[文章编号]1000-1832(2018)02-0058-08
[DOI]10.16163/j.cnki.22-1123/n.2018.02.011

一种基于 SV-NN 的哈萨克语文本分类方法

古丽娜孜·艾力木江[1,2],乎西旦·居马洪[1],孙铁利[3],梁义[1]
(1.伊犁师范学院电子与信息工程学院,新疆 伊宁 835000;
2.东北师范大学地理科学学院,吉林 长春 130024;
3.东北师范大学信息科学与技术学院,吉林 长春 130117)

[摘　要] 根据哈萨克语语法规则设计实现哈萨克语文本的词干提取,完成哈萨克语文本的预处理.提出基于最近支持向量机的样本距离公式,结合 SVM 与 KNN 分类算法实现了哈萨克语文本的分类.结合构建的哈萨克语文本语料库的语料进行文本分类仿真实验,结果表明所提出的算法是有效的.

[关键词] 词干提取;支持向量机;文本分类;分类精度
[中图分类号] TP 391.1　　[学科代码] 520·10　　[文献标志码] A

随着企业与数字图书馆的快速增长,文本分类已成为文本数据组织与处理的关键技术.文本分类(Text Classification,TC)是基于机器学习的学习任务[1],它是信息检索技术非常活跃的研究领域.TC 的任务是为一个文档自动分配一组预定义的类别或应用主题.数字化数据有不同的形式,它可以是文字、图像、空间形式等,其中最常见和应用最多的是文本数据,我们阅读的新闻、社交媒体上的帖子和信息主要以文本形式出现.文本自动分类在网站分类[2-3]、自动索引[4-5]、电子邮件过滤[6]、垃圾邮件过滤[7-9]、本体匹配[10]、超文本分类[11-12]和情感分析[13-14]等许多信息检索应用中起到了重要的作用.数字化时代,在线文本文档及其类别的数量越来越巨大,而文本分类是从数据海洋中挖掘出具有参考价值数据的应用程序.[15-16]文本挖掘工作是许多应用领域里书面文本的分析过程,朴素贝叶斯、K 紧邻、支持向量机、决策树、最大熵和神经网络等基于统计与监督的模式分类算法在文本分类研究中已被广泛应用.针对迅速发展的 Web 数据的开发应用,提高文本分类效率的算法研究具有重要意义.

一般来说,合理的词干有助于提高文本分类的性能和效率[17-18],特别是对像哈萨克语这样构词和词性变化较复杂语言的文本分类而言词干的准确提取极其重要.由于从同一个词干可以派生许多单词,因此通过词干提取还可以对语料库规模进行降维.文本文档数量的巨大化和包含特征的多样化给文本挖掘工作带来一定的困难.目前,众多文本分类研究都是基于英文或中文,而基于少数民族语言为基础的文本分类研究相对较少.[19]然而国外阿拉伯语的文本分类工作相对于中国少数民族语言文本分类较成熟.[20-21]

哈萨克语言属于阿尔泰语系突厥语族的克普恰克语支,中国境内通用的哈萨克文借用了阿拉伯语和部分波斯文字母,而哈萨克斯坦等国家用的哈萨克文是斯拉夫文字.哈萨克文本跟中文不同的是哈萨克文文本单词以空格分开的,而这点类似于英文,但由于两种语言语法体系不一样,英文词干提取规则不能直接用到哈萨克语文本分类问题上,需要研究适合哈萨克语语法体系的词干提取规则之后才能实

[收稿日期] 2017-09-24
[基金项目] 国家自然科学基金资助项目(61663045);新疆高校科研计划项目(XJEDU2014I043);伊犁师范学院重点项目(2016YSZD04);伊犁师范学院基金资助项目(2016WXYB0004).
[作者简介] 古丽娜孜·艾力木江(1972—),女,博士,副教授,主要从事模式识别、文本分类、遥感图像处理研究.

现哈萨克语文本的分类工作. 哈萨克语具有丰富的形态和复杂的拼字法，所以实现哈萨克语文本分类系统并不是一件容易的事. 为了实现文本分类任务需要一定规模的语料库，而语料库里语料的质量直接影响文本分类的精度. 但是，到目前为止还没有一个公认的哈萨克文语料库，也有不少人认为新疆日报（哈文版）上的文本可以当做文本分类语料. 本文为了保证文本分类语料的规范化和文本分类工作的标注化，经过认真挑选中文标准语料库里的部分语料文档并对其进行翻译和挑选新疆日报（哈文版）上的部分文档来自行搭建了本文研究的语料. 在之前研究[22-23]进行优化改善的基础上，本文给出新的样本测度指标与距离公式，并结合 SVM 与 KNN 分类算法实现了哈萨克语文本分类.

1　文本特征提取

1.1　文本预处理

文本预处理在整个文本分类工作中扮演最重要的角色，其处理程度直接影响到后期进行的文本分类精度. 因为它是从文档中抽取关键词集合的过程，而关键词的单独抽取因语言语法规则的不同而不同，所以这是属于技术含量较高的基础性工作，需要设计人员熟练掌握语言语法规则和计算机编程能力.

哈萨克语文字由 24 个辅音字母和 9 个元音字母组成. 哈萨克语文本词与词之间有空格分开，所以不需要用分词处理，但要用词干提取. 由于哈萨克语语法形式由在单词原形的前后附加一定的成分来完成，所以哈萨克语言属于黏着语，即跟英文类似，一个哈萨克语单词对应多种链接形式，因此对其一定要进行词干提取.

我们前期基本完成了哈萨克语文本词干提取以及词性标注工作，完成了哈萨克语文本词干表的构建. 该表收录了由新疆人民出版社出版的《哈萨克语详解词典》中的 6 万多个哈萨克语文本词干（见图 1）和 438 个哈萨克语文本词干附加成分（见图 2）.

id	word	pos
1	ئۇزۇشىتىشكەرە	v
2	ئۇزالجىك	adj
3	ئۇزۇ	n
4	ئۇزىمر	v
5	ئۇزىمرل	vc
6	ئۇزىمرلۇ	va

图 1　哈萨克语词干

index	type	suffix	btype
215	adj	مك	gc
201	adj	ۇلى	gc
228	adj	سىز	gc
227	adj	پە	gc
226	adj	پا	gc

图 2　哈萨克语附加成分

本文给出 3 种词性的有限状态自动机，并采用词法分析和双向全切分相结合的改进方法实现哈萨克语文本词干的提取和单词构形附加成分的细切分. 改进逐字母二分词典查询机制对词干表进行搜索，提高词干提取的效率. 以概率统计的方法对歧义词和未登陆词进行切分. 在此基础上，设计实现了哈萨克语文本的词法自动分析程序，完成哈萨克语文本的读取预处理. 处理结果如图 3 所示，上半部显示的是待切分的文档原文，下半部显示的是词干切分后的结果.

1.2　特征处理

特征就是文本分类时判别类别的尺度. 模式识别的不同分类问题有不同的特征选择方法，而在文本分类问题中常用到的方法有互信息（MI）、X^2 统计量（CHI）、信息增益（IG）、文档频率（DF）等几种.[24] 这些方法各具特色和不足. MI、IG 和 CHI 倾向于低频词的处理，而 DF 则倾向于高频词的处理. 目前，也有许多优化改进方法[25-27]，其中，文本频率比值法 DFR（Document Frequency Ratio，DFR）以简单、快捷等优点克服了以上几种方法存在的问题，综合考虑了类内外文本频率，其计算公式为

$$\mathrm{DFR}(t,C_i)=\frac{(N-n_i)\times \mathrm{DF}_i}{n_i\times \mathrm{DF}_i'}. \tag{1}$$

其中：t，N 是训练文本数；n_i 是 C_i 类别中的文本数；DF_i 是 C_i 类别中包含词 t 的文本数；DF_i' 是除了 C_i 类以外的类别中包含词 t 的文本数.

通过对词频统计、词权重计算和文档向量化表示等一系列的预处理之后才能运用分类算法，所以对于文本分类而言这些都是非常重要的阶段性基础工作. 每类文档里（如体育类文档中）每一个单词（如

"排球"词)的总出现次数见图4.词的权重计算结果见图5,即统计某词在判别文档类别所属关系中的隶属度,隶属度越高说明该词在文档分类时的贡献越大.最后对文档进行形式向量化表示(见图6),生成分类问题的文档向量,即"XX号特征词:该特征词的特征向量"形式向量化表示.

图3 哈萨克语文本词干切分结果示例

图4 词频统计结果

图5 词权重计算结果

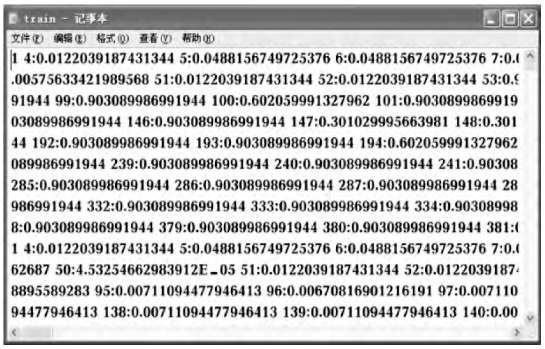


图6 文本向量文件

2 SVM与KNN方法

2.1 SVM方法

SVM是由C. Cortes等[28]在1995年首次提出来的一种模式识别分类技术.它是在统计学习理论(Statistical Leaning Theory,SLT)原理的基础上发展起来的机器学习算法.SVM方法的重点是在高维特征空间中构造函数集VC维尽可能小的最优分类面,使不同类别样本通过超平面在分类风险上界最小化,从而保证分类算法的最优推广能力.在有限训练样本情况下,SVM在学习机复杂度和学习机泛化能力之间找到一个平衡点[29],从而保证学习机的推广能力.

SVM方法模型见图7,图7b是线性可分的,图7c是线性不可分的,即根据样本分布情况与样本集维数,SVM分类算法的判别函数原理大致可由图7(b,c)2种形式表示.

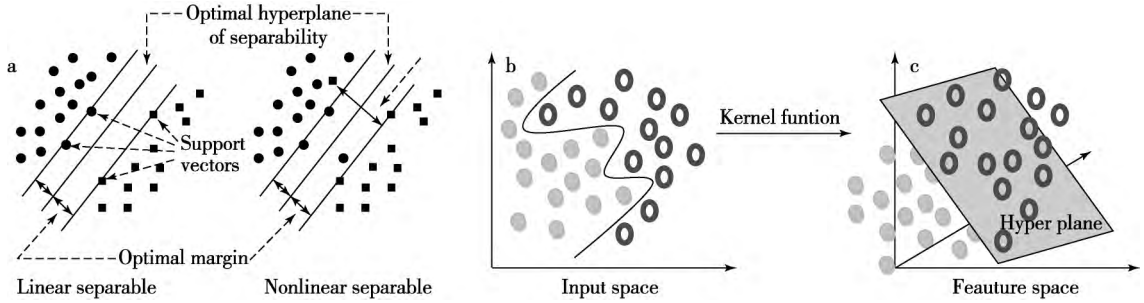


图7 SVM分类原理示意图

2.1.1 线性可分

训练样本集的SVM线性可分分类问题的数学模型为

$$S=\{(x_i\cdot y_i),i=1,2,\cdots,r\},x_i\in\mathbf{R}^n,y_i\in\{+1,-1\}. \tag{2}$$

(2)式还可表达为

$$\min\varphi(\boldsymbol{\omega})=\frac{1}{2}\|\boldsymbol{\omega}\|^2,\ \text{s.t.}\ \ y_i[\boldsymbol{\omega}x_i+b]-1\geqslant 0,i=1,2,\cdots,n. \tag{3}$$

假如,对 n 维空间中的分类界面为 $\boldsymbol{\omega}\cdot x+b=0$,使得与此分类界面最近的两类样本之间的距离 $\text{Margin}=\frac{2}{\|\boldsymbol{\omega}\|}$ 最大,即 $\|\boldsymbol{\omega}\|$ 为最小,则该分类界面就称为最优分类界面;$\boldsymbol{\omega}$ 为权重向量(是 $f(x)$ 的法向量),b 为函数阈值. 最终可得到所求的最优分类函数为

$$f(x)=\text{sign}\Big(\sum_{i=1}^{n}\boldsymbol{a}_i y_i(x_i\cdot x)+b\Big). \tag{4}$$

其中对应 $\boldsymbol{a}_i\neq 0$ 时的样本点就是支持向量. 因为最优化问题解 $\boldsymbol{a}_i$ 的每一个分量都与一个训练点相对应,显然所求得的划分超平面,仅仅与对应 $\boldsymbol{a}_i\neq 0$ 时的训练点$(x_i\cdot x)$相关,而跟 $\boldsymbol{a}_i=0$ 时的训练点无关. 相应于 $\boldsymbol{a}_i\neq 0$ 时的训练点$(x_i\cdot x)$输入点 x_i 就是支持向量,通常是全体样本中的很少一部分. 最终分类分界面的法向量 $\boldsymbol{\omega}$ 只受支持向量的影响,与非支持向量训练点的无关.

2.1.2　非线性可分

SVM 通过运用合适的非线性映射,如 $\varphi:x_i\rightarrow\varphi(x_i)$ 把分类问题原训练样本点转变(映射)到新特征空间中,使得原样本在这新特征空间(目标高维空间)中能够线性可分,然后利用线性可分问题求出最终的最优分类超面.

为此,需要在(3)式中增加一个松弛变量 ξ_i 和惩罚因子 C,从而(3)式变为

$$\min\varphi(\boldsymbol{\omega},\xi)=\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{n}\xi_i; \tag{4}$$

$$\text{s.t.}\ \ y_i[\boldsymbol{\omega}x_i+b]-1+\xi_i\geqslant 0,\xi_i\geqslant 0,i=1,2,\cdots,n. \tag{5}$$

其中 C 为控制样本对错分程度的调整因子,通常称为惩罚因子. C 越大,惩罚越重.

分类问题的训练样本不充足或不能保证训练样本质量情形下,确定非线性映射是很困难的,SVM 通过运用核函数概念解决这些困难.

SVM 通过引入一个核函数 $K(x_i,x)$,将原低维的分类问题空间映射到高维的新问题空间中,让核函数代替 $\boldsymbol{\omega}\cdot\varphi(x)$ 内积运算,这个高维的新问题空间就称 Hilbert 空间. 引入核函数以后的最优分类函数为

$$f(x)=\text{sign}\Big(\sum_{i=1}^{n}a_i y_i K(x_i\cdot x)+b\Big). \tag{6}$$

2.2　KNN 方法

KNN(K Nearest Neighbor,KNN)分类法是基于实例的学习算法,它需要所有的训练样本都参与分类.[30]在分类阶段,利用欧氏距离公式,将每个测试样本与和邻近的 k 个训练样本进行比较,然后将测试样本归属到票数最多的那一类里.[31]KNN 方法是根据测试样本最近的 k 个样本点的类别信息来对该测试样本类型进行判别,所以 k 值的选定非常重要. k 值太小,测试样本特征不能充分体现;k 值太大,与测试样本并不相似的个别样本也可能被包含进来,这样反而对分类不利. 在分类决策上只依据最邻近的 k 个样本的类别来决定待分样本的所属类. 目前,对于 k 值的选取还没有一个全局最优的筛选方法,这也是 KNN 方法的弊端,具体操作时,只好根据先验知识先给出一个初始值,然后根据仿真分类实验结果重新调整,而重复调整 k 值的操作一直到进行到分类结果满足用户需求为止. 该方法原理可表示为

$$y(d_i)=\underset{k}{\text{argmax}}\sum_{x_j\in\text{KNN}}y(x_j,c_k). \tag{7}$$

(7)式表明将测试样本 d_i 划入到 k 个邻近类别中成员最多的那个类里.

在使用 KNN 算法时,还可利用其他策略生成测试样本的归属类,其函数为

$$y(d_i)=\underset{k}{\text{argmax}}\sum_{x_j\in\text{KNN}}\text{Sim}(d_i,x_j)y(x_j,c_k). \tag{8}$$

其中:d_i 是测试样本,而 x_j 是 k 个最近邻之一;$y(x_j,c_k)\in\{0,1\}$ 表明 x_j 是否属于 c_k 类,即当 $x_j\in c_i$ 时,

$y(x_j,c_i)=1$,当 $x_j\notin c_i$ 时,$y(x_j,c_i)=0$;$\mathrm{Sim}(d_i,x_j)$是测试样本 d_i 和它最近邻 x_j 之间的余弦相似度.余弦相似度测量是由一个向量空间中2个向量之间夹角余弦值来定义的.(8)式说明测试样本 d_i 被归属到 k 个最近邻类里相似性最大的那个类里.

一般情况下,不同类别训练样本的分布是不均匀的,同样不同类别的样本个数也可能不一样.所以,在分类任务中,KNN 中 k 值可能会导致不同类别之间的偏差.例如,对于(7)式,一个较大的 k 值使得方法过拟合,反过来一个较小的 k 值使得方法模型不稳定.实际上,k 值通常由交叉验证技术来获取.然而,像在线分类等某些情况下,不能用交叉验证技术,只能给出经验值,总之 k 值的选定很重要.

KNN 虽然是简单有效的分类方法,但不能忽略以下两方面的问题:一方面,由于 KNN 需要保留分类过程中的所有相似性计算实例,随着训练集规模的增多,方法计算量也会增长,在处理较大规模数据集的分类时方法的时间复杂度会达到不可接受的程度[32],这也是 KNN 方法的一个很大缺点;另一方面,KNN 方法分类的准确性可能受到训练数据集中特性的无关性和噪声数据的影响,若考虑这些因素分类效果也许更好.

3 基于 SV-NN 的哈萨克语文本分类算法

3.1 SV-NN 算法描述

假设共有 n 个类,每个类别含有 m 个支持向量.

训练集:$\boldsymbol{T}_1=\{x_1,x_2,\cdots,x_t\}$;

测试集:$\boldsymbol{T}_2=\{x_1,x_2,\cdots,x_l\}$,且 $\boldsymbol{T}=\boldsymbol{T}_1\cup\boldsymbol{T}_2$.

SV-NN 分类算法描述:

Start:

{integeri,j,k,l;

$i=1;j=1;k=1;//i=1,2,\cdots,n;j=1,2,\cdots,m$;

SVM:$\boldsymbol{T}_1\rightarrow \boldsymbol{sv}_{ij}$;//通过使用 SVM 定义每个类的支持向量.

while($k<l$)

{ 输入 x_k;

计算 x_k 与 $\boldsymbol{sv}_{ij}$ 之间的距离(D_k);

计算 x_k 与 $\boldsymbol{sv}_{ij}$ 之间的平均距离($\mathrm{aver}D_k$);

计算 x_k 与 $\boldsymbol{sv}_{ij}$ 之间的最小平均距离($\min\limits_k(\mathrm{aver}D_k)$);

将 x_k 划入到基于$\min\limits_k(\mathrm{aver}D_k)$的最近类别;

$k=k+1$;

}

}

End.

3.2 SV-NN 算法实现

步骤1:将所有训练点映射到向量空间,并通过传统 SVM 确定每一个类别的支持向量.

$$\begin{pmatrix}\boldsymbol{sv}_{11} & \cdots & \boldsymbol{sv}_{1m}\\ \vdots & \ddots & \vdots\\ \boldsymbol{sv}_{n1} & \cdots & \boldsymbol{sv}_{nm}\end{pmatrix},i=1,2,\cdots,n,j=1,2,\cdots,m. \tag{9}$$

其中支持向量 $\boldsymbol{sv}_{ij}$ 是从输入文档中提取的(共有 n 个类,每个类别含有 m 个支持向量).确定每一类的支持向量 $\boldsymbol{sv}_s$ 之后,其余的训练点可以消除.

步骤2:使用欧氏距离公式

$$D_{kj}=\sum_{k=1}^{l}\sqrt{\Big(\sum_{i=1}^{n}\Big(x_k-\sum_{j=1}^{m}\boldsymbol{sv}_{ij}\Big)^2\Big)},i=1,2,\cdots,n;j=1,2,\cdots,m;k=1,2,\cdots,l \tag{10}$$

计算测试样本 x_k 与由步骤1生成的每一类支持向量 $\boldsymbol{sv}_{ij}$ 之间的距离.

步骤 3：计算测试样本 x_k 与每一类支持向量 sv_{ij} 之间的平均距离，公式为

$$\mathrm{aver}D_k = \frac{\sum_{j=1}^{m} D_{kj}}{m}, j = 1,2,\cdots,m; k = 1,2,\cdots,l. \tag{11}$$

步骤 4：计算最短平均距离 $\min D$，并将测试样本 x_k 划入到最短平均距离对应的一类中，公式为

$$\min D = \min_k(\mathrm{aver}D_k), k = 1,2,\cdots,l. \tag{12}$$

即输入点被确认为输入点与 sv_{ij} 之间最短平均距离值对应的正确类.

重复步骤 2～4，直到所有的测试样本分类完为止.

4　实验结果与评价

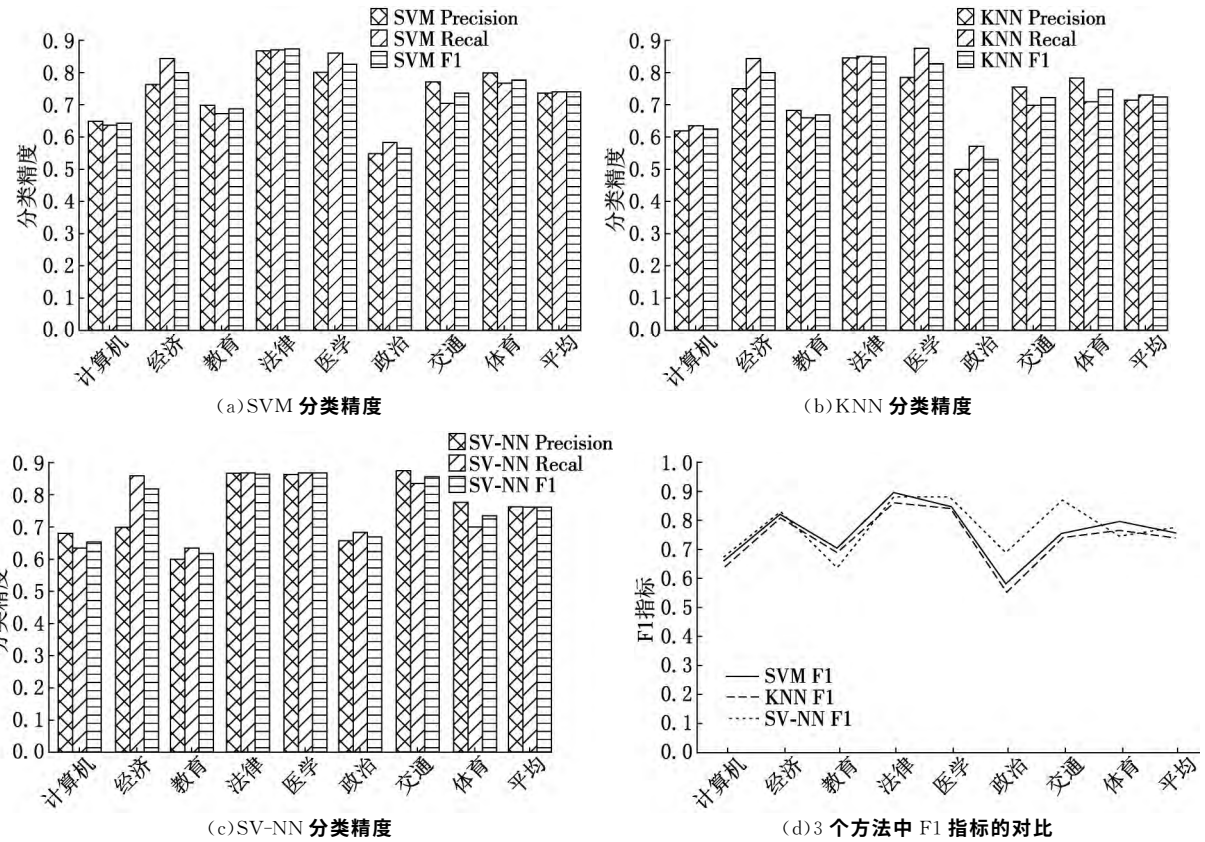
通常语料库里语料的质量与数量直接影响文本分类算法的分类性能. 本文考虑到文本分类工作的规范性和语料的标准性，由中文标准语料库里的部分文档的翻译和新疆日报（哈文版）上的部分文档的筛选搭建了本文研究的语料库. 这次是对前期语料集的补充和优化完善. 原来的语料集语料文档只有 5 类文档，这次扩充到 8 类文档. 通过语言学专家们的多次沟通，选择了具有代表性的文档，同时对词干提取程序解析规则上也做了些适当的调整. 对于本文研究所构建语料库上还不能用得上“标准”这词语，但现阶段对哈萨克语文本分类任务的完成具有实际应用价值.

本文在前期系列研究的基础上，把以前的语料集规模扩大到由计算机、经济、教育、法律、医学、政治、交通、体育等 8 类共 1 400 个哈萨克语文档组成的小型语料数据集（见表 1）. 数据集被分为 2 个部分. 880 个文档（63%）用于训练数，520 个文档用于测试（37%）.

表 1　8 类小型语料数据集

类别	文档总数	训练文档数	测试文档数
计算机	175	110	65
经济	175	110	65
教育	175	110	65
法律	175	110	65
医学	175	110	65
政治	175	110	65
交通	175	110	65
体育	175	110	65
总计	1 400	880	520

本文文本分类实验评价指标采用了分类精度、召回率和 F13 种评价方法. 期望获得较高的分类精度和召回率. 在前期系列研究中所搭建的哈萨克文语料集的补充完善和词干提取程序提取规则细节的优化改善基础上实现了哈萨克语文本的分类. 运用 SVM、KNN 与本文提出的 SV-NN 算法，并对 3 种算法分类精度进行了较全面的对比分析，分析结果见图 8. 通过对图 8 的仿真实验数字的对比分析，发现 SVM 算法优于 KNN 算法，而 SV-NN 算法优于 SVM 算法. SV-NN 方法 F1 指标除了教育类和法律类以外在其他类上的 F1 指标都高于 SVM、KNN. SVM、KNN 和 SV-NN 平均分类精度分别为 0.754，0.731 和 0.778，说明本文提出的算法对所有类别文档词的召回率和区分度较稳定. 在有限样本情况下，该算法模型已继承 SVM 算法，获得较好分类精度，而且没有定义 KNN 算法的 k 参数，也没有跟所有类全部训练样本进行距离运算. 所以，本文提出的算法无论从算法复杂度的分析还是算法收敛速度的分析都是有效的. 当然，总体精度没有中英文等其他语言文本分类精度高，但是目前获得的分类精度比较理想，本文算法的文本分类性能和召回率有了很大的提升，对于影响分类精度的以上几方面的问题将继续研究，并努力争取得到满意的分类精度.



(a)SVM 分类精度 (b)KNN 分类精度

(c)SV-NN 分类精度 (d)3 个方法中 F1 指标的对比

图 8 SVM、KNN、SV-NN 分类精度及 F1 指标的对比

5 结论

本文在前期系列研究的基础上实现了哈萨克语文本的分类. 运用了模式识别的 3 种分类算法,并对 3 种算法分类精度进行了较全面的对比分析. 通过仿真实验,证明本文提出方法具有一定的优越性. 本文算法对所有类别文档词的召回率和区分度较稳定. 不需要设置 k 参数,保证了分类算法的收敛速度,获得了较高的分类精度和召回率.

[参 考 文 献]

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

[2] AHMADI A,FOTOUHI M,KHALEGHI M. Intelligent classification of web pages using contextual and visual features[J]. Applied Soft Computing,2011,11(2):1638-1647.

[3] MARTINEZ CAMARA E,MARTIN VALDIVIA MT,URENA LOPEZ LA,et al. Polarity classification for Spanish tweets using the COST corpus[J]. Journal of Information Science,2015,41(3):263-272.

[4] PERCANELLA G,SORRENTINO D,VENTO M. Automatic indexing of news videos through text classification techniques [C]// Proceedings of the 3rd International Conference on Pattern Recognition and Image Analysis (Part II). Berlin:Springer, 2005:512-521.

[5] RONG HU,BRIAN MAC NAMEE,SARAH JANE DELANY. Active learning for text classification with reusability[J]. Expert Systems With Applications,2016,45(3):438-449.

[6] SAKURAI S,SUYAMA A. An e-mail analysis method based on text mining techniques[J]. Applied Soft Computing,2006,6 (1):62-71.

[7] ALKABI M,WAHSHEH H,ALSMADI I,etal. Content-based analysis to detect Arabic web spam[J]. Journal of Information Science,2012,38(3):284-296.

[8] ADEL HAMDAN,RAED ABUZITAR. Spam detection using assisted artificial immune system[J]. International Journal of Pattern Recognition and Artificial Intelligence,2011,25(8):1275-1295.

[9] RAED ABUZITAR,ADEL HAMDAN. Application of genetic optimized artificial immune system and neural networks in spam detection[J]. Applied Soft Computing,2011,11(4):3827-3845.

[10] MING M,YEFEI P,MICHAEL S. Ontology mapping:as a binary classification problem[J]. Concurrency and Computation: Practice and Experience,2011,23(9):1010-1025.

[11] YANG Y，SLATTERY S，GHANI R. A study of approaches to hypertextcategorization[J]. Journal of Intelligent Information Systems，2002，18(2/3)：219-241.
[12] REN FUJI，LI CHAO. Hybrid Chinese text classification approach using general knowledge from Baidu Baike[J]. IEEJ Transaction on Electrical and Electronic Engineering，2016，11(4)：488-498.
[13] DUWAIRI R，ELORFALI M. A study of the effects of preprocessing strategies on sentiment analysis for Arabic text[J]. Journal of Information Science，2014，40(4)：501-513.
[14] 张冬梅. 文本情感分类及观点摘要关键问题研究[D]. 济南：山东大学，2012.
[15] 杨杰明. 文本分类中文本表示模型和特征选择算法研究[D]. 长春：吉林大学，2013.
[16] CNNIC. 第 37 次中国互联网络发展状况统计报告[R]. 北京：中国互联网络信息中心，2016.
[17] SYIAM MM，FAYED ZT，HABIB MB. An intelligent system for Arabic text categorization[J]. Journal of Intelligent Computing and Information Sciences，2006，6(1)：1-19.
[18] DUWAIRI R，ALREFAI M，KHASAWNEH N. Stemming versus light stemming as feature selection techniques for Arabic text categorization[J]. International Conference on Innovations in Information Technology，2008，25(9)：446-450.
[19] HE HUI，WANG JUNYI. Study of active learning support vector machine and its application on mongolian text classification [J]. Acta Sientiarum Naturalium Universitatis NeiMongol，2006，37(5)：560-563.
[20] ABDULLAHI O ADELEKE，NOOR A SAMSUDIN，AIDA MUSTAPHA，et al. Comparative analysis of text classification algorithms forautomated labelling of quranicverses[J]. International Journal on Advanced Science Engineering Information Technology，2017，7(4)：119-1427.
[21] ADEL HAMDAN MOHAMMAD，TARIQ ALWADA'N，OMAR AL MOMANI. Arabic text categorization using supportvector machine，naïve bayes and neural network[J]. GSTF Journal on Computing (JOC)，2016，5(1)：108-115.
[22] GULINAZI，SUN TIE LI，YILIYAER，et al. Research into text categorization of kazakh based on support vector machine[J]. CAAI Transaction on Intelligent Systems，2011，6(3)：261-267.
[23] GULNAZ，SUN TIE LI，YILIYAR. Text categorization of Kazakh text based on SVM-modified KNN[J]. Journal of Northwest Normal University，2014，50(5)：48-53.
[24] 旺建华. 中文文本分类技术研究[D]. 长春：吉林大学，2007.
[25] JOACHIMS T. Text categorization with support vector machines：Learning with many relevant features[C]//In Proceedings of The 10th European Conference on Machine Learning(ECML). Berlin：Springer，1998：137-142.
[26] WANG ZIQIANG，SUN XIA，ZHANG DEXIAN，et al. An optimal svm-based text classification algorithm[C]//5th International Conference on Machine Learning and Cybernetics. Dalian：IEEE，2006：13-16.
[27] MONTANES E，FERANDEZ J，DIAZ I，et al. Measures of rule quality for feature selection in text categorization[C]// 5th International Symposium on Intelligent Data Analysis. Berlin：Springer，2003：589-598.
[28] CORTES C，VAPNIK V. Support-vector networks[J]. Machine Learning，1995，20(3)：273-297.
[29] WANG XUESONG，HUANG FEI，CHENG YUHU. Computational performance optimization of support vector machine based on support vectors[J]. Neurocomputing，2016，211：66-71.
[30] COVER T M，HART P E. Nearest neighbor pattern classification[J]. IEEE Transactionson Information Theory，1967，13(1)：21-27.
[31] HASTIE T，TIBSHIRANI R，FRIEDMAN J H. The elements of statistical learning：data mining，inference and prediction[J]. Journal of the Royal Statistical Society，2009，173(3)：693-694.
[32] QING MIN MENG，CHRIS J，CIESZEWSKI，et al. Knearest neighbor method for forest inventory using remote sensing data [J]. GIS Science and Remote Sensing，2007，44(2)：149-165.

An approach based on SV-NN for Kazakh language text classification

GULNAZ Alimjan[1,2]，HURXIDA Jumahun[1]，SUN Tie-li[3]，LIANG Yi[1]

(1. School of Information Science and Technology，Yili Normal University，Yining 835000，China；
2. School of Geographical Science，Northeast Normal University，Changchun 130024，China；
3. School of Computer Science and Information Technology，Northeast Normal University，Changchun 130117，China)

Abstract：According to the rules of Kazakh grammar，the text stemming of Kazakh language was designed，and the pretreatment of Kazakh language was completed. This paper proposes a sample distance formula which is based on recent support vector. Combining SVM with KNN classification algorithm implementeds Kazakh language text classification. The text classification simulation experiment of the Kazakh language corpus has been carried out，the numerical experiments show the effectiveness of the proposed algorithm and confirm the theoretical results.

Keywords：stemming；support vector machines；text categorization；classification accuracy

（责任编辑：石绍庆）