基于SVM的中文文本分类算法*

冀胜利,李 波

(重庆工学院 电子信息与自动化学院,重庆 400050)

摘要:提出了一种基于支持向量机的中文文本分类算法,介绍了文本分类过程中的文本表示、特征提取和SVM算法等关键技术.最后进行了实验和分析,由实验结果可以看出,该方法在精确率和召回率等方面能够达到比较好的效果.

关 键 词:支持向量机;特征提取;文本分类
中图分类号:TP311 文献标识码:A 文章编号:1671-0924(2008)07-0084-04

Chinese Text Categorization Algorithm Based on SVM

JI Sheng-li, LI Bo

(School of Electronic Information and Automation, Chongqing Institute of Technology, Chongqing 400050, China)

Abstract: A Chinese text categorization algorithm based on SVM is presented, and its key techniques in the process of text organizing-text expression, feature selection, SVM algorithm, etc. are introduced. The experiments show that it works well in precision rate, recall rate, etc.

Key words: support vector machine; feature selection; text categorization

文本分类属于有指导的机器学习,是指在给定的分类系统下,根据文本的内容或属性,将大量文本归到一个或多个类别的过程.它是为降低查询时间,提高个性化搜索质量,方便用户快速有效获取文本而产生的文本处理技术.20世纪90年代以来,众多的统计方法和机器学习方法应用于文本分类.现有的分类方法主要有:向量空间模型(Vector Space Model,VSM)、决策树(Decision Tree),Rocchio,KNN和支持向量机(Support Vector Machine,SVM)等[1].SVM是Vapnik等人根据统计学理论(Statistical Learning Theory,SLT)提出的一种基于结构风险最小化原理(structural risk minimization,SRM),具有高泛化性能的通用学习机器.这是一种专门研究小样本情况下机器学习规律的理论,这种理论具有坚实的统计学理论基础,并在实际应用中显示了独特的优越性,如手写数字识别(handwritten digit recognition)、文本分类[2]等.

1 文本分类的预处理

1.1 文本分类的任务描述

一般来说,文本分类系统的任务是:在给定的分类体系下,根据文本的内容或属性自动的确定文本的类别.从另一个角度来看,文本分类是一个

* 收稿日期:2008-04-28
基金项目:重庆市自然科学基金资助项目(CSTC,2006BB2084).
作者简介:冀胜利(1981—),男,山东鄄城人,硕士研究生,主要从事测试与控制技术研究;李波,男,博士,教授,主要从事信息安全与计算机网络研究.

映射过程,即将未知类别的文本映射到已有的类别中.该映射可以是一对一,也可以是一对多.文本分类中的映射规则是系统根据给定的训练样本类别信息,分析总结并结合相应算法而得出的判别式或判别规则,然后再应用于新的待测文本,根据总结出的判别式或判别规则,确定其类别.基于SVM的中文文本分类模型如图1所示.

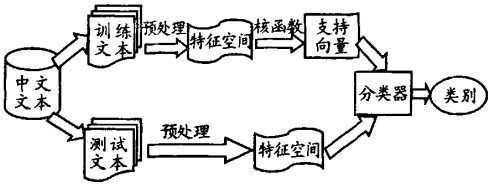

图1 基于SVM的中文文本分类模型

1.2 文本表示

在特征向量提取前先要对文本进行预处理,对于中文文本而言,词与词之间没有明显的切分标志,所以需要分词.分词的方法主要有基于理解的方法,基于字符串匹配的方法和基于统计的方法.同时对切分词还要进行词性的标注、人名地名等专有名词的识别.

计算机不具有人类的智能.为了进行后续计算机处理,文本文档必须要有一种有效的表示方式.目前,在信息处理领域,文本表示的模型主要有Dumais等人提出的潜在语义索引(LSI)[3]模型和Salton等人提出的向量空间模型(Vector Space Model,VSM)[4]等.本研究采用后者,其基本思想[5]是把文档简化为以特征项的权重为分量的向量表示:$(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$,其中$t_i$为文档中第$i$个特征项,$w_i$为第$i$个特征项的权重.目前有多种TF-IDF公式[6],这里给出一个常用的归一化公式:

$$w_{ik}=\frac{tf_{ik}\times\log(N/df_k+0.01)}{\sqrt{\sum_{T_k\in D_i}[tf_{ik}\times\log(N/df_k+0.01)]^2}} \tag{1}$$

其中:w_{ik}表示特征项t_k在文档D_i中的权重;tf_{ik}表示特征项t_k在文档D_i中出现的频率,tf_{ik}越高,意味着特征项对文档越重要;df_{ik}表示出现特征项的文档频率,df_{ik}越高,意味着特征项t_k在衡量文档之间相似性方面的作用越低;N为全部文档的数量;分母为归一化因子.

1.3 特征提取

在中文文本分类中使用较多的特征抽取方法包括文档频率(Document Frequency,DF)、互信息(Mutual Information,MI)、信息增益(Information Gain,IG)和χ^2统计等.本研究采用DF作为特征抽取的方法.为了减少信息的噪声,对文本文档进行处理:① 对文本分词以后,取出文档中包含的所有字词;② 除去停用词,比如的、了等;③ 统计每个字词出现的频率;④ 根据需要过滤出现频率较高的那部分字词和出现频率较低的那部分词;⑤ 余下的词作为特征项(term)进行唯一标注.

由于DF具有相当于训练语料规模的线性计算复杂度,因此能够容易被用于大规模语料统计.在对文档进行处理时,过滤了出现频率较低的那部分词,但在信息检索(Information Retrieval,IR)研究中通常认为DF值低的特征词相对于DF值高的特征词具有更大的信息量,不应该将它们过滤掉.

2 基于支持向量机的文本分类算法

2.1 支持向量机

统计学习理论[7-8]给出了线性分类器边缘与其泛化误差之间关系的形式化解释,我们称这种理论为结构风险最小化(SRM)理论.该理论给出了分类器泛化误差的一个上界,即在概率$1-\eta$下,分类器的泛化误差在最坏情况下满足

$$R\leqslant R_{emp}+\varphi\left(\frac{h}{N},\frac{\log(\eta)}{N}\right) \tag{2}$$

其中:h是反映学习机器学习能力(复杂度)的参数,称为VC维,不等式右侧的第2部分通常称置信范围.学习的目标是最小化经验风险和置信范围的和,也就是结构风险最小.机器学习通常要在经验风险和学习机器的复杂度之间折衷选择.

2.2 支持向量机算法

SVM是从线性可分情况下的最优分类面发展而来的,基本思想[9]由图2的二维情况说明,方框和圆点分别代表2类样本,H是分类线,H1,H2分别是穿过2类离分类线最近的样本且平行于分类线的直线,2条直线之间的距离叫做分类间隔(Margin).所谓最优分类线就是要求分类线下不但能将2类样本正确分开(训练错误率为0),而且使分类间隔最大.

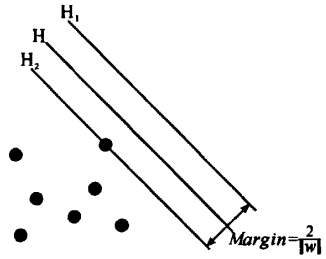

图2 最优分类面

设有 N 个样本，每个样本表示为一个二元组 $(x_i, y_i)(i=1,2,\cdots,N)$，其中 $x_i \in \mathbf{R}^n$，令 $y_i \in \{-1,1\}$ 表示它的类标号.一个线性分类器的决策边界可以写成：

$$W \cdot X + b = 0 \tag{3}$$

其中 W 和 b 是模型的参数.

SVM 的训练阶段包括从训练数据中估计决策边界的参数 W 和 b，选择的参数必须满足 2 个条件：

$$W \cdot X_i + b \geqslant 1 \quad \text{如果 } y_i = 1 \tag{4}$$

$$W \cdot X_i + b \leqslant -1 \quad \text{如果 } y_i = -1 \tag{5}$$

以上 2 个不等式可概括为：

$$y_i(W \cdot X_i + b) \geqslant 1, i = 1,2,\cdots,N \tag{6}$$

此时分类间隔等于 $\frac{2}{\|W\|}$，使其最大化等价于最小化下面的目标函数

$$f(W) = \frac{\|W\|^2}{2} \tag{7}$$

SVM 的学习任务可以形式化的描述为

$$\min_W \frac{\|W\|^2}{2} \tag{8}$$

$$\text{s.t.: } y_i(W \cdot X_i + b) \geqslant 1, i = 1,2,\cdots,N;$$

利用拉格朗日乘子(Lagrange Multiplier)方法可以得到：

$$\min_W \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i \tag{9}$$

$$\text{s.t.: } \sum_{i=1}^{N} y_i \alpha_i = 0 \text{ and } \alpha_i \geqslant 0, i = 1,2,\cdots,N;$$

那么可求决策函数

$$f(x) = \text{sgn}((w \cdot x) + b) \tag{10}$$

当 SVM 应用于非线性决策边界数据集上时，要引入非线性映射.通过非线性映射 ϕ，把数据从原来的特征空间映射到一个高维特征空间 Z，决策边界在这个空间下成为线性.在特征空间 Z 中构造最优化超平面时，训练算法都涉及计算向量之间的点积 $\phi(x_i) \cdot \phi(x_j)$，这样的运算可能导致维灾难问题.为了解决这类问题，引入一种核技术(kernel trick)的方法，只要能找到函数 K 满足 $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$，这样在高维空间中进行的内积运算就可以用原空间中的函数来实现，维灾难问题就解决了.在非线性 SVM 中使用的核函数必须满足 Mercer 定理，因此不需要知道映射函数 ϕ 的确切形式.在 SVM 中，构造的复杂度取决于支持向量的数目，而不是特征空间的维数.SVM 的原理如图 3，其中 y 为决策函数.

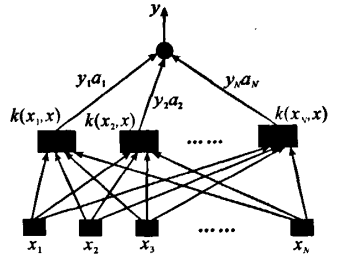

图3 支持向量机的原理

2.3 算法描述

本研究是对“一对多”的 SVM 文本分类算法[10]的改进，提出了中文文本分类算法，具体可描述为：

1) 设训练样本集 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$，其中 $x_i \in \mathbf{R}^n, y_i \in \{1,2,\cdots\cdots,N\}$；

2) 选择适当的核函数 $K(x,y)$ 和适当的参数，利用二次规划求解最优化问题.

常用的核函数有：

a) 多项式核函数

$$K(x,y) = (x \cdot y + c)^d, (d = 1,2,\cdots) \tag{11}$$

b) 径向基核函数

$$K(x,y) = \exp(-\gamma x - y^2) \tag{12}$$

c) Sigmoid 核函数

$$K(x,y) = \tanh(kx \cdot y + \sigma) \tag{13}$$

本研究选择多项式作为核函数，选取参数 c 为 1.那么最优化问题为

$$\min_\alpha \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (K(x_i, x_j) + m\lambda_{ij}) - \sum_{i=1}^{N} \alpha_i \tag{14}$$

$$\text{s.t.: } \sum_{i=1}^{N} y_i \alpha_i = 0 \text{ 且 } \alpha_i \geqslant 0, i = 1,2,\cdots,N;$$

$$\text{其中: } \lambda_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{其他} \end{cases} \tag{15}$$

使用二次规划技术可解得最优解

$$\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*)^T \quad (16)$$

3) 选择 α^* 的一个正分量 α_j^*,并据此计算

$$b^* = y_i(1 - m\alpha_j^*) - \sum_{i=1}^{N} y_i \alpha^i K(x_i, x_j) \quad (17)$$

4) 构造决策函数,判断 x 的类别

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{N} \alpha_i^* y_i K(x_i, x) + b^*) \quad (18)$$

在这里,并没有通过对应 k 个 SVM 分类器构造 k 次限制问题,而是构造了 k 个两类规则,这样就可以只解决一个优化问题从而节省了计算时间.

3 中文文本分类的实现

3.1 测试指标

因为文本分类从根本上说是一个映射过程,所以评估分类系统性能的标志是映射的准确程度和映射的速度.在这里我们就采用准确率(precision)和召回率(recall)2个性能指标来评估该分类系统.

2个指标可表示为:

$$\text{准确率} = \frac{\text{正确分到某类的文本数}}{\text{实际分到某类的文本数}} \quad (19)$$

$$\text{召回率} = \frac{\text{正确分到某类的文本数}}{\text{实际应分到某类的文本数}} \quad (20)$$

准确率和召回率反映了分类质量的2个不同的方面,2个必须综合考虑,不可偏废,因此,选取 F 作为两者综合考虑的评估指标.可表示为

$$F = \frac{\text{准确率} \times \text{召回率} \times 2}{\text{准确率} + \text{召回率}} \quad (21)$$

3.2 测试分析

目前,国内还没有标准的分类测试语料库.在实验中,选择了复旦大学语料库来进行分类测试实验.实验中的训练语料选择了计算机、体育、环境等五大类,共6 451篇文档.测试语料在每类中任意选择了200篇,共1 000篇.在实验中,本研究采用 Matlab 7.1 作为仿真软件.

在训练阶段,首先对实验语料按照上述步骤进行预处理.分词以后,去掉停用词,得到特征空间.选择合适的参数构造分类器,直到分类器具有较好的分类能力,分类器就构造好了.

在测试阶段,用构造好的分类器去评测测试语料,采用上述通用的准确率和召回率对系统进行测试.测试结果如表1所示,其中平均准确率为90.86%,平均召回率为88.68%,平均 F 值为89.76%.通过实验可以看出,该方法能够达到比较好的效果.

表1 实验结果

类别	计算机	体育	环境	政治	经济
训练样本数	1 357	1 253	1 217	1 024	1 600
测试样本数	200	200	200	200	200
准确率	94.3%	97.1%	89.4%	88.2%	85.3%
召回率	92.7%	96.3%	86.6%	85.4%	82.4%
F	93.5%	96.7%	88.0%	86.8%	83.8%

4 结束语

本研究根据中文分类系统的特点,提出了一个基于支持向量机的中文文本分类算法,并对真实语料进行了实验测试,结果表明支持向量机是一种具有较好泛化能力、性能优越的技术.同时该算法达到了令人满意的效果.

参考文献:

[1] 高洁,吉根林.文本分类技术研究[J].计算机应用研究,2004(7):28-30.

[2] Joachims T. Text categorization with support vector machines[C]//in Proc of European Conference on Machine Learning(ECML).[S.l.]:[s.n.],1998.

[3] Deerwester S, Dumais S T A. Indexing by latent semantic analysis[J]. Journal of the Society for Information Science, 1990,41(6):391-407.

[4] Salton G. Developments in automatic text retrieval[J]. Science, 1991, 253(23):974-980.

[5] Salton G, Buckley C. Term weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988,24(5):513-523.

[6] Church K W, Gale W A. Inverse document frequency (IDF):A measure of deviations from Poisson[C]//in Proceedings of the 3rd Workshop on Very Large Corpora. Boston:[s.n.], 1995:121-130.

[7] Vapnik V N. 统计学习理论的本质[M].张学工,译.北京:清华大学出版社, 2000.

[8] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.

[9] Pang-Ning Tan. 数据挖掘导论[M].范明,范宏健,译.北京:人民邮电出版社, 2006.

[10] 牛强,王志晓,陈岱,等. 基于SVM的中文网页分类方法的研究[J].计算机工程与设计, 2007,28(8):1893-1895.

(责任编辑 陈 松)