基于 SVM 藏文文本分类的研究与实现

文/贾宏云　群诺　苏慧婧　次仁罗增　巴桑卓玛

摘要

利用 SVM（支持向量机）技术对复杂繁琐的汉文文本资源进行快速分类已经相当的成熟，但其在藏文文本分类中的应用还处于研究阶段，因此实验目的在于测试该方法是否在藏文文本分类中具有良好的性能。主要过程包括：文本向量空间模型化，获取 SVM 中核函数的参数并进行常用核函数分类性能对比，最后与 Logistic 回归分类器进行同等条件下的实验对比，验证了支持向量机模型在藏文文本分类中具有良好的分类效果。

【关键词】藏文文本分类　支持向量机　Logistic 回归

1 引言

由于文本记录着时代变迁的痕迹，文本的数量在历史的长河中不断增加，因人们在查阅和学习的过程中对相关文本的寻找显得十分麻烦，所以文本分类的有效性对上述问题的解决具有重要作用。同时伴随着科学技术的发展，人们开始利用计算机的高效性进行文本自动分类，因此对分类数学模型的选择变得更为重要。

目前，机器学习算法已成为主流的方法，尤其在中文文本分类算法的研究上已经相当成熟，特别是 SVM 算法利用最小结构风险的原理使得分类性能更加优异，在武汉理工大学熊浩勇 [1] 的硕士毕业论文中已经详细描述，虽然其具有对模型参数的设置相当复杂并且耗费时间长等不足之处，但所获取的模型参数十分精确。由于 SVM 的核函数很多，因此不同结构的文本使用的核函数具有差异性，藏文文本也存在这种情况。因此实验目的在于测试该方法是否在藏文文本分类中具有良好的性能。主要过程包括：文本向量空间模型化，获取 SVM 中核函数的参数并进行常用核函数分类性能对比，最后与 Logistic 回归分类器进行同等条件下的实验对比，验证了支持向量机模型在藏文文本分类中具有良好的分类效果。

2 SVM模型分类原理

SVM 是一种二分类模型，但可以在多分类中进行多次二分类，它的基本模型定义在样本特征数据空间上的间隔最大的线性分类器，有效的解决样本特征数据在低维空间中非线性（线性也是非线性的特殊情况）的情况下，通过核函数把样本数据映射到高维空间中，利用经验风险和结构风险最小化原理找到线性超平面实现样本分类。由于藏文文本特征的高稀疏性和低维空间中样本的不可分类性，所以本文选择一定的惩罚参数 C 和核函数来构造 SVM 分类器。

2.1 SVM分类器构建算法

2.1.1 确定目标函数

构建最优分类面来分割属于两类的训练集：$(x_i, y_i), i = 1,2...,n, x_i \in R^d, y_i \in \{+1,-1\}$的问题，可转化为解下述二次规划问题：在约束条件 $y_i(w.x_i + b) \geq 1, i = 1,2....,n$下，为了使分类器具有更好的泛化能力和良好的分类效果，求 w 和 b 的优化条件是使两类样本到超平面的距离之和 $\frac{2}{\|w\|}$最大值，其等价于求目标函数$\phi(w) = \frac{1}{2}\|w\|^2$的最小值。

2.1.2 引入松弛变量，构建惩罚参数

<< 上接 143 页

4 实验结果与分析

4.1　实验设计和结果

才旦夏茸大师文集第一册至十三册作为实验语料，其中第一册到第三册为训练语料，用于建设消歧知识库和训练属格的 5 个助词的检错正则表达式，第四册至第六册内容作了修改作为测试语料。语料的规模如表 3。

衡量指标选用了准确率 P:

$$P = \frac{返回正确结果数}{返回结果总数} \times 100\% \quad (8)$$

方法 1 代表基于正则表达式的属格助词自动检错算法 1，方法 2 代表基于正则表达式和知识库的属格助词自动检错算法，实验结果如表 4。

4.2 实验结果分析

从计算的结果可以看出，采用方法 2 检错率比方法 1 的检错率高，虽然只增加了消歧知识库，但准确率明显提高，但方法 2 比方法 1 增加了时间复杂度 T(n)=O(log2(n))。

在实验过程中也发现一些有待解决的问题：消歧知识库的规模不大，还得扩充知识库。

5 结束语

纵观当前少数民族语言文字发展的现状，我们可以清楚地看到，要想句法分析向语义分析阶段顺利迈进，目前最重要的问题就是处理好藏文的虚词，藏文虚词的研究成果可以在各个层面上推广应用。下一步工作计划是，扩充藏文歧义虚词知识库，提高藏文虚词识别和检错的准确性。

总体而言藏文属格助词的识别及其自动检错算法的研究达到了可实用的水平。

参考文献

[1] 卓玛吉，安见才让．藏文不自由虚词的自动识别研究 [J]．商业文化，2014(05)．
[2] 高定国，扎西加，赵栋材．计算机识别藏语虚词的方法研究 [J]．中文信息学报，2014, 28(01): 113-05.
[3] 吴朔平．科技英语虚词分析法简介 [J]．系统工程与电子技术学报，1986(07)．
[4] 杨慧玲．英语虚词在常规句和疑难句中的翻译比较分析 [J]．昆明师范高等专科学校学报，2006, 28(01): 86-88.
[5] 多拉．藏语语义理解中功能性虚词研究 [J]．西藏大学学报（社会科学版），2011, 26(04): 106-107.
[6] 胡书津．简明藏文文法：藏汉对照—2 版 [M]．云南民族出版社，2000(10)．
[7] 格桑局冕，格桑央金．实用藏文文法教程 [M]．四川民族出版社，2004(11)．
[8] 才旦夏著．藏文文法详解 [M]．青海民族出版社，1954, 5: 18-45.

作者单位

拉毛措（（1988-），女，藏族，西藏大学信息科学技术学院，硕士，主要从事自然语言信息处理研究。

作者单位

西藏大学信息科学技术学院　西藏自治区拉萨市　850000

表 1：藏文语料样本数据 单位（篇）

语料收集和预处理								
	教育类	时政类	法律类	政务类	人文类	经济类	民生类	总样本
初始语料集	15000	15000	15000	10000	10000	7000	5000	77000
预处理后语料	700	1400	700	1100	500	500	500	5400

表 2：训练样本集 单位（篇）

训练样本（篇）						
教育类	人文类	政务类	时政类	经济类	法律类	民生类
487	372	820	990	348	477	378

表 3：测试样本集

测试样本（篇）						
教育类	人文类	政务类	时政类	经济类	法律类	民生类
213	128	280	410	152	223	122

由于部分样本点可能离超平面很近，会产生一定的分类误差，为了降低此类误差，*Vapnik* 引入非负的松弛变量 $\xi_i, i=1,2,...,n$ 将 $y_i(w \cdot x_i + b) \geq 1$ 放松为：$y_i(w \cdot x_i + b) \geq 1 - \xi_i$，其中 $\sum_{i=1}^{n} \xi_i$ 是训练错误的一个度量。目标函数改为求 $\phi(w, \xi) = \frac{1}{2}\|w\|^2 + C(\sum_{i=1}^{n} \xi_i)$ 的最小值。其中 C 为惩罚参数，它是一个可调参数，C 越大对错误的惩罚越重。

2.1.3 构建 Lagrange 函数 $L(w,b,a,\xi,\beta)$

首先在式（1）约束条件下利用对偶性求取 Lagrange 函数如（2）所示 $L(w,b,a,\xi,\beta)$ 中 (w,b,ξ,β) 的最小值，得到函数 $Q(a)$ 如公式（3）所示对其优化，求取该函数中 a 的最大值。

$$0 \leq a_i \leq C, i = 1,2,...,n \text{和} \sum_{i=1}^{n} a_i y_i = 0 \quad (1)$$

$$L(w,b,a,\xi,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n} a_i\{y_i(w \cdot x_i + b) - 1 + \xi_i\} - \sum_{i=1}^{n}\beta_i \xi_i \quad (2)$$

$$Q(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j (x_i, x_j) \quad (3)$$

2.1.4 最优分类函数

求解上述问题得到的最优分类函数如（4）所示：

$$f(x) = \mathrm{sgn}\{(w^* \cdot x) + b^*\} = \mathrm{sgn}\{\sum_{i=1}^{n} a_i^* y_i (x_i \cdot x) + b^*\} \quad (4)$$

为了解决低空间样本不可分问题，根据泛函分析中的度量空间理论，倘若有一种函数 $K(x_i,x_j)$ 满足 Mercer 条件的核函数替换线性算法的内积就可以找到原输入空间中对应的非线性算法。如果用特征空间的 $\varphi(x)$ 代替 x, 则 $K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$，把 $L(w,b,\xi,\beta)$ 最优参数代入 $L(w,b,a,\xi,\beta)$ 中此时寻优函数变为 Q(a) 如式（5）所示，然后再求取函数 Q(a) 中 a 的最大值，最后构造分类器 F(x) 如式（6）所示：

$$Q(a) = \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j \varphi(x_i) \cdot \varphi(x_j) \quad (5)$$

$$F(x) = \mathrm{sgn}\{(w^* \cdot x) + b^*\} = \mathrm{sgn}\{\sum_{i=1}^{n} a_i^* y_i K(x_i \cdot x) + b^*\} \quad (6)$$

3 文本语料的收集和预处理

本次基于 SVM 模型的藏文文本分类实现的语料来源于对中国西藏新闻网 (藏文版)，人民网（藏文版），西藏日报（藏文版）等网站相关文章信息的收集而成共 77000 篇，主要以剔除文本数据较小的文本并适当组合形成新语料的方式进行预处理，语料数据数据如表 1。

4 建立向量空间模型

向量空间模型是对文本信息的组合，通常以特征项的权重作为向量空间元素，向量空间模型对分类模型的运算较为便捷。

4.1 特征项定义

特征项是文本的属性包含着文本的信息，在藏文文本中字是其属性的一个元素，字与字之间主要由音节点来分离，字可分为一个字符、多个字符，多个音节点之间又有一定的关联，为了统计的方便和对模型的测试，本次试验选择一个音节点里包含的藏字作为特征项，并且假设每个音节点之间的藏字相互独立。

4.2 特征项权重设置与选择

特征项的权重设置方法很多，通常以该特征项在样本中出现的频度作为其特征权重，首先统计出上述类别文本中各个字出现频率 fn，使藏字特征数据化成 X1 向量，对特征项进行选择时，可以选择相对高频字与相对低频字作为待选特征集，并利用信息增益算法对待选特征集降维，从待选特征集中选择部分信息增益相对大的特征项作为待提取特征集。最后利用欧氏距离算法对待提取特征集中的特征进行聚类，使待提取特征集中的特征形成特征簇，并加权平均特征簇内特征，最终得到文本类别分类特征项。本次基于 SVM 模型的藏文文本分类实现共分 7 个类别，共使用 36 个特征构成分类特征项向量 X=（ཁྲིམས,ཚན,དངོས,སྤྱི,སྲུང,སློངས,འཛིན,དམངས,ཞིབ,ངལ,སློབ,ཡིན,དངུལ,ཞིང,ལོ,ཁང,ཐོན,ས,ཤུགས,འཛུགས,ཆེ,གཞི,འཕེལ,བཏང,ལག,ཏང,པར,གནས,རྩལ,ཞིག,མང,ཁག,གིས,ནང,མང,ཚང），数据化特征向量 X1=(f1,f2,...f35,f36)，X1 ∈ Xn,n 为样本数量，类别向量集合 Y=（教育类：1，人文类：2，政务类：3，时政类：4，经济类：5，法律类：6，民生类：7）标记所有样本成 X=（X1 ∈ Xn，Y1 ∈ Y），构成整个完整的样本向量空间模型数据集。

5 实验结果与评价

实验前需要把所有文本进行向量空间模型化，形成样本数据集，本次实验由三部分内容组成，分别是分类模型对训练集样本数据进行训练获取模型参数、对测试样本数据进行分类预测、对实验结果进行评价。

5.1 模型训练与参数获取

通常分类模型对训练样本数据的训练是获取该模型参数的主要途径，训练样本数据越多，对模型参数的获取就越精确，对训练样本数据的训练有效性对后续测试结果有着直接的影响，训练样本集如表 2。

5.1.1 获取核函数参数

实验中 , 把线性函数、多项式函数、RBF 函数和 SIGMOID 函数作为模型的核函数，通过对训练样本集构成的训练样本数据集进行多次训练和 10 折交叉验证法进行评估，获得核函数参数如下：

LINEAR, 线性，$K(x_i, x_j) = x_i^t x_j$

POLY, 多项式，$K(x_i, x_j) = (2x_i^t x_j + 1)^3$

RBF, 径向基函数，

$K(x_i, x_j) = \exp(-2\|x_i - x_j\|^2)$

SIGMOID,SIGMOID 函数，

$K(x_i, x_j) = \tanh(2x_i^t x_j + 1)$

5.2 模型测试

5.2.1 SVM 模型下不同核函数实验对比

模型测试是对模型训练学习性能的一种检测，模型从训练数据集中获取到参数后，在其它参数相同的条件下，分别对惩罚系数 C 取不同的值，并在表 3 测试样本集构成的测试样本数据集下进行测试，表 4、表 5‘--’记号符表示当前值为零。

5.2.2 SVM 与 Logistic 回归模型测试结果对比

Logistic 回归模型在以前的文本分类实验中，分类效果较好，因此本次实验选择 Logistic 回归模型与 SVM 模型对藏文文本分类性能进行对比，其中实验的平台、语料、文本特征项、特征项权重以及测试和训练数据集完全相同。

5.3 实验结果评价

（1）从上述表 4 与表 5 实验数据可以看出，当核函数选择 LINEAR 和 POLY 时比选择 RBF 和 SIGMOID 的分类效果好，并且选择核函数 LINEAR 和 POLY 自身分类效果较好。

（2）设定不同的惩罚参数 C 对分类效果具有一定影响，LINEAR 和 POLY 变化趋势相

表 4：测试结果

C=0.1 时不同核函数下文本分类的准确率，召回率，F1 值

SVM	LINEAR			POLY			RBF			SIGMOID		
文本类别	精确率	召回率	F1 值	精确率	召回率	F1 值	精确率	召回率	F1 值	精确率	召回率	F1 值
人文类	97.58%	94.53%	0.96	99.17%	93.75%	0.96	--	--	--	--	--	--
教育类	98.54%	94.84%	0.97	96.62%	93.90%	0.95	--	--	--	--	--	--
政务类	97.68%	90.36%	0.94	95.88%	91.43%	0.94	--	--	--	--	--	--
时政类	98.56%	83.66%	0.91	93.06%	81.71%	0.87	26.83%	100.00%	--	--	--	--
经济类	99.29%	92.11%	0.96	97.96%	94.74%	0.96	--	--	--	--	--	--
法律类	99.52%	93.72%	0.97	99.53%	94.17%	0.97	--	--	--	--	--	--
民生类	98.99%	80.33%	0.89	97.96%	78.69%	0.87	--	--	--	7.98%	100.00%	--

表 5：测试结果

C=10 时不同核函数下文本分类的准确率，召回率，F1 值

SVM	LINEAR			POLY			RBF			SIGMOID		
文本类别	精确率	召回率	F1 值	精确率	召回率	F1 值	精确率	召回率	F1 值	精确率	召回率	F1 值
人文类	97.58%	94.53%	0.96	99.17%	93.75%	0.96	0.78%	0.78%	0.01	--	--	--
教育类	98.06%	94.84%	0.96	95.73%	94.84%	0.95	0.47%	0.47%	0.00	--	--	--
政务类	97.32%	90.71%	0.94	98.03%	88.93%	0.93	3.57%	3.57%	0.04	--	--	--
时政类	98.57%	83.90%	0.91	95.94%	80.73%	0.88	27.08%	100.00%	0.43	--	--	--
经济类	98.60%	92.76%	0.96	99.30%	93.42%	0.96	--	--	--	--	--	--
法律类	98.57%	92.83%	0.96	99.06%	94.17%	0.97	--	--	--	--	--	--
民生类	98.06%	82.79%	0.90	97.92%	77.05%	0.86	1.64%	1.64%	0.02	7.98%	100.00%	--

表 6：测试结果对比

C=0.1 时不同核函数下 SVM 和 Logistic 回归文本分类性能对比

SVM	LINEAR			POLY			Logistic		
文本类别	精确率	召回率	F1 值	精确率	召回率	F1 值	精确率	召回率	F1 值
人文类	97.58%	94.53%	0.96	99.17%	93.75%	0.96	98.49%	92.02%	0.95
教育类	98.54%	94.84%	0.97	96.62%	93.90%	0.95	95.35%	96.09%	0.96
政务类	97.68%	90.36%	0.94	95.88%	91.43%	0.94	95.06%	89.29%	0.92
时政类	98.56%	83.66%	0.91	93.06%	81.71%	0.87	93.90%	78.78%	0.86
经济类	99.29%	92.11%	0.96	97.96%	94.74%	0.96	99.31%	94.74%	0.97
法律类	99.52%	93.72%	0.97	99.53%	94.17%	0.97	94.37%	97.76%	0.96
民生类	98.99%	80.33%	0.89	97.96%	78.69%	0.87	91.00%	74.59%	0.82

似。

（3）由选择的特征向量中的值比较大，使特征向量内积和差值相对很大，因此 RBF 和 SIGMOID 的分类效果不好。

（4）从表 6 测试结果可以看出，当 SVM 核函数选择为 LINEAR 和 POLY 并且在上述参数下，从整体参考值上看 SVM 的藏文本分类效果好于 Logistic 回归文本分类效果。

6 总结

本文采用基于 SVM 模型的藏文文本分类实现过程中，为了降低模型的复杂度，对藏文文本特征提取时，忽略词与词之间联系，因此假定词与词之间的互信息为零。实验验证了 SVM 模型对藏文文本具有良好的效果，因此后期会继续研究藏文文本结构形式，增大特征信息量，提高分类的效果。

（通讯作者：群诺）

参考文献

[1] 熊浩勇．基于 SVM 的中文文本分类算法研究与实现 [D]．武汉理工大学，2008.
[2] 李航．统计学习方法 [M]．北京：清华大学出版社，2012.
[3] 崔建明，刘建明，廖周宇．基于 SVM 算法的文本分类技术研究 [J]．计算机仿真，2013.
[4] 高定国，珠杰．藏文信息处理的原理与应用 [M]．成都：西南交通大学出版社，2015.
[5] 杨玉珍，刘培玉，朱振方，邱烨．应用特征项分布信息的信息增益改进方法研究 [J]．山东大学学报（理学版），2009.
[6] 杨杰明．文本分类中文本表示模型和特征选择算法研究 [D]．吉林大学，2013.

作者简介

贾宏云（1990-），男，四川省成都市人。硕士研究生在读，西藏大学。主要研究方向为自然语言处理。
苏慧婧（1989-），女，四川省眉山市人。硕士研究生在读，西藏大学。主要研究方向为自然语言处理。
次仁罗增（1993-），女，藏族，西藏自治区日喀则市人。硕士研究生在读，西藏大学。主要研究方向为自然语言处理。
巴桑卓玛（1991-），女，藏族，西藏自治区山南市人。硕士研究生在读，西藏大学。主要研究方向为自然语言处理。

通讯作者简介

群诺（1972-），男，藏族，西藏自治区拉萨市人。副教授，西藏大学。主要研究方向为自然语言处理。

作者单位

西藏大学信息科学技术学院　西藏自治区拉萨市　850000

● 基金项目：西藏自治区科技计划重大科技专项（ZDZX2017000136）；西藏大学“珠峰学者人才发展支持计划”项目；西藏大学研究生“高水平人才培养计划”项目（2016-GSP-028）。