统计学理论在邮件分类中的应用研究

汤 伟,程家兴,纪 霞
(安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:分类问题,尤其是文本自动分类一直是机器学习与数据挖掘研究中的研究热点与核心技术,其中如朴素贝叶斯、KNN等近年来得到了广泛的关注和快速的发展。文中在统计学理论的基础上给出了一种基于支持向量机方法的文本分类算法,并设计出了相应的垃圾邮件过滤系统。实验证明与朴素贝叶斯方法相比,该算法极大地提高了分类准确率和查全率,具有应用推广的价值。

关键词:机器学习;文本分类;垃圾邮件
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2008)12-0231-04

Research and Design of a Spam Filtering System Based on Statistical Learning Theory

TANG Wei, CHENG Jia-xing, JI Xia
(Ministry of Education Key Lab. of Intelligent
Computing & Signal Processing, Anhui University, Hefei 230039, China)

Abstract: Classification is one of the most important research fields in data mining and machine learning. In recent years, there have been extensive studies and rapid progresses in automatic text categorization. Proposes a SVM text categorization on the basis of statistic theory, and designs a corresponding spam email filtering system. Compared with the naive Bayes, the validity of this system is proved. At last some future directions of the research are given.

Key words: machine learning; text classification; spam

0 引 言

信息技术和网络的飞速发展,使信息产生和传播的速度较之十几年前有了巨大的提高。无论是政府、企业等机构还是个人每天都在接收和发送大量的数据。在信息技术给人们带来极大便利的同时,海量的信息也在一定程序上妨碍了人们对信息的有效识别和获取。如何从大量的数据中准确及时地找到所需的信息,甚至从中寻找出规律来对未来数据进行预测是数据挖掘中一个重要的研究领域。其中基于统计学理论基础的机器学习便是数据挖掘技术中的一项重要研究课题。作为当前信息的一个主要表现形式——文本信息具有类别多样、数据噪音多、样本易采集等特点,通过研究机器学习方法对文本信息进行处理既具富有挑战性也具有很大的实用价值。

统计学是机器学习方法的重要理论基础之一,支持向量机(Support Vector Machine,SVM)是20世纪90年代中期在统计学习理论基础上发展起来的一种机器学习方法。它采用结构风险最小化准则(Structural Risk Minimization,SRM)训练学习[1,2],将学习问题归结成为一个凸二次规划问题,通过非线性变换将数据映射到高维特征空间,使数据在高维空间中可以用线性判别函数分类巧妙地解决维数问题,算法复杂度与样本维数无关具有简洁的数学形式和直观的几何解释,人为设定的参数少,便于理解和使用。支持向量机建立在严格的理论基础之上,较好地解决了非线性、高维数、局部极小点等问题。

1 统计学理论与支持向量机分类

机器学习的目的可以用图1简单地表示,即根据给定的训练样本求系统输入输出之间的依赖关系,一般地可表示为变量 y 与 x 之间存在的未知的依赖关

收稿日期:2008-04-02
基金项目:国家自然科学基金(60273043);安徽大学研究生创新基金(20073053)
作者简介:汤 伟(1982-),男,安徽肥西人,硕士研究生,研究方向为机器学习、智能计算;程家兴,教授,博士生导师,研究方向为智能计算、算法分析及设计及最优化方法。

系。

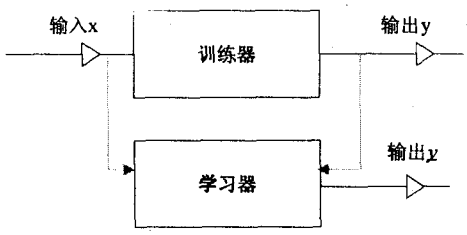

图 1 机器学习过程示意图

学习的问题就是从给定的函数集 $f(x,a)$，$a\in\Lambda$（Λ 表示参数集合）中选择能够最好逼近训练器响应的函数。训练集由根据联合概率 $F(x,y)$ 抽取出的 l 个独立同分布样本：$(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)$ 组成，通过训练从一组函数$\{f(x,a)\}$中选出的最优函数 $f(x,a_0)$，并对其进行依赖关系估计，使期望风险最小[3]。

$$R(a)=\int L(y,f(x,a))dF(x,y) \tag{1}$$

上面的期望风险公式中，$\{f(x,a)\}$称作预测函数集，$a\in\Lambda$（Λ 表示参数集合）。$L(y,f(x,a))$为由于用 $f(x,a)$对 y 进行预测而造成的损失。在传统的学习方法中，采用了所谓的经验风险最小化(Empirical Risk Minimization ERM)准则，即用样本定义经验风险：

$$R_{emp}(a)=\frac{1}{l}\sum_{i=1}^{l}L(y_i,f(x_i,a)) \tag{2}$$

通过设计学习算法经训练使得 $R_{emp}(a)$最小化，作为对式(1)的估计。

但在实际应用中，可以发现经验风险(式(2))最小不等于期望风险(式(1))最小，不能保证分类器的推广能力，即不能取得小的实际风险。经验风险只有在样本数无穷大时才能趋向于期望风险，需要非常多的样本才能保证分类器的性能，需要找到经验风险和推广能力的平衡点[4]。

支持向量机是由 Vapnik 领导的 AT&T Bell 实验室提出的一种新的非常有潜力的知识发现方法，它开辟了学习高维数据新的天地，在分类方面具有良好的性能[5]。

在线性可分情况下(见图 1)，设两类问题训练样本集为$(x_1,y_2),(x_2,y_2),\cdots,(x_l,y_l)$，其中 $x_i\in R^n$，$y\in\{+1,-1\}$，l 为样本数，存在着超平面$(w\cdot x)+b=0$，使得训练样本中的正类输入和负类输入分别位于该超平面的两侧。其中“·”是向量点积。分类如下：

$$(w\cdot x_i)+b\geqslant 0,y_i=+1 \tag{3}$$

$$(w\cdot x_i)+b<0,y_i=-1 \tag{4}$$

w 是超平面的法线方向，最优超平面就是训练数据在无误关的划分前提下，使得每一类数据都与超平面距离最近的微量与超平面之间的距离最大。在线性可分的情况下，求解最优超平面的问题就是求解二次型规划的问题[6]，在满足约束条件：

$$y_i[(w\cdot x_i)+b]\geqslant 1,i=1,\cdots,l \tag{5}$$

根据给定的训练样本找到权向量 w 与偏移 b 的最优值，获得最小化泛函：

$$\Phi(w)=\frac{1}{2}||w||^2 \tag{6}$$

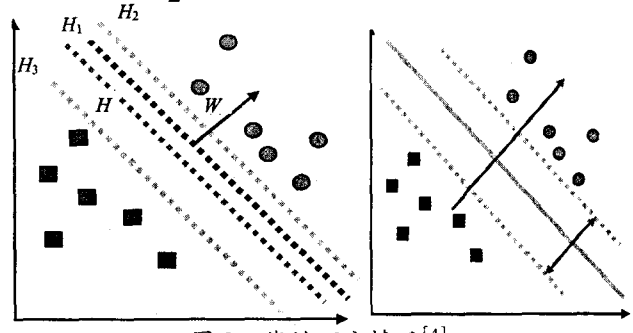

图 2 线性可分情况[4]

这是一个二次优化问题，可在满足约束条件(式(5))的情况下，使用 Lagrange 乘子法求解。

$$\min L(w,b,a)=\frac{1}{2}||w||^2-\sum_{i=1}^{l}a_i[y_i(w^{\mathrm{T}}\cdot x_i+b)-1] \tag{7}$$

通过求偏导，代入可得到求解原问题的对偶式，求解对偶式

$$\max Q(a)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i,j=1}^{l}a_ia_jy_iy_j(\vec{x_i}\cdot\vec{x_j}) \tag{8}$$

$$\text{s.t. }\sum_{i=1}^{l}y_ia_i=0,0\leqslant a_i\leqslant C,i=1,\cdots,l$$

b_0 可由约束条件 $a_iy_i[(w\cdot x_i)+b]-1=0$ 求得，其中 $a_i>0$ 的样本即为支持向量，可得到分类函数：

$$d(x)=\sum_{i=1}^{l}y_ia_i^*(\tilde{x}\cdot x_i)+b^* \tag{9}$$

其中 $\tilde{x}$ 是输入的样本，根据 $d(x)$就可以确定 $\tilde{x}$ 的归属。

在线性不可分的情况下(见图 3)，Vapnik 等提出了用核函数的方法解决，基本思想是选择非线性映射 $\phi(x)$将样本 x 映射到高维特征空间 Z，在 Z 中构造最

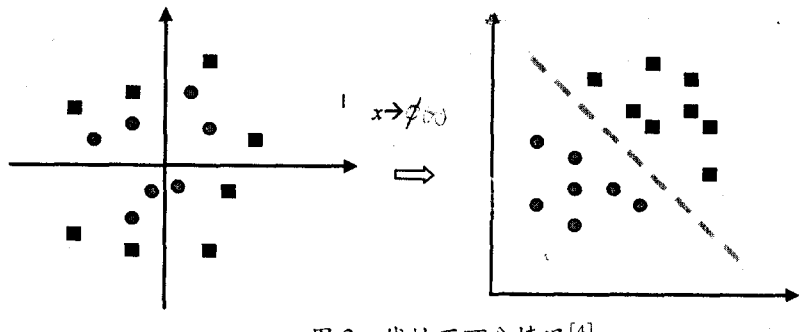

图 3 线性不可分情况[4]

优超平面,此时的目标函数式变为

$$\max\ Q(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\vec{x}_i,\vec{x}_j) \quad (10)$$

分类函数也变为

$$d(x)=\sum_{i=1}^{l}y_i a_i^{*}K(\bar{x},x_i)+b^{*} \quad (11)$$

2　分类算法设计

2.1　文本分类

按 Pawlak 的说法,人的智能主要就是分类能力。文本分类是指按照预先定义的类别,为文本集合中的每一个文本确定一个类别,文本分类是文本挖掘的一个重要组成部分。长期以来,文本分类都是数据挖掘领域的一个重要的研究方向。笔者进行文本分类研究的目的则是对垃圾邮件的识别和过滤,将未知邮件分入两类 H类(有效邮件),S类(垃圾邮件)中的一类。

2.2　文本表示

在文本处理领域,通常采用向量空间模型(VSM, Vector Space Model)表示文本,一篇文本可以表示为一个 n 维向量$\{w_1,w_2,w_3,\cdots,w_n\}$,其中 $w_i(i=1,2,\cdots,n)$ 表示 VSM 的第 i 个特征项的权值,特征项可以是字、词、短语或者某种自定义的概念单位,但常用的文本分类方法大多采用词作为特征项,文中就是采用词作为特征项。权重有多种计算方法,最简单的是布尔权重(也叫二值权重),即权重为 1(该特征项在文本中出现)或者为 0(该特征项没有在文本中出现);更通常的情况下,VSM中的权重计算采用词频(TF:表示该特征词在文本中出现的次数)和文档频次(DF:表示在所有的文本中出现该特征词的文档数量)的某种组合来进行计算。

所要做的工作就是要通过学习,将一篇使用 VSM 表示的邮件 $m_i=\{w_1,w_2,w_3,\cdots,w_n\}$ 在保证准确率的情况下准确划分到 H类或是 S类中去。

2.3　特征向量降维

在对邮件内容进行向量空间表示过程中,如果将所有的词、句都作为特征进行处理,会使最终需要处理的向量维数非常大,会使计算量和存储空间急剧增加,处理速度慢,失去实际的使用价值。而且在实验中发现,特征向量粒度并不是越小越好,如果按字为单位进行提取,分类效果很差甚至不如将文本处理后以二个字为单位切分提取。

在这里,使用一种信息增益方法(Information Gain)来提取特征向量,信息增益方法是指在过滤问题中用于度量已知一个特征是否出现于某一主题相关文本中对于该主题预测有多少有用信息,通过计算信息增益可以得到那些在正例样本中出现频率高而在反例样本中出现频率低的特征,以及那些在反例样本中出现频率高而在正例样本中出现频率低的特征,这种方法的缺点是会增加一些统计计算量,信息增益的计算公式如下:

$$IG(t)=-\sum_{i=1}^{|c|}P(c_i)\log P(c_i)+P(t)\sum_{i=1}^{|c|}P(c_i\mid t)\log P(c_i\mid t)+P(\bar{t})\sum_{i=1}^{|c|}P(c_i\mid\bar{t})\log P(c_i\mid\bar{t}) \quad (12)$$

$P(\bar{t})$ 表示词 t 不出现的概率,$\sum_{i=1}^{|c|}P(c_i\mid t)$ 表示词 t 出现的情况下文本属于 c_i 类的概率,$\sum_{i=1}^{|c|}P(c_i\mid\bar{t})$ 表示词 t 不出现的情况下文本属于 c_i 类的概率。下面的公式中相应变量的含义与此相同。$IG(t)$ 的值反映了该词为整个分类所提供的信息量。

除此之外,可以将一些样本中常用的联接词、符号(除了叹号外)在向量描述的时候忽略掉,尽量降低训练的计算量和耗费的时间。

分类算法流程见图 4。

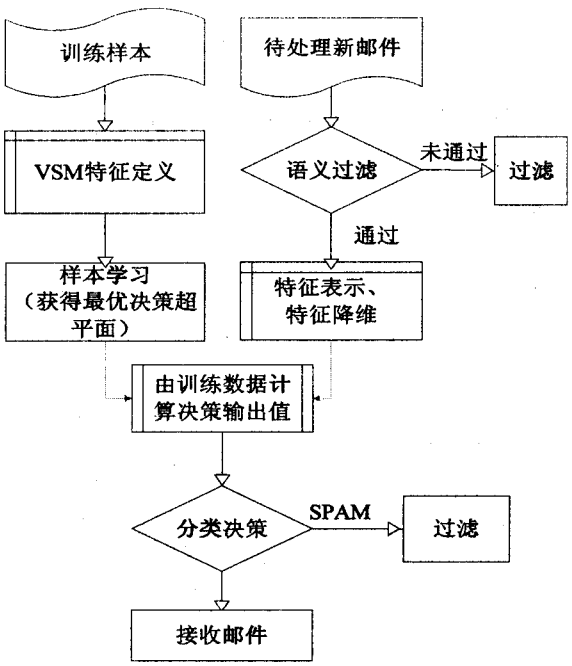

图 4　分类算法流程

2.4　实验与分析

在实验中采用两份数据集,一份是实验中收集的 1000 封邮件 KeyLib,另外一份是 CCERT 公布的 June 2005 邮件语料库,它是由 http://www.ccert.edu.cn 提供给非商用用途使用的测试邮件,其中包含了 25088 封垃圾邮件和 9272 封非垃圾邮件,在里面抽取了 2000 封垃圾邮件和 2000 封非垃圾邮件作为训练样本集,又随机抽取了 200 封作为测试样本。将二份邮件分别分成二部分,75% 作为训练集,25% 作为测试集。评判的效果使用下列一些指标[7]:

(1)查全率 R (Recall)。Recall = 正确过滤掉的邮件数/应该过滤掉的垃圾邮件。数值越高,表示漏网的垃圾邮件就越少。

(2)准确率 P (Precision)。Precision = 正确过滤掉的邮件数/实际过滤掉的邮件。数值越高,表示将合法邮件误判为垃圾邮件的可能性越小。

(3)F 测试值。$F = (2 * R * P)/(R + P)$。

实验结果见表 1。

表 1 邮件过滤系统实验测试结果(%)

数据集	查全率	准确率	F 值
KeyLib (1000)	96.2	90.3	93.156
CDSCE (1830)	94.5	87.1	90.649

实验显示,训练样本的选择的典型性直接影响着最终的过滤效果,与之前在采用朴素贝叶斯方法进行邮件分类相比较,在查全率和准确率上都有一定的改善,基本上可以满足垃圾邮件的过滤任务。

3 结束语

在目前广泛取得应用的大多数客户端垃圾邮件过滤系统中(如 FoxMail 6.0 等)都是采用的朴素贝叶斯方法[8],朴素贝叶斯方法的优点是算法简单、执行速度快,但对样本的要求很高,而且分类精度上提升空间有限。文中给出了一种基于统计理论的垃圾邮件过滤方法——支持向量机进行垃圾邮件分类,它具有严格的理论基础,又能较好地解决小样本、非线性、高维数和局部极小点等实际问题。最后通过实验证明这种方法在垃圾邮件分类中已经可以很好地完成分类任务,确定系统中的过滤精度,在查全率与准确率上较之朴素贝叶斯分类有了很大的提升。但是支持向量机在计算时间和空间要求较高,这也从一定程序上影响了它的应用推广。如何在保持分类精确度的前提下,提高算法的效率,这是下一步需研究的课题。

参考文献:

[1] Christopher J, Burges C. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998(2):121 - 167.

[2] Hsu Chih - Wei, Lin Chih - Jen. A comparison of methods for multi - class support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2):415 - 425.

[3] 史忠植. 知识发现[M]. 北京:清华大学出版社, 2002.

[4] Ronnie, Rifkin R. Improving Multi - class Text Classification with the Support Vector Machine[DB/OL]. 2002 - 05 - 23. www.ai.mit.edu / research / abstractabstracts2001 /machine - learning.

[5] 安金龙,王正欧. 一种新的支持向量机多类分类方法[J]. 信息与控制, 2004(3):262 - 267.

[6] 王 斌,潘文锋. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报, 2005, 19(5):1 - 3.

[7] 曹麒麟. 反垃圾邮件的研究[D]. 北京:清华大学电子工程系, 2002.

(上接第 227 页)

[5] 陶树平,钱 挺. 一种网格平台数据挖掘服务模式及其算法[J]. 计算机工程, 2005(5):109 - 111.

[6] 佟 强. 科学数据网格中数据挖掘技术研究[D]. 北京:中国科学院计算技术研究所, 2006:23 - 41.

[7] 吕 品,陈年生,董武世. 一种网格数据挖掘应用系统的设计[J]. 计算机技术与发展, 2007, 17(1):158 - 160.

[8] 侯文国,傅秀芬,谢翠萍. 网格的数据挖掘[J]. 计算机应用研究, 2004(10):241 - 243.

[9] 林 雯,段小斌,谢晓兰. 数据挖掘技术在中小型制造企业 CRM 中的应用[J]. 计算机技术与发展, 2007, 17(11):247 - 250.

[10] 刘天鹏,周 娅. P2P 网络中的数据挖掘[J]. 计算机应用, 2008(1):162 - 164.

(上接第 230 页)

5 结束语

用参数化的分形树与随机函数相结合,能生成形态逼真三维树木;在虚拟场景中进行复杂的树木建模时,根据不同的距离选择不同的模型,从而加快了渲染的速度。如何加入更多影响树木形态的参数,例如光照、树木的自重等,使树木的逼真度提高;如何使树木在风的作用下摇曳,都是值得进一步研究的课题。

参考文献:

[1] 王永皎,莫国良. 植物的三维建模研究进展[J]. 计算机应用研究, 2005(11):1 - 4.

[2] Weber J, Penn J. Creation and Rendering of Realistic Trees [J]. SIGGRAPH, 1995, 64(8):119 - 127.

[3] 康孟珍,De Reffye P,胡包钢,等. 快速构造植物几何结构的子结构算法[J]. 中国图形图像学报, 2004, 9(3):79 - 86.

[4] 李庆忠, 韩金姝. 基于 IFS 的树木形态模拟真实感的研究[J]. 微机发展(现更名为:计算机技术与发展), 2005, 15(7):86 - 88.

[5] 孙家广. 计算机图形学[M]. 北京:清华大学出版社, 2002:369 - 373.