DOI: 10.11992/tis.201711007
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.tp.20180423.1540.016.html

基于支持向量的最近邻文本分类方法

古丽娜孜·艾力木江[1,2,3], 乎西旦·居马洪[1], 孙铁利[2], 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 计算机科学与技术学院, 吉林 长春 130117; 3. 东北师范大学 地理科学学院, 吉林 长春 130024）

摘　要: 文本分类为一个文档自动分配一组预定义的类别或主题。文本分类中, 文档的表示对学习机的学习性能有很大的影响。以实现哈萨克语文本分类为目的, 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, 避免 k 参数的选定, 以 SVM 与 KNN 分类算法的特殊组合算法（SV-NN）实现了哈萨克语文本的分类。结合自己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, 数值实验展示了提出算法的有效性并证实了理论结果。

关键词: 词干提取; 预处理; 支持向量机; 文本分类; 分类精度
中图分类号: TP309　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)05−0799−09

中文引用格式: 古丽娜孜·艾力木江, 乎西旦·居马洪, 孙铁利, 等. 基于支持向量的最近邻文本分类方法[J]. 智能系统学报, 2018, 13(5): 799–807.
英文引用格式: GULNAZ Alimjan, HURXIDA Jumahun, SUN Tieli, et al. The nearest neighbor text classification method based on support vector[J]. CAAI transactions on intelligent systems, 2018, 13(5): 799–807.

The nearest neighbor text classification method based on support vector

GULNAZ Alimjan[1,2,3], HURXIDA Jumahun[1], SUN Tieli[2], LIANG Yi[1]
(1. Department of Electronics and Information Engineering, Yili Normal University, Yining 835000, China; 2. School of Information Science and Technology, Northeast Normal University, Changchun 130117, China; 3. Department of Geographical Science, Northeast Normal University, Changchun 130024, China)

Abstract: Text categorization automatically assigns a set of predefined categories or topics to a document. In text classification, the representation of the document has a great influence on the learning performance of the learning machine. The aim is to achieve Kazakh text classification, according to Kazakh grammar rules, the stemming of Kazakh texts is designed to complete the preprocessing of Kazakh text. A sample distance formula based on the latest support vector machine (SVM) is proposed to avoid the selection of k-parameters. The Kazakh texts are classified by special combination of SVM and KNN classification algorithms (SV-NN) . Combining the corpus of Kazakh text corpora constructed by himself, text categorization simulation experiments were conducted. Numerical experiments showed the effectiveness of the proposed algorithm and confirmed the theoretical results.

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

文本分类 (text classification, TC) 是对一个文档自动分配一组预定义的类别或应用主题的过程[1]。随着数字图书馆的快速增长, TC 已成为文本数据组织与处理的关键技术。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, 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

收稿日期: 2017−11−02.　**网络出版日期:** 2018−04−24.
基金项目: 伊犁师范学院一般项目 (2016WXYB0004); 国家自然科学基金项目 (61663045); 新疆高校科研计划重点研究项目 (XJEDU2014I043); 伊犁师范学院重点项目 (2016YSZD04).
通信作者: 古丽娜孜·艾力木江. E-mail: alay328@163.com.

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 TC 是从数据海洋当中挖掘出具有参考价值数据的应用程序[15-16]。文本挖掘工作是政府工作、科学研究、办公业务等许多应用领域里书面文本的分析过程。朴素贝叶斯、k 近邻、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。提高文本分类效率的算法研究对 web 数据的开发应用具有重要意义。

合理的词干有助于提高文本分类的性能和效率[17-18]，特别是对于哈萨克语这样的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。从同一个词干可以派生出许多单词，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，基于少数民族语言的文本分类研究相对较少[19]；但是国外对于阿拉伯语的文本分类工作比中国少数民族语言文本分类工作成熟[20–21]，投入研究的人员也较多。

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一点是哈萨克文文本单词以空格分开的，这点类似于英文，都需要文本词干提取过程。由于哈萨克语与英语语法体系不一样，英文词干提取规则还不能直接用到哈萨克语文本分类问题上，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，因此哈萨克语文本分类系统的实现是有难度的。为了实现文本分类任务需要一定规模的语料库，语料库里语料的质量直接影响文本分类的精度。到目前为止在哈萨克语中还没有一个公认的哈萨克文语料库，当然，也有不少人认为新疆人民日报 (哈文版) 上的文本可以当作文本分类语料库。本文为了保证文本分类语料库的规范化和文本分类工作的标准化，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新疆人民日报上的部分文档来自行搭建了本研究的语料库。本文在对前期研究里词干提取程序词干解析规则[22-24]进行优化改善的基础上实现本研究的文本预处理，提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着最重要的角色，其处理程度直接影响到文本分类精度。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，所以这层工作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。目前存在一个现实问题，即包括作者在内的很多编程人员因研究工作的需要一般从事于中英文文字资料上的研究，所以对母语 (哈萨克语) 语法规则的细节不精通，对从小开始在汉语授课学校上学的编程人员情况则更严重，所以要实现词干解析需要向语言学专家或相关人员全面请教，这也是影响哈萨克语文本分类工作进展的一个客观问题。

哈萨克语文字由 24 个辅音字母和 9 个元音字母的共有 33 个字母组成。因为哈萨克语语法形式是在单词原形前后附加一定附加成分来完成的，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。

本文前期系列研究工作基本完成了哈萨克语文本词干提取以及词性标注工作，已完成哈萨克语文本词干表的构建。该词干表收录了如图 1 所示的由新疆人民出版社出版的《哈萨克语详解词典》中的 60 000 多个哈萨克语文本词干和如图 2 所示的 438 个哈萨克语文本词干附加成分。

id	word	pos
1	ۇشۇرشىتىشكىكەرە	v
2	ۇشۇلجىك	adj
3	ۇشۇي	n
4	ۇشۇير	v
5	ۇشۇيرىل	vc
6	ۇشۇيرىلۇ	va

图 1　哈萨克语词干
Fig. 1　Kazakh text stem

index	type	suffix	btype
215	adj	دەك	gc
201	adj	ۇلى	gc
228	adj	سىز	gc
227	adj	پەز	gc
226	adj	پا	gc

图 2　哈萨克语附加成分
Fig. 2　Additional components in Kazakh text

本文在前期准备研究工作的基础上, 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与单词构形附加成分的细切分。以改进的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记载词进行切分。在此研究基础上, 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, 上半窗体上显示的是待切分的文档原文, 下半窗体上显示是词干切分后的结果。

图 3　哈萨克语文本词干切分结果示例

Fig. 3　Example segmentation results of the Kazakh text stem

1.2　特征处理

特征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)、卡方统计等[25]。这些方法各具优点和不足之处。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[26-28], 其中文本频率比值法 (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

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

式中, 对于词 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, 而DF'_i显然是除了 C_i 类以外的别的类别中包含词 t 的文本数。

通过词频统计、词权重计算和文档向量化表示等一系列的预处理工作之后才能运用分类算法, 所以对文本分类工作而言这些都是非常重要的阶段性基础工作。图 4 所示的是每类文档里 (如体育类文档中) 每一个单词 (如“排球”) 的总出现次数。图 5 所示的是词的权重计算结果, 即统计某词在判别文档类别所属关系中的隶属度, 当然隶属度越高说明该词在文档分类时的贡献越大。最后把文档由如图 6 所示的形式向量化表示, 生成分类问题的文档向量, 即“X 号特征词: 该特征词的特征向量”形式向量化表示。

图 4　词频统计结果

Fig. 4　Term frequency statistical result

图 5　词权重计算结果

Fig. 5　Term weight computed result

图 6　文本向量文件

Fig. 6　Text vector files

2　SVM 与 KNN 方法

2.1　SVM 方法

支持向量机 (support vector machine, SVM) 是

在 1995 年由 Cortes 和 Vapnik 首次提出来的一种模式识别分类技术[29]。SVM 是在统计学习理论 (statistical learning theory, SLT) 原理的基础上发展起来的机器学习算法。SVM 方法的重点在于在高维特征空间中能构造函数集 VC 维尽可能小的最优分类面，使得不同类别样本在这分类面上分类上界最小化，从而保证分类算法的最优推广能力。图 7 所示的是 SVM 方法的分类原理示意图。SVM 在有限训练样本情况下，在学习机复杂度和学习机泛化能力之间找到一个平衡点，从而保证学习机的推广能力[30]。

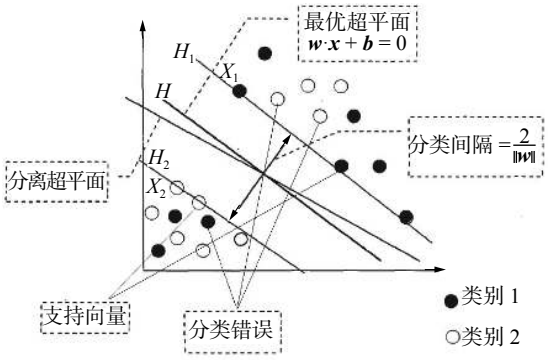


图 7　SVM 分类原理示意图

Fig. 7　SVM classification schematic diagram

根据样本分布情况与样本集维数，SVM 算法的判别函数原理大致可分为线性可分与非线性可分两种形式。

1) 线性可分

带有以式 (2) 所示训练样本集的 SVM 线性可分分类问题的数学模型可通过式 (3) 来表示：

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

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

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

$$f(x)=\text{sign}\left(\sum_{i=1}^{n}a_iy_i(x_i\cdot x)+b\right) \tag{4}$$

式中对应 $a_i\neq 0$ 时的样本点就是支持向量。因为最优化问题解 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 $a_i\neq 0$ 时的那些训练点 $(x_i\cdot x)$ 相关，而跟 $a_i=0$ 时的那些训练点没有任何关系。相应于 $a_i\neq 0$ 时的训练点 $(x_i\cdot x)$ 里的输入点 x_i 就是支持向量，通常它们是全体样本中的很少一部分。得出结论，最终分类分界面的法向量 $\boldsymbol{\omega}$ 只受支持向量的影响，不受非支持向量训练点的影响。

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) 到新的地方 (新特征空间)，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，然后在这新的映射空间中利用线性可分问题求解原理求出最终的最优分类面。

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

$$\min\varphi(\omega,\xi)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i,\ \text{s.t.}y_i[\omega x_i+b]-1+\xi_i\geqslant 0 \tag{5}$$

式中：$\xi_i\geqslant 0$；$i=1,2,\cdots,n$；C 为控制样本对错分的调整因子，通常称为惩罚因子。C 越大，惩罚越重。

虽然原理看起来简单，然而在分类问题的训练样本不充足或不能保证训练样本质量的情形下确定非线性映射是很困难的，那么如何确定非线性映射呢？SVM 通过运用核函数概念解决这个问题，核函数是 SVM 的其他分类算法无法替代的独特功能。

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

$$f(x)=\text{sign}\left(\sum_{i=1}^{n}a_iy_iK(x_i\cdot x)+b\right) \tag{6}$$

2.2　KNN 方法

KNN(K nearest neighbor) 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[31]。在分类阶段，利用欧氏距离公式，将每个测试样本与和它邻近的 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[32]。KNN 算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 值的选取非常重要。k 值若太小，测试样本特征不能充分体现；k 值若太大，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。KNN 算法在分类决策上只凭 k 个最邻近样本类别确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，根据先验知识先给出一个初始值，然后需要根据仿真分类实验结果重新调整，调整 k 值的操作有时一直到分类结果满足用户需求为止。该方法原理可由式

(7) 表示:

$$y(d_i)=\arg\max_k \sum_{x_j\in \mathrm{KNN}} y(x_j,c_k) \tag{7}$$

式 (7) 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别里。

在使用 KNN 算法时, 还可由其他策略生成测试样本的归属类, 如式 (8) 也是被广泛使用的公式:

$$y(d_i)=\arg\max_k \sum_{x_j\in \mathrm{KNN}} \mathrm{Sim}(d_i,x_j)y(x_j,c_k) \tag{8}$$

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两个向量之间角余弦值来定义的。式 (8) 说明测试样本 d_i 被归到 k 个最近邻类里相似性最大那个类别里。

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数量也可能不一样。所以, 在分类任务中, KNN 中 k 参数的一个固定值可能会导致不同类别之间的偏差。例如, 对于式 (7), 一个较大的 k 值使得方法运行结果过拟合, 反过来一个较小的 k 值使得方法模型性能不稳定。实际上, k 的值通常由交叉验证技术来获取。然而, 像在线分类等情况, 就不能用交叉验证技术, 只能给出经验值, 因此 k 值的选定很重要。

KNN 虽是简单有效的分类方法, 但不能忽略以下两方面的问题: 1) 由于 KNN 需要保留分类过程中的所有相似性计算实例, 从而随着训练集规模的增多, 其计算量也会增加, 在处理较大规模数据集分类时方法的时间复杂度会达到不可接受的程度[33], 这是 KNN 方法的主要缺点; 2) 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, 若考虑这些因素其分类效果也许更好。

3　基于 SV-NN 的哈萨克语文本分类算法

本文提出一种组合分类方法, 把 SVM 算法当作 KNN 算法的训练阶段, 这样可以避免 k 参数的选择。组合分类方法结合了 SVM 算法的训练和 KNN 算法的学习阶段。首先运用 SVM 算法对所有训练样本进行一次训练获得每一类别的少量的支持向量 (support vectors, SVs), 在测试阶段使用最近邻分类器进行测试并分类测试样本, 即计算出新测试样本与每个类别 SVs 平均距离值后对其进行对比分析, 该测试样本与哪一类别 SVs 平均距离值点离得最近就把它归为该类别中。分类决策依据是各类别 SVs 平均距离值后对其与测试点之间距离的数值分析, 所以简称该算法为支持向量与最近邻方法 (the support vector of nearest neighbor, SV-NN)。

3.1　SV-NN 算法描述及流程图

假设共有 n 个类别, 每个类别含有 m 个支持向量。

训练集: $T_1=\{x_1,x_2,\cdots,x_t\}$。

测试集: $T_2=\{x_1,x_2,\cdots,x_l\}$。

SV-NN 分类算法:

Start:

{ integer i,j,k,l;

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

SVM: $T_1\rightarrow \mathrm{sv}_{ij}$; //通过使用 SVM 定义每个类别的支持向量。

while(k<l)

{ 输入 x_k;

利用式 (9) 计算 x_k 与 sv_{ij} 之间的距离 (D_k);

利用式 (10) 计算 x_k 与 sv_{ij} 之间的平均距离 (averD_k);

利用式 (11) 计算 x_k 与 sv_{ij} 之间最小平均距离 ($\min_k(\mathrm{aver}D_k)$);

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

k=k+1;

}

}

End

SV-NN 分类方法的工作流程如图 8 所示。

3.2　SV-NN 算法实现

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。

$$\begin{pmatrix} \mathbf{sv}_{11} & \mathbf{sv}_{12} & \cdots & \mathbf{sv}_{1m} \\ \mathbf{sv}_{21} & \mathbf{sv}_{22} & \cdots & \mathbf{sv}_{2m} \\ \vdots & \vdots & & \vdots \\ \mathbf{sv}_{n1} & \mathbf{sv}_{n2} & \cdots & \mathbf{sv}_{nm} \end{pmatrix},\ i=1,2,\cdots,n;\ j=1,2,\cdots,m \tag{9}$$

式中: 支持向量 $\mathbf{sv}_{ij}$ 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之后, 其余的训练点可以消除。

2) 使用欧氏距离公式 (9) 计算测试样本 x_k 与由 1) 生成的每一类支持向量 $\mathbf{sv}_{ij}$ 之间的距离。

$$D_{kj}=\sum_{k=1}^{l}\sqrt{\left(\sum_{i=1}^{n}\left(x_k-\sum_{j=1}^{m}\mathbf{sv}_{ij}\right)\right)^2} \tag{10}$$

式中: $i=1,2,\cdots,n$; $j=1,2,\cdots,m$; $k=1,2,\cdots,l$。

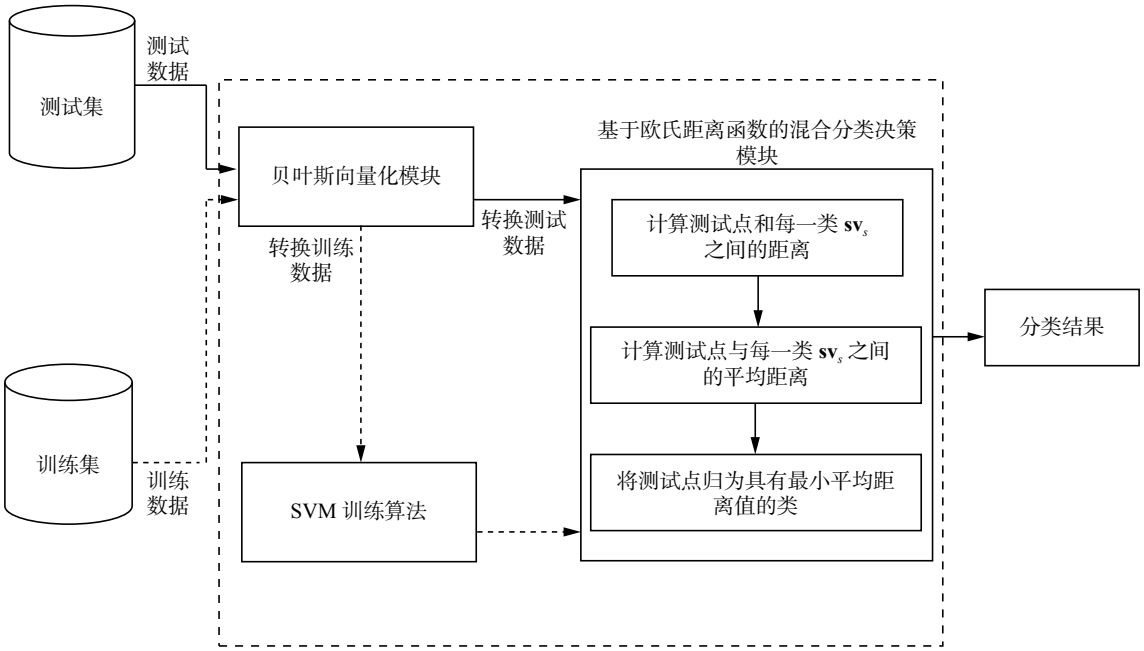


图 8　SV-NN 分类方法工作原理

Fig. 8　SV-NN classification approach working principle diagram

3) 使用式 (10) 计算测试样本 x_k 与每一类支持向量 $\mathbf{sv}_{ij}$ 之间的平均距离：

$$\text{aver}D_k = \frac{\sum_{j=1}^{m} D_{kj}}{m},\ j = 1,2,\cdots,m; k = 1,2,\cdots,l \quad (11)$$

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那一类中。

$$\min D = \min_k(\text{aver}D_k),\ k = 1,2,\cdots,l \quad (12)$$

即输入点被确认为与 $\mathbf{sv}_{ij}$ 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。中、英文等其他语言文本分类研究都有标准的语料库，而哈萨克语文本分类工作却还没有一个公认的标准语料库。本文考虑到文本分类工作的规范性和语料的标准性，由中文标准语料部分文档的翻译和挑选新疆日报 (哈文版) 上的部分文档来搭建了本研究的语料库。在前期研究里，同样是通过翻译收集语料集的，只是其规模小了点，本文的语料工作算是对前期研究语料集的补充和优化完善。前期研究语料集语料文档只有 5 类文档，本文扩充到 8 类文档。通过跟语言学专家们的多次沟通，选择具有代表性的文档，同时对词干提取程序解析规则也作了适当的调整。虽然本文所构建的语料库还不能称得上“标准”词语，但对现阶段哈萨克语文本分类任务的完成具有实际应用价值。

本文把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集，如表 1 所示。数据集被分为两个部分。880 个文档 (63%) 用于训练，520 个文档用于测试 (37%)。

表 1　数据集

Table 1　Data set

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召回率、精度和 F_1 这 3 种评价方法。精度评价是指比较实际文本数据与分类结果，以确定文本分类过程的准确程度，是文本分类结果是否可信的一种度量。高精度意味着一个算法返回更相关的结果，高召回率代表着一个算法返回最相关的结果，所以文本分类工作期望获得较高的精度和召回率。

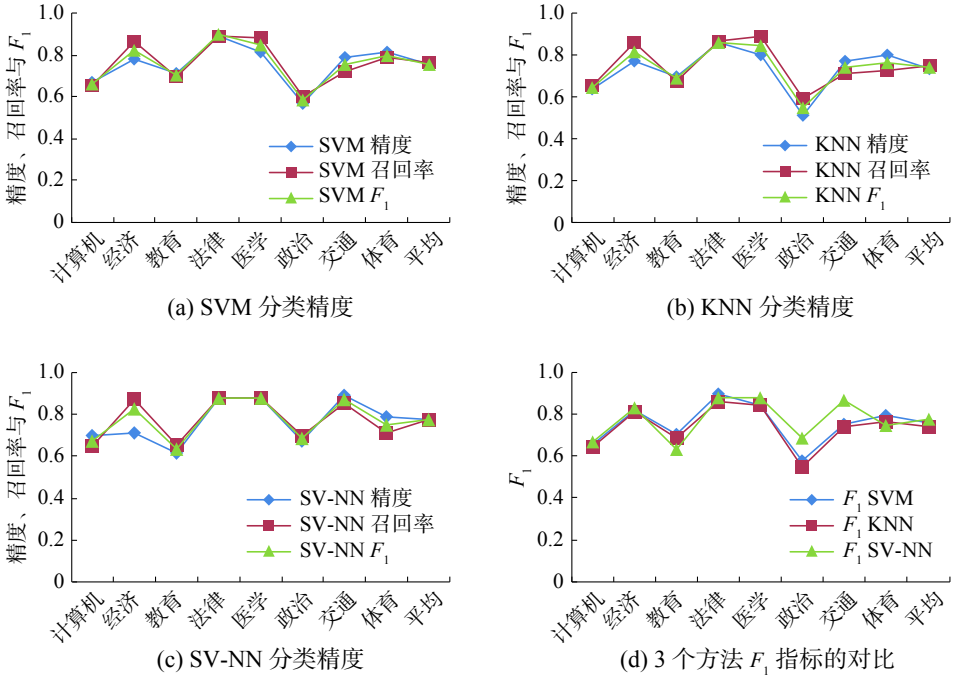
本文在前期研究中搭建的哈萨克文语料集的补充完善以及对其词干提取程序提取规则细节的

优化改善基础上实现了本研究哈萨克语文本的预处理。分类任务的实现运用了 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。通过对表 2 和图 9 上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 F_1 指标除了教育类和法律类以外在其他类上的 F_1 指标都高于都 SVM、KNN 对应指标。SVM、KNN 和 SV-NN 平均分类精度分别为 0.754、0.731 和 0.778, 这说明本文提出算法对所有类别文档词的召回率和区分度较稳定。本研究提出的算法模型继承了 SVM 算法在有限样本情况下也能获得较好分类精度的优点, 另外, 本算法没有去定义 KNN 算法的 k 参数, 也没有跟所有类所有训练样本进行距离运算。所以, 本研究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, 总体精度还是没有像中、英文等其他语言文本分类精度那么理想, 因为涉及很多方面的因素, 如研究语料库语料文档数量、每一类文档本身的质量、词干表里已录用的词干数量和质量、词干提取程序解析规则的细节等, 但目前所获得的分类精度比前期系列研究成果理想, 本算法的文本分类性能有了很大的提升也较好地提高了召回率。

表 2　SVM、KNN、SV-NN 的分类精度对比

Table 2　SVM KNN and SV–NN comparison of classification accuracy

类别名	SVM			KNN			SV-NN		
	精度	召回率	F_1	精度	召回率	F_1	精度	召回率	F_1
计算机	0.665	0.651	0.658	0.636	0.649	0.642	0.696	0.647	0.671
经济	0.781	0.862	0.819	0.768	0.858	0.811	0.710	0.870	0.830
教育	0.715	0.691	0.703	0.698	0.675	0.686	0.615	0.651	0.632
法律	0.889	0.891	0.895	0.860	0.865	0.862	0.879	0.881	0.878
医学	0.812	0.881	0.845	0.799	0.892	0.843	0.877	0.881	0.879
政治	0.563	0.598	0.580	0.513	0.589	0.548	0.673	0.698	0.685
交通	0.791	0.721	0.754	0.769	0.711	0.739	0.891	0.851	0.870
体育	0.811	0.785	0.796	0.801	0.726	0.762	0.791	0.711	0.749
平均	0.754	0.759	0.756	0.731	0.746	0.738	0.778	0.774	0.775



(a) SVM 分类精度　(b) KNN 分类精度

(c) SV-NN 分类精度　(d) 3 个方法 F_1 指标的对比

图 9　分类精度的对比分析（每一类别均含 175 篇文档）

Fig. 9　Comparative analysis of classification accuracy (each category contains 175 documents)

5 结束语

本文在前期系列研究中所搭建的哈萨克文语料集和词干提取程序的优化完善基础上实现了哈萨克语文本的预处理。分类任务的实现上运用了模式识别的 3 种分类算法，并对 3 种分类算法分类精度进行了较全面的对比分析。通过仿真实验客观数字的对比分析，说明本文提出算法的优越性。本文算法对所有类别文档词的召回率和区分度较稳定。本文算法在继承 SVM 算法的分类优越性基础上，还有效避免了 KNN 算法设置 k 参数的麻烦和跟所有训练样本进行距离计算而带来的巨大时间复杂度，进而保证了分类算法的收敛速度。

本研究仍有许多待优化完善的问题，本文接下来的研究工作中将系统地研究并解决影响文本分类精度的阶段性问题，获得满意的分类精度。

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] MARTÍNEZ-CÁMARA E, MARTÍN-VALDIVIA M T, UREÑA-LÓPEZ L A, et al. Polarity classification for Spanish tweets using the COST corpus[J]. Journal of information science, 2015, 41(3): 263–272.

[4] PERCANNELLA G, SORRENTINO D, VENTO M. Automatic indexing of news videos through text classification techniques[M]//SINGH S, SINGH M, APTE C, et al. Pattern Recognition and Image Analysis. Berlin: Springer, 2005: 512–521.

[5] HU Rong, NAMEE B M, DELANY S J. Active learning for text classification with reusability[J]. Expert systems with applications, 2016, 45: 438–449.

[6] SAKURAI S, SUYAMA A. An e-mail analysis method based on text mining techniques[J]. Applied soft computing, 2005, 6(1): 62–71.

[7] AL-KABI M, WAHSHEH H, ALSMADI I, et al. Content-based analysis to detect Arabic web spam[J]. Journal of information science, 2012, 38(3): 284–296.

[8] ZITAR R A, MOHAMMAD A H. Spam detection using genetic assisted artificial immune system[J]. International journal of pattern recognition and artificial intelligence, 2011, 25(8): 1275–1295.

[9] MOHAMMAD A H, ZITAR R A. Application of genetic optimized artificial immune system and neural networks in spam detection[J]. Applied soft computing, 2011, 11(4): 3827–3845.

[10] MAO Ming, PENG Yefei, SORING M. Ontology mapping: As a binary classification problem[J]. Concurrency and computation: practice and experience, 2011, 23(9): 1010–1025.

[11] YANG Yiming, SLATTERY S, GHANI R. A study of approaches to hypertext categorization[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 transactions on electrical and electronic engineering, 2016, 11(4): 488–498.

[13] DUWAIRI R, EL-ORFALI 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.
ZHANG Dongmei. Research on key problems in text sentiment classification and opinion summarization[D]. Ji'nan: Shandong University, 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.
YANG Jieming. The research of text representation and feature selection in text categorization[D]. Changchun: Jilin University, 2013.

[16] 张晓娜. CNNIC 发布第 37 次中国互联网络发展状况统计报告[N]. 民主与法制时报, 2016-01-23(001).

[17] SYIAM M M, FAYED Z T, HABIB M B. An intelligent system for Arabic text categorization[J]. International journal of cooperative information systems, 2006, 6(1): 1–19.

[18] DUWAIRI R, AL-REFAI M, KHASAWNEH N. Stemming versus light stemming as feature selection techniques for Arabic text categorization[C]//Proceedings of the 4th International Conference on Innovations in Information Technology. Dubai, 2007: 446–450.

[19] 贺慧, 王俊义. 主动支持向量机的研究及其在蒙文文本分类中的应用[J]. 内蒙古大学学报: 自然科学版, 2006, 37(5): 560–563.
HE Hui, WANG Junyi. Study of active learning support vector machine and its application on mongolian text classification[J]. Acta scientiarum naturalium universitatis neimongol, 2006, 37(5): 560–563.

[20] ADELEKE A O, SAMSUDIN N A, MUSTAPHA A, et al. Comparative analysis of text classification algorithms for automated labelling of quranic verses[J]. International journal on advanced science engineering information

technology, 2017, 7(4): 1419–1427.

[21] MOHAMMAD A H, ALWADA N T, AL-MOMANI O. Arabic text categorization using support vector machine, naïve bayes and neural network[J]. GSTF journal on computing, 2016, 5(1): 1–8.

[22] 古丽娜孜·艾力木江, 孙铁利, 伊力亚尔·加尔木哈, 等. 一种基于主动学习支持向量机哈萨克文文本分类方法[J]. 智能系统学报, 2011, 6(3): 261–267.
GU Linazi Ai Limujiang, SUN Tieli, Yi Liyaer Jia Ermuhamaiti, et al. An approach to the text categorization of the Kazakh language based on an active learning support vector machine[J]. CAAI transactions on intelligent systems, 2011, 6(3): 261–267.

[23] 古丽娜孜·艾力木江, 孙铁利, 乎西旦·居马洪, 等. 一种基于改进 KNN 的哈萨克语文本分类[J]. 东北师大学报: 自然科学版, 2014, 46(2): 63–68.
GU Linazi Ai Limujiang, SUNTieli, HU Xidan Ju Mahong, et al. Textcategorization of kazakh text based on improved KNN[J]. Journal of northeast normal university: natural science edition, 2014, 46(2): 63–68.

[24] 古丽娜孜·艾力木江, 孙铁利, 乎西旦·居马洪, 等. 一种基于 SVM-修正 KNN 算法的哈萨克语文本分类[J]. 西北师范大学学报: 自然科学版, 2014, 50(3): 48–53.
GU Linazi Ai Limujiang, SUN Tieli, HU Xidan Ju Mahong, et al. An approach to the text categorization of the Kazakh language based on SVM-modified KNN algorithm[J]. Journal of northwest normal university: natural science, 2014, 50(3): 48–53.

[25] 旺建华. 中文文本分类技术研究[D]. 长春: 吉林大学, 2007.
WANG Jianhua. Research on classification of Chinese documents[D]. Changchun: Jilin University, 2007.

[26] JOACHIMS T. Text categorization with support vector machines: Learning with many relevant features[M]// NÉDELLEC C, ROUVEIROL C. Machine Learning: ECML-98. Berlin: Springer, 1998: 137–142.

[27] WANG Ziqiang, SUN Xia, ZHANG Dexian, et al. An optimal SVM-based text classification algorithm[C]//Proceedings of 2006 International Conference on Machine Learning and Cybernetics. Dalian, China, 2006: 13–16.

[28] MONTAÑÉS E, FERÁNDEZ J, DÍAZ I, et al. Measures of rule quality for feature selection in text categorization [M]//International Symposium on Advances in Intelligent. Berlin: Springer,, 2003: 589–598.

[29] CORTES C, VAPNIK V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273–297.

[30] WANG Xuesong, HUANG Fei, CHENG Yuhu. Computational performance optimization of support vector machine based on support vectors[J]. Neurocomputing, 2016, 211: 66–71.

[31] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE transactions on information theory, 1967, 13(1): 21–27.

[32] FRANKLIN J. The elements of statistical learning: data mining, inference and prediction[J]. The mathematical intelligencer, 2005, 27(2): 83–85.

[33] MENG Qingmin, CIESZEWSKI C J, MADDEN M, et al. *K* nearest neighbor method for forest inventory using remote sensing data[J]. GIScience & remote sensing, 2007, 44(2): 149–165.

作者简介:

古丽娜孜·艾力木江, 女, 1972 年生, 副教授, 博士,主要研究方向为机器学习、模式识别、智能信息分类与图像处理。参与国家级、省部级科研项目 3 项, 承担院级重点项目 4 项。发表学术论文 20 余篇。

乎西旦·居马洪, 女, 1966 年生, 教授, 主要研究方向为智能信息处理、人脸识别。承担国家级、省部级科研项目 4 项。发表学术论文 20 余篇, 出版教材 1 部。

孙铁利, 男, 1956 年生, 教授, 博士生导师, 主要研究方向为智能用户接口、智能信息挖掘。承担国家级、省部级科研项目 12 项。发表学术论文 150 余篇, 出版专著及教材 10 部。