t=1}^{k}\beta_{lt} = \sum_{t=1}^{k}\exp\left(\frac{\delta_l-\gamma}{\gamma}\right)\cdot\exp\left(\frac{-D_{lt}}{\gamma}\right) = \exp\left(\frac{\delta_l-\gamma}{\gamma}\right)\cdot\sum_{t=1}^{k}\exp\left(\frac{-D_{lt}}{\gamma}\right) = 1$$

因此,$\exp\left(\frac{\delta_l-\gamma}{\gamma}\right) = \frac{1}{\sum_{t=1}^{k}\exp\left(\frac{-D_{lt}}{\gamma}\right)}$

最终整理可得:

$$\beta_{lt} = \frac{\exp\left(\frac{-D_{lt}}{\gamma}\right)}{\sum_{t=1}^{k}\exp\left(\frac{-D_{lt}}{\gamma}\right)}$$

### 参考文献


[1] Wang J, Wang Z, Zhang D, et al. Combining knowledge with deep convolutional neural networks for short text classification[A]. Proceedings of the 26th International Joint Conference on Artificial Intelligence[C]. New York: ACM, 2017. 2915 – 2921.

[2] Li C, Wang H, Zhang Z, et al. Enhancing topic modeling for short texts with auxiliary word embeddings[J]. ACM Transactions on Information Systems, 2017, 36(2): 11:1 – 11:30.

[3] 张雄,陈福才,黄瑞阳. 基于双词主题模型的半监督实体消歧方法研究[J]. 电子学报,2018,46(3):607 – 613.
ZHANG Xiong, CHEN Fu-cai, HUANG Rui-yang. Semi-supervised entity disambiguation method research based on biterm topic model[J]. Acta Electronica Sinica, 2018, 46(3): 607 – 613. (in Chinese)

[4] 马慧芳,刘芳,夏琴,等. 基于加权超图随机游走的文献关键词提取算法[J]. 电子学报,2018,46(6):1410 – 1414.
MA Hui-fang, LIU Fang, XIA Qin, et al. Keywords extraction algorithm based on weighted hypergraph random walk[J]. Acta Electronica Sinica, 2018, 46(6): 1410 – 1414. (in Chinese)

[5] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[A]. Proceedings of Advances in Neural Information Processing Systems 27th Annual Conference on Neural Information Processing Systems[C]. Cambridge: MIT Press, 2013. 3111 – 3119.

[6] Boom C, Canneyt S, Demeester T, et al. Representation learning for very short texts using weighted word embedding aggregation[J]. Pattern Recognition Letters, 2016, 80(C): 150 – 156.

[7] Gao L, Zhou S, Guan J. Effectively classifying short texts by structured sparse representation with dictionary filtering[J]. Information Sciences, 2015, 323: 130 – 142.

[8] Jing L, Ng M K, Huang J Z. An entropy weighting k-Means algorithm for subspace clustering of high-dimensional sparse data[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(8): 1026 – 1041.

[9] Song Q, Ni J, Wang G. A fast clustering-based feature subset selection algorithm for high-dimensional data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 1 – 14.

[10] Sharma N, Dileep A D, Thenkanidiyoor V. Text classification using hierarchical sparse representation classifiers[A]. Proceedings of the 16th IEEE International Conference on Machine Learning and Applications[C]. Piscataway: IEEE, 2017. 1015 – 1019.

[11] Sharma N, Sharma A, Thenkanidiyoor V, et al. Text classification using combined sparse representation classifiers

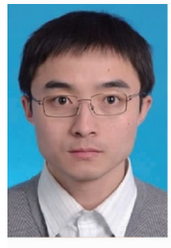
and support vector machines[A]. International Symposium on Computational and Business Intelligence[C]. Piscataway:IEEE,2016. 181 – 185.

## 作者简介

**脱　婷**　女,1990 年 9 月出生,甘肃庆阳人. 自 2016 年进入西北师范大学计算机科学与工程学院学习,现为硕士研究生,主要从事自然语言处理与分类算法方面研究.
E-mail:nwnutuot@ yeah. net

**马慧芳(通信作者)**　女,1981 年 7 月出生,甘肃兰州人. 博士,硕士生导师,现为西北师范大学计算机科学与工程学院教授,主要从事机器学习与数据挖掘等方面研究工作.
E-mail:mahuifang@ yeah. net

**李志欣**　男,1971 年 10 月出生,广西桂林人. 博士,博士生导师. 现为广西师范大学计算机科学与信息工程学院教授,主要从事图像理解与机器学习等方面的研究.
E-mail:lizx@ gxnu. edu. cn



**赵卫中**　男,1981 年 10 月出生,山东菏泽人. 博士,硕士生导师,现为华中师范大学计算机学院副教授,主要从事机器学习与数据挖掘等方面研究工作.
E-mail:zhaoweizhong@ gmail. com