第16卷 第11期
2006年11月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.11
Nov. 2006

基于支持向量机的中文文本分类模型研究

马忠宝,刘冠蓉
(武汉理工大学 计算机科学与技术学院,湖北 武汉 430070)

摘 要:支持向量机是在统计学习理论基础上发展起来的新一代学习算法,适宜构造高维有限样本模型,具有很好的分类精度和泛化性能。文中介绍了中文文本分类过程,将支持向量机应用于中文文本分类模型中,对分类器参数选择进行了分析和讨论。实验分析表明,该系统在较小训练集条件下可以取得较好的分类效果。

关键词:支持向量机;文本分类;模型
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2006)11-0070-03

Research on Chinese Text Classification Model Based on SVM

MA Zhong-bao, LIU Guan-rong
(School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China)

Abstract: Support vector machine(SVM) is a new learning algorithm based on statistics theory, and it is proved very useful for text classification. In this paper a model of Chinese text model based on SVM is built and different type of kernel functions is used. According to the experiment, it is showed that this model has good result for text classification.

Key words: support vector machine; text classification; model

0 引 言

随着中国经济社会的不断发展,越来越多的领域要求对中文信息能够有效地进行处理,并迅速对之加以利用。文本分类的任务就是将文本根据内容分为预先定义的若干类别,在邮件过滤、信息检索、办公自动化等方面都有广泛和深入的应用。国内对此研究起步较国外晚些,不过随着近年来国内中文信息处理技术的发展,在某些方面已经取得了令人瞩目的研究成果。

支持向量机(support vector machine)是一种基于统计学习理论的机器学习方法,由 Boser, Guyon, Vapnik 等人在 COLT(Computational Learning Theory)-92 上首次提出,在文本分类、图像识别、生物信息等领域都取得了成功的应用。相比较传统的分类检测方法,支持向量机在求解小样本、非线性、高维空间、局部极小点等问题上表现出了较好的性能[1]。支持向量机根据结构风险最小化原则,具有全局最优解,同时提高了分类器的泛化能力。

文中首先介绍了文本分类的流程,分析了文本分类系统的组成,然后提出了一种基于支持向量机的中文文本分类模型,并用在互联网上收集的文本训练集和测试集对该模型进行训练和测试,详细讨论了该模型的工作原理和实验结果。最后对所进行的工作进行小结,对将来的工作进行了展望。

收稿日期:2006-03-14

作者简介:马忠宝(1977-),男,安徽淮南人,硕士研究生,研究方向为智能计算、机器学习;刘冠蓉,教授,研究方向为并行算法、网络计算。

1 文本预处理

1.1 汉语分词

中文文本预处理首先要进行分词,根据"贝叶斯假设",假定组成文本的字或词在确定文本类别作用上相互独立,这样就可用文本中出现的字或词的集合来代替文本。通过分词获得的全部原始特征构成了文本特征全集,进而可以将文本用向量形式表示[2],以便计算机处理。汉语分词主要有3种算法:机械分词法、语义分词法和人工智能法。文中采用机械分词法中的双向匹配法进行分词,算法描述如下:

(1) 建立分词词典;

(2) 定义类 ASM(d, a, m);//Automatic Segmentation Model

(3) d 表示匹配方向,+1 表示正向,-1 表示逆向;a 表示匹配字符增量;m 表示匹配模式,1 表示最大匹配,-1 表示最小匹配;

(4) 遍历文本,打印分词及间隔标志。

1.2 特征提取

特征提取是文本分类过程中的一个关键环节,目前国际上对文本特征提取多数通过采用某种评估函数,计算特

征属性的权重,选取权重在一定数目或某个值范围之内的特征项作为文本的特征子集,常用的评估函数有 X^2 统计、信息增益、互信息等[3]。其中 X^2 统计经过众多文献证明是一种有效的特征提取方法,具体算法如下:

$$X^2(t,c) = \frac{N \times (A \times D - C \times B)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

公式中 N 表示文档集合的文档总数,A 表示包含 t 且属于 c 类的文档频数,B 表示包含 t 且不属于 c 类的文档频数,C 表示不包含 t 且属于 c 类的文档频数,D 表示不包含 t 且不属于 c 类的文档频数。特征项 t 对文档类别 c 的 $X^2(t,c)$ 统计值越高,与该类别之间的相关性就越大。如果有 m 个类别,每个 t 就会有 m 个值,取其平均值作为该特征项对所有类别的 $X^2(t)$ 统计值,即,$X^2_{avg}(t) = \sum_i P(c_i)X^2(t,c_i)$ 从原始特征空间中删除低于特定值的特征项,将高于该值的特征项构成文本向量的特征子集。

2 支持向量机

2.1 线性可分

设文本特征向量为 $\boldsymbol{X}$,所属类别为 Y,如图 1 所示,对于样本集 (x_i, y_i),$1 \leqslant i \leqslant n$,$x_i \in R^d$,$y_i \in (-1, +1)$,$H$ 为最优分类面,在 d 维空间中,若要求所有样本分类正确且分类间隔最大,则应满足如下约束条件[4]:

$$\min(\frac{1}{2} \| w \|^2) \tag{1}$$

$$y_i(w \cdot x_i + b) \geqslant 1 \tag{2}$$

其中分类间隔等于 $2/\| w \|$,间隔最大等价与 $\frac{1}{2} \| w \|^2$ 最小。满足式(1),(2)的分类面就是最优分类面,使得 $y_i(w \cdot x_i + b) = 1$ 的训练样本点称作支持向量(support vector)。引入 Lagrange 函数:

$$L(w,b,\alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{l} \alpha_i [y_i(w \cdot x_i) + b) - 1], \alpha_i \geqslant 0 \tag{3}$$

对 w 和 b 取 L 的偏微分,求式(3)的极小值可得:

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i, \alpha_i \geqslant 0 \tag{4}$$

每个样本向量(即每篇文档)对应一个 α_i,其中 $\alpha_i > 0$ 的训练样本即为支持向量,l 为支持向量的个数。将式(4)代入式(2)得到样本分类函数:

$$f(x) = \mathrm{sgn}(w \cdot x + b) = \mathrm{sgn}[\sum_{i=1}^{l} \alpha_i^* y_i (x_i \cdot x) + b^*] \tag{5}$$

其中 sgn() 为符号函数,α_i^* 为最优解,b^* 为阈值,可由原始约束条件(2)得到:

$$b^* = -\frac{1}{2}[\max_{y_i=-1}(w \cdot x_i) + \min_{y_i=1}(w \cdot x_i)] \tag{6}$$

对于经过预处理的未知类别的测试向量 x,将其值代入到分类函数式(5)中即可判别出其类别。

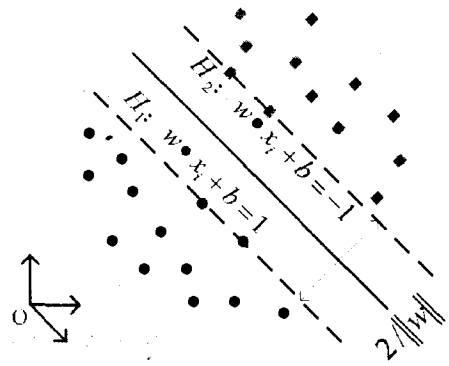

图 1　线性支持向量机

2.2 线性不可分

对于线性不可分情况,引入缓冲量 $\xi_i (\xi_i > 0)$,把式(2)进行变化,如图 2 所示。

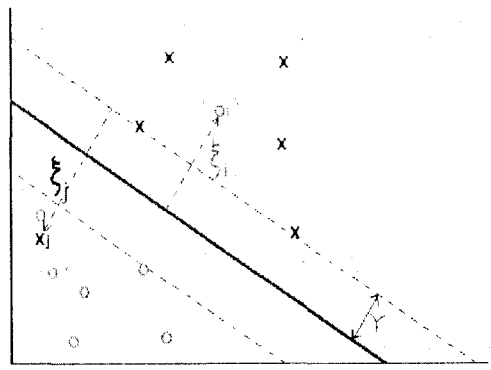

图 2　线性不可分支持向量机

$$y_i(w \cdot x_i + b) \geqslant 1 - \xi_i \tag{7}$$

相应 Lagrange 函数:

$$L(w,b,\alpha) = \frac{1}{2}(w \cdot w) + c\sum_{i}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i [y_i(w \cdot x_i) + b) - 1 + \xi_i] - \sum_{i}^{l} u_i \xi_i \tag{8}$$

求解方法与线性可分类似,只是目标函数多了缓冲量 ξ_i,同时要求 $c \geqslant \alpha_i \geqslant 0$,即在最大分类间隔和最少错分数量之间取得平衡。

2.3 非线性

对于非线性问题,支持向量机采用特征映射方法,利用线性问题的计算框架来实现[5],其原理如图 3 所示。通过引入核函数 K,使得 $K(x_i, x_j) = \phi(x_i)\phi(x_j)$,实现非线性变化后的线性分类,相应的分类函数为:

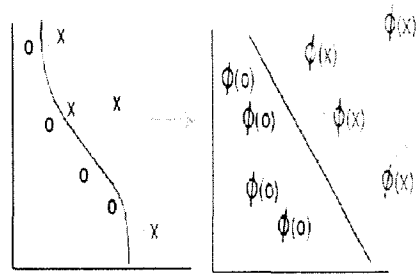

图 3　非线性支持向量机

$$f(x) = \mathrm{sgn}[\sum_{i=1}^{l} \alpha_i^* y_i K(x_i \cdot x) + b^*] \tag{9}$$

常见的核函数有：

* 多项式核函数：$K(x,y)=[s(x\cdot y)+c]^d, d\in N$

* RBF(Radial Basis Function) 核函数：$K(x,y)=\exp[-\|x-y\|^2/(2\sigma^2)]$

* Sigmoid 核函数：$K(x,y)=\tanh[s(x,y)+c]$

3 实验与分析

实验中分类模型所用的训练和测试文本来自互联网，按类别存放。每个类别的训练集和测试集中都含有 300 个文本文件。首先采用双向匹配法对文本进行分词，再使用 X^2 统计评估函数对分词所得的特征全集进行特征提取，将文本表示成多维空间向量，最后使用支持向量机分类算法建立文本分类器。评判分类效果一般用查全率(recall)和查准率(precision)表示分类效果，文中综合了查全率和查准率在评判中的因素，引入 F 测度值作为分类效果的评判指标，定义如下：

查全率 = 事实属于此类且被分类正确的文档数 / 属于此类的总文档数；

查准率 = 事实属于此类且被分类正确的文档数 / 被判为此类的文档数；

F 测度值 = 2×查全率×查准率 /(查全率 + 查准率)。

实验结果如表 1 所示。

从表中可以看出，这种采用支持向量机算法建立的分类模型普遍具有较高的 F 测度值，其中以线性 SVM 和多项式核 SVM 这两种分类器的分类效果最为平均，Sigmoid 核 SVM 分类器具有相对最高的单类别 F 测度值，RBF 核 SVM 分类器具有相对最低的单类别 F 测度值。由此可见，SVM 分类器的参数选择对分类效果的影响与训练集和测试集的选择有很大关联，参数选择对分类效果的影响并不具备明显的规律性。

表 1 分类效果实验结果

文本类别	政治	经济	农业	环境	计算机	体育
线性 SVM	92.73%	86.27%	90.47%	89.40%	95.987%	92.23%
多项式核 SVM	92.88%	85.13%	89.25%	88.77%	94.98%	91.58%
RBF 核 SVM	96.94%	91.13%	78.96%	81.86%	93.65%	97%
Sigmoid 核 SVM	96.68%	91.09%	79.77%	84.26%	95.47	97.27%

4 结束语

支持向量机是一种已被证明有效的学习算法，在文本分类领域中有着广泛和深入的应用。文中介绍了中文文本分类实现过程，阐述了线性和非线性支持向量机的原理，设计出了基于向量机的文本分类模型并进行了实验分析。此外，还对不同支持向量机分类器的性能作了比较。实验表明，这种模型在小样本条件下获得了较好的分类效果。

参考文献：

[1] 贾 洞，梁永帧. 基于支持向量机的中文网页自动分类[J]. 计算机工程，2005，31(10)：145－147.

[2] 邱均平，文庭孝，周黎明. 汉语自动分词与内容分析法研究[J]. 情报学报，2005，24(3)：309－315.

[3] 孙国菊，张 杰. 中文文本分类的特征选取评价[J]. 哈尔滨理工大学学报，2005，10(1)：76－78.

[4] 李 亮，刘万春，徐泉清，等. 一种基于支持向量机的专业中文网页分类器[J]. 计算机应用，2004，24(4)：58－61.

[5] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York：Springer－Verlag，2000.

(上接第 69 页)

依据文中编码规则“晋”字标准编码：

0541212000052110000 或 0541212000025110000

根据图 1(b)～(e)，对“晋”字提取编码：

0541212000052110000

5 结 论

文中采用汉字统计特征粗分类和笔划结构细分类相结合的方法，模仿人认识汉字的过程，提出了一种适于机器识字的汉字容错编码方法。

(1)定义了仿人拆字的字元集，给出了易混淆笔划字元的容错编码。

(2)给出了笔划字元的顺序判断规则。

(3)归结了 37 类简单常用的部首为子结构编码，并给出冗余的容错编码。

(4)建立了仿人构字的编码规则。

实验和仿真结果表明，该编码方法的字元易于稳定提取，能很好地表征和区分汉字集。

参考文献：

[1] 朱 辉，杨 扬，颉 斌，等. SVM 在小字符集手写体汉字识别中的应用研究[J]. 微计算机信息，2004，20(4)：74－75.

[2] 陈友斌，丁晓青，吴佑寿. 非特定人脱机手写汉字识别[EB/OL]. 2005－03－20. 中国 OCR 信息网. http://www.chinaocr.net/show_hdr.php?xname=TVKUIV0&xpos=6&dname=.

[3] 钱自拓. 汉字图像识别研究[D]. 合肥：合肥工业大学，2005：34－38.

[4] 陈治平，林亚平，李军义. 基于笔划和笔顺的汉字识别算法[J]. 湖南大学学报，2000，27(4)：103－104.

[5] 王竹林. 图像字符的仿人认知特征的机理研究[D]. 合肥：合肥工业大学，2005：34－38.