文章编号:1001 - 9081(2010)06 - 1676 - 03

基于 SVM 的哈萨克语文本分类

王　花,古丽拉·阿东别克,吴守用
(新疆大学 信息科学与工程学院,乌鲁木齐 830046)
(wangdianyuan007@ sina. com)

摘　要:介绍了支持向量机(SVM)和 *k*-最近邻法(kNN)分类算法的思想和两种哈萨克语特征提取方法。对 SVM、kNN 和 Bayes 算法在哈萨克语文本分类的实验进行了比较。实验结果表明:在处理哈萨克语文本分类问题上,SVM 较 kNN 和 Bayes 有较好的分类效果。由于哈萨克文单词的语素和构形的特点,若对哈萨克语词缀进行切分,则会降低文本分类的准确率和查全率。

关键词:文本分类; 支持向量机; 特征选择; *k*-最近邻法
中图分类号: TP391.1　　**文献标志码**:A

Study on Kazak text categorization based on SVM

WANG Hua, GULILA Altenbek, WU Shou-yong
(*College of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang* 830046, *China*)

Abstract: This paper introduced the basic theory of the Support Vector Machine (SVM) and *k*-Nearest Neighbor (kNN) algorithm and two different features selection methods in Kazak natural language. An empirical study of using the SVM, kNN, Bayes algorithm to categorize the Kazak text was conducted. The experimental results show that compared with kNN, Bayes, SVM has better categorization of the Kazak text. Due to the characteristics of Kazak's morpheme and configuration, the precision and recall will be lowered if the word is cut with affix.

Key words: text categorization; Support Vector Machine (SVM); feature selection; *k*-Nearest Neighbor (kNN)

0　引言

随着互联网技术的迅速普及和发展,大量的文字信息开始以计算机可读的形式存在。如何对浩如烟海的文献资料和数据(很大一部分是文本)进行自动分类,已成为重要的研究课题。

目前文本分类的技术有许多,使用较多的是基于统计的学习方法。比较著名的有:贝叶斯分类法、支持向量机(Support Vector Machine, SVM)算法、基于最大熵的文本分类法、*k*-最近邻法(*k*-Nearest Neighbor, kNN)[1-2]等,虽然这些分类算法的分类质量和执行效率都各有不同,但它们的共同点是在文本分类前都需要对文本进行分词。哈萨克语属于黏着语,对哈萨克文进行词的切分工作是相对容易的,因为哈萨克语的词与词之间是以空格分开,切分后的词、构造和变化大多都体现在词缀上,词缀的不同变化,会产生许多不同的词。

汉语的文本分类技术已相当成熟,但是哈萨克语的文本分类仍处于起步阶段。对哈萨克语文本分类,有 3 个要解决的问题[4]:文本的表示、分类器的设计和文本分类器性能的评测。本文重点研究哈萨克语的文本表示、分类器选择和设计。

1　哈萨克语的特点

使用哈萨克语的人口约 130 万,不同于维吾尔语、柯尔克孜语等语言,它具有自己特点,哈萨克语的语素类型分为词根、词干和附加成分 3 种。虽然哈萨克语的形态变化十分丰富,但无论是构词还是构形,语素的附加成分都始终不变。在词根和词缀的组合规律方面尤为突出。

哈萨克语中不再分解为意义上单位的词,而属于词根和词干。词干是词去掉构形附加成分和构词附加成分后剩下的部分,它包含着词的词汇意义。在文本分类之前必须对文本进行词的切分,然后再对词进行词形分析,最后才能进行词语的特征提取。

2　哈萨克语的文本的表示

一个文本表现为文字和标点符号组成的字符串。哈萨克语同样也是由字或字符组成,其中词组成短语,进而形成句、段、节、章、篇的结构。目前文本表示有 *N* 元语法表示法、词组表示法、概念表示法。其中词组表示法虽然提高了特征向量的语义含量,但却降低了特征向量的统计质量,使得特征向量变得更加稀疏,难以从中提取用于分类的统计特征。概念表示法同词组表示法类似,不同之处在于前者用概念作为特征向量来表示,而后者用词组作为特征向量来表示。用概念表示法的时候,需要额外的语言资源(主要是一些词义词典)。

SVM[1-2]的原理是在线性可分的情况下寻找一个最优的超平面,使其在误判率最低的情况下达到最优的分类效果。图 1 是给定的一个线性可分的训练集平面[5]。

图 1 的方框和圆圈分别表示两类文本,实线为分类面,虚

收稿日期:2009 - 12 - 16;**修回日期**:2010 - 04 - 09。
基金项目:国家自然科学基金资助项目(60763005);国家教育部/国家语委民族语言文字规范标准建设及信息化科研项目(MZ115-92)。
作者简介:王花(1978 -),女,甘肃武威人,硕士,主要研究方向:自然语言处理; 古丽拉·阿东别克(1962 -),女(哈萨克族),新疆阿勒泰人,教授,主要研究方向:自然语言信息处理、人工智能; 吴守用(1983 -),男,安徽蚌埠人,硕士,主要研究方向:自然语言处理。

线是平行于实线的分类平面,这个分类平面是在经过训练样本中,离分类面最近的平面。最优超平面[7] 不仅能将所有的训练样本正确划分,而且距该平面最近的异类向量之间的距离最大,用 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{y}=0$ 表示。

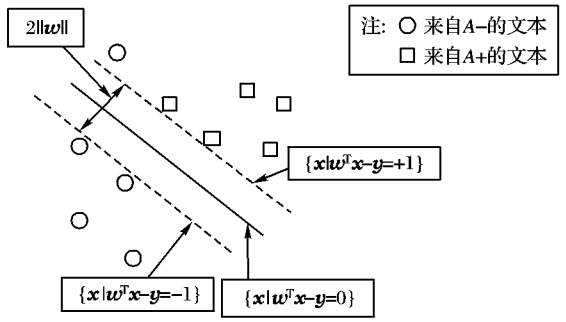

图 1　二维训练集平面

图 1 所示的两个边界超平面 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{y}=1$ 到原点的距离是 $|-\boldsymbol{y}-1|/\|\boldsymbol{w}\|$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{y}=-1$,到原点的距离是 $|-\boldsymbol{Y}+1|/\|\boldsymbol{w}\|$。所以,两个边界超平面的距离是 $2/\|\boldsymbol{w}\|$。

训练集的样本通常表示为:

$$S=((x_1,y_1),I(x_1,y_1))(\boldsymbol{X}\times\boldsymbol{Y})$$

其中:I 是文本的数目;x_1 是文本;y_1 是它们的标记;$\boldsymbol{X}$ 是输入空间,$\boldsymbol{Y}$ 表示输出域。决策平面由式(1) 给出:

$$w_1x_1+w_2x_2-y=0 \tag{1}$$

决策平面可以通过求解一个约束的二次优化问题确定。采用式(2):

$$\boldsymbol{w}=\sum_{i=1}^{k}a_i\boldsymbol{x}_i \tag{2}$$

将训练模式的一个子集展开:

$$\boldsymbol{w}=\begin{bmatrix}w_1\\w_2\\\vdots\\w_n\end{bmatrix}=a_1\begin{bmatrix}x_{11}\\x_{12}\\\vdots\\x_{1n}\end{bmatrix}+a_2\begin{bmatrix}x_{21}\\x_{22}\\\vdots\\x_{2n}\end{bmatrix}+\cdots+a_k\begin{bmatrix}x_{k1}\\x_{k2}\\\vdots\\x_{kn}\end{bmatrix} \tag{3}$$

把这一训练模式的子集称为支持向量,它包含了关于分类问题的所有相关信息,训练集中的决策函数用:

$$f(\boldsymbol{x})=\mathrm{sign}\Big(\sum_{i=1}^{k}a_i(\boldsymbol{x}_i^{T}\boldsymbol{x})-\boldsymbol{y}\Big) \tag{4}$$

在输入空间中,如果训练的文本数据不是线性可分的,支持向量机通过非线性映射 $\phi:\mathbf{R}^n\to F$ 将数据映射到特征空间 F。训练算法采用序列最小优化算法(Sequential Minimal Optimization)[3],然后在 F 中执行上面的算法。

3　哈萨克语文本的特征选择方法

在向量空间模型中,表示文本的特征项可以选择字、词、短语,甚至"概念"等多种元素。如何选择特征,在所选择的特征中赋予多大的权值,以及选择不同的特征对文本分类的性能影响是不同的。目前已有的特征选择方法比较多,常用的有:基于文档频率的特征提取法(Document Frequency, DF)、信息增益(Information Gain, IG)、χ^2 统计量(Chi-square, CHI)和互信息(Mutual Information, MI)方法。

3.1　χ^2 统计量的特征方法

χ^2 统计量(CHI) 衡量的是特征项 t_i 和类别 c_j 之间的相关程度,并假设 c_j 和 t_i 之间符合具有一阶自由度的 χ^2[6] 分布特征对某类的 χ^2 统计值越高,它与该类的相关性越大,携带的信息量越大,反之就越少。如果令 N 表示训练语料库中的文档总数,A 表示属于 C_i 类且包含 t_i 的文档频数,B 表示不属于 C_j 类但包含 t_i 的文档频数,C 表示属于 C_j 类但不包含 t_i 的文档频数,D 是既不属于 C_j 也不包含 t_i 的文档频数,N 为总的文本数量。特征项 t_i 对 C_j 的 CHI 值可表示为:

$$\chi^2(t_i,C_i)=\frac{N\times(A\times D-C\times B)}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{5}$$

对于多类问题采用的是分别计算 t_i 对于每个类别的 CHI 的值,然后在整个训练语料库上计算:

$$\chi^2_{\max}(t_i)=\max_{j=1}^{m}\{\chi^2(t_i,c_i)\} \tag{6}$$

其中 M 是类别数从原始特征空间中去除统计量低于阈值的特征,保留统计量高于给定的阈值的特征,并将其作为文档的特征。

3.2　互信息法的特征方法

互信息量越大,特征 t_i 和类别 C_i 共现的程度越大,A、B、C、D 的含义和3.1 节中的含义相同,那么互信息可以由式(7)计算:

$$I(t_i,c_i)=\log\frac{p(t_i,c_i)}{p(t_i)p(c_i)}\approx\log\frac{A\times N}{(A+C)\times(A\times B)} \tag{7}$$

如果特征 t_i 和类 C_i 无关,则 $p(t_i,c_i)=p(t_i)\times p(c_j)$,那么 $I(t_i,c_i)=0$。对于多类文档识别有用的特征与上面的方法类似,采用式(8):

$$I_{\max}(t_i)=\max_{j=1}^{M}[p(C_j)\times I(t_i,C_j)] \tag{8}$$

除了以上的两种哈萨克文特征提取方法之外还有 DTP 法[7]、期望交叉熵法、优势率法、组合特征提取方法[8] 等。

4　哈萨克语分类器的设计

4.1　朴素的贝叶斯分类器

概率的方法是最早用于文本分类的分离器算法,基本思想是利用特征项和类别的联合概率估计给定文档的类别概率。其中文档 D 属于 C_i 的概率为:

$$p(C_i\mid D)=\frac{P(D\mid C_i)\times P(C_i)}{p(D)} \tag{9}$$

其中:文档 D 采用 DF 向量的表示法,其分量为一个布尔值,0 表示特征没有在该文档中出现,1 表示特征在该文档中出现。于是:

$$P(C_I\mid D)=\frac{P(C_i)\prod_{t_j\in v}p(D(t_j)\mid C_i)}{\sum_i[p(C_i)\prod_{t_j\in v}p(D(t_j)\mid C_i)]} \tag{10}$$

其中:$p(C_i)$ 为 C_i 类文档的概率,$p(D(t_j)\mid C_i)$ 是对 C_i 类文档中特征 t_i 出现的条件概率。可以用式(11) 来估计:

$$p(D(t_j)\mid C_I)=\frac{1+N(D(t_j)C_i)}{2+|D_{CI}|} \tag{11}$$

其中 $N(D(t_j)\mid C_i)$ 是 C_i 类文档中特征 t_i 出现的文档数,$|D_{Ci}|$ 为 C_i 类文档所包含的文档的个数。

贝叶斯分类器易于理解,计算也相对简单,分类效果基本能够满足哈萨克语的要求。

4.2　k- 最近邻法

给定一个测试的文档,系统在训练集中查找离它最近的 k 个邻近的文档,并且根据这些邻近文档的分类来给该文档的候选类别评分。把邻近文档和测试文档的相似度,作为邻近

文档所在类别的权重，也就是说，如果在 k 个文档中，有多个文档属于同一类，则该类的分值为这些文档与测试文档之间的相似度之和，对这 k 个文档所属的类的分值统计完毕后，可按分值进行排序，同时还要给定一个阈值，只有当分值超过阈值时才予以考虑，决策规则用式(12)：

$$y(\vec{x},C_i) = \sum_{\vec{d}_i \in \text{kNN}} sim(\vec{x},\vec{d}_i)y(\vec{d}_i,C_i) - b_i \tag{12}$$

$$y(d_i,C_i) = \begin{cases} 1, & d_i \in C_i \\ 0, & d_i \notin C_i \end{cases} \tag{13}$$

其中：$sim(\vec{x},\vec{d}_i)$ 表示测试文档 $\vec{x}$ 和训练文档 $\vec{d}$ 之间的相似度，b_i 是阈值，b_i 可以通过一个验证文档集来进行调整。验证文档集是训练文档集的一部分。

基于 kNN 的方法还可以实现不同方法的组合变化，可以将基于概括的方法和基于实例的方法结合，在这种方法中，先对训练样本进行聚类得到一般实例，然后再根据一般实例代替训练文档作为 kNN 的输入即可。

4.3 哈萨克语的测试

实验采用 2008 年 1 月《新疆日报》的电子文本数据进行统计(约含哈萨克语单词 30 万个，文档大小为 4.62 MB)。在计算机录入哈萨克文所使用的输入法中，都带有维、哈、柯 3 种文字的输入法，但由于各种输入法所使用的 Unicode 的编码不相同，导致输入的文本在不同的计算机显示时会出现乱码。本文在哈萨克文词处理时也同样遇到了文字编码这类问题，例如：在本文中涉及到的哈萨克文语料使用 elpida 输入法和 Alkatip 输入法，这两种输入法采用的 Unicode 编码和哈萨克文的标准 Unicode 编码不同。在实验中采用 elpida 输入法。同在进行文本中词的切分前要去除包含汉语的哈萨克语翻译，目的是保持统计哈萨克语词组的“纯洁”同时进行内码转换。对哈萨克语词按空格切分后要去掉停用词，然后进行词形分析，最后按照上述方法提取哈萨克语词特征向量。

5 性能评价及结果分析

针对哈萨克语的文本，常采用的分类器性能评价方法包括召回率、正确率、平衡点[10]和平均正确率等[11]。其最基本的准则是更快、更准确地对文本进行分类。

$$P = \frac{a}{a+b} \times 100\% \tag{14}$$

$$r = \frac{a}{a+c} \times 100\% \tag{15}$$

$$F_1 = \frac{2 \times P \times r}{P+r} \tag{16}$$

其中：a 为正确分类的文本数，b 为文本分类数，c 为分类器将文本错误的排除在某个类别之外的个数，P 为准确率，r 为查全率，F_1 为测度值。

本文采用准确率、查全率和 F_1 来评价。数据集采用《新疆日报》上的 340 篇文章作为训练样本，120 篇文章作为测试样本，共计 460 篇。将哈萨克文按环境、教育、经济、体育和政治共分为 5 类，性能比较结果如表 1 所示。

表 1 哈萨克语分类方法的结果对比

算法	封闭测试			开放测试		
	P	r	F_1	P	r	F_1
SVM	0.8847	0.8675	0.8760	0.8211	0.8201	0.8205
kNN	0.7654	0.7586	0.7619	0.7431	0.7126	0.7275
Bayes	0.7112	0.7103	0.7107	0.7005	0.6896	0.6950

从数据结果来看，SVM 的性能在哈萨克语中分类性能最优。由于哈萨克语属于“黏着”语言的特点使 SVM 的方法比起处理汉语的工作相对简单。但在实验中也发现：对于哈萨克语，将词按空格分开后，对词进行切分(去掉词的附加成分)后会使 P、r、F_1 下降，这是由于哈萨克语的词是由根素和缀素构成，根素构成了哈萨克语词意的核心，缀素的作用表现在：1)缀素不能单独出现，只能与根素结合，或与词干结合，改变原词词汇的意义，使其变成另外一个词；2)改变原词的词性。如果在实验中对哈萨克语词缀切分，结果如表 2 所示。

表 2 哈萨克语词缀变化

哈萨克语词语	汉语	哈萨克语词缀
bill	知道	无词缀
billim	知识	lm—名词词缀
bilimpaz	博学者	Paz—词缀表示人
bilimazadap	学识渊博	dap—抽象名词词缀

通常在对哈萨克语词性分析，尤其在哈萨克语词性标注时都会对切分出的词再次切分(去掉词缀)，但在哈萨克语文本分类中不能再次切分。在实验中可以看出：Bayes 方法在哈萨克语的分类中性能不是很理想，这种分类器的性能很容易受到分类任务大小的影响[9]。

6 结语

本文采用基于 SVM 的方法来表示文本，并采用 χ^2 统计量和互信息法方法对哈文进行特征提取，采用 kNN 分类器和 Bayes 分类器对文本进行分类。下一步要研究的问题是改进和完善上面所提到的模型和算法，以及当提高哈萨克语特征维度时，使文本分类的查全率和准确率进一步提高。

参考文献：

[1] MITCHELL T M. Machine learning [M]. New York: McGraw Hill, 1997.

[2] VAMNIK V. Statistical learning theory [M]. New York: Wiley, 1998.

[3] OSUNA E, FREUND R, GIROSI F. Support vector machines: Training and applications, AI Memo 1602[R]. Cambridge: MIT, 1997.

[4] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社, 2008.

[5] SOMAN K P. 数据挖掘基础教程[M]. 范明, 牛常勇, 译. 北京：机械工业出版社, 2009.

[6] DUNNING T. Accurate methods for the statistics of surprise and coincidence [J]. Computational Linguistics, 1993, 19(1): 61 - 74.

[7] MOYOTL-HERNANDEZ E, JIMENEZ-SALAZAR H. Enhancement of DTP feature selection method for text categorization [C]// CICLing 2005. Washington, DC: IEEE, 2005: 719 - 722.

[8] 代六玲, 黄河燕. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报, 2004, 18(1): 26 - 32.

[9] 石志伟, 吴功宜. 改善朴素贝叶斯在文本分类中的稳定性[C]// NCIRCS2004. 上海: [出版者不详], 2004: 137 - 145.

[10] AAS K, EIKVIL L. Text categorisation: A survey (1999) [EB/OL]. [2009 - 10 - 10]. http//citeseer. ist. pus. edu/aas99text. html.

[11] TAGHVA K, BORSACK J, LUMOS S, *et al*. A comparison of automatic and manual zoning: An information retrieval prospective [J]. International Journal on Document Analysis and Recognition, 2004, 6(4): 230 - 235.