基于优化的多核学习方法的 Web 文本分类的研究

江 伟[1,2],潘 昊[1]
(1. 武汉理工大学 计算机科学与技术学院,湖北 武汉 430070;
2. 武汉科技大学城市学院 信息工程学部,湖北 武汉 430083)

摘 要:Web 文本分类技术是数据挖掘中一个重要研究领域,为了能从海量信息中快速检索遍布网络各处的文档,需要提高 Web 文本分类技术的性能。多核学习方法是当前机器学习领域的一个热点,可以显著提升分类识别能力和学习推广能力,而核方法是解决高维非线性模式分析的有效方法之一。利用多核代替单核能增强决策函数的可解释性并获得更优的性能。文中分析研究了一种基于优化的多核学习的支持向量机,在此基础上结合通用的 Web 文本分类模型,提出了一种基于多核学习支持向量机的 Web 分类方法。通过实验测试表明,该方法具有良好的效果,对比一致组合的多核学习方法,所提出的方法具有较高的准确率。

关键词:支持向量机;数据挖掘;多核学习;Web 文本分类
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2013)10-0080-03
doi:10.3969/j.issn.1673-629X.2013.10.020

Research of Web Document Classification Based on Optimized Multiple Kernel Learning Method

JIANG Wei[1,2], PAN Hao[1]
(1. College of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China;
2. Department of Information Engineering, City College of Wuhan University of Science and Technology, Wuhan 430083, China)

Abstract: Web document classification has been considered as an important research field in data mining, it's necessary to improve the performance of technique of Web document classification for quickly retrieving the documents from the massive information spread all over the network. Multiple-kernel learning is a focus in current machine learning community, which is able to develop the capability of classification and learning extension, while kernel method is one of effective approaches for solving high dimension and non-linear pattern analysis. By using the advantage of multiple kernel can boost interpretability of decision function and obtain better performance. In this paper, propose a Web document classification based on multiple kernel learning after a research of a SVM based on multiple kernel learning. According to the result of the experiment, this approach presented in this paper has high efficiency and more accurate rate compared with simple consistent combination multiple kernel learning method.

Key words: SVM; data mining; multiple kernel learning; Web document classification

0 引 言

随着互联网络信息的爆炸式增长,大范围地从网络上检索文档变得愈发困难,人们需要能从大规模的信息中准确迅速地获取有用信息,所以就需要提高 Web 文档分类系统的性能。由 Vapnik 等人提出的基于统计学习理论的支持向量机是一种新的机器学习方法[1],体现了结构风险最小化(Structural Risk Minimization, SRM)原则,使其成为了当前机器学习和数据挖掘领域的重要工具[2]。它集成了最大间隔超平面凸二次优化、稀疏解和松弛变量等多项技术,在有限样本

收稿日期:2013-01-11 **修回日期**:2013-04-16 **网络出版时间**:2013-07-24
基金项目:湖北省自然科学基金(2011CDB257)
作者简介:江 伟(1980-),男,湖北随州人,讲师,博士研究生,研究方向为智能计算、数据挖掘、Web 应用;潘 昊,教授,博士生导师,研究方向为智能计算、数据挖掘。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130724.0945.005.html

信息上的机器学习能力、复杂性以及泛化能力都有显著优势。而且,通过引入核函数,利用核函数良好的非线性扩展性能,可以很容易地在核空间里把类似 SVM 这样的线性算法转化为非线性算法[3],如核主成分分析(Kernel PCA)、核聚类分析和核 Fisher 判别分析(Kernel FDA)等,这些算法都是在传统的算法之上结合 SVM 的成功应用,对核方法进行的大量推广和改进,使得基于核方法的应用渗透到机器学习的众多领域。多核学习提出了一种多个基核(basis kernel)的组合形式,使得它和单核 SVM 相比具有许多优异性能,如核函数的自动选择、预测性能的提升等,进一步提高了分类器的学习效率和准确率。文中提出了一种高性能多核学习支持向量机,将其应用于 Web 分类系统,并且通过实验结果表明,该方法具有良好的效果。

1　Web 文本分类的原理

Web 文本分类是根据事先定义好的分类体系,按照相关的主题类别或者属性信息,把待归类的文档归入一个或多个类别中。所以可以把这种分类过程看作是一个映射,即从页面文本属性到归类类别空间的映射过程,这种映射可以是一对一或一对多映射,这是一种典型的机器学习过程。文本分类的过程,用数学方式可以这样描述:设文档集表示为 $D=\{d_1,d_2,\cdots,d_i\}$,预定义的类集为 $C=\{c_1,c_2,\cdots,c_n\}$,这样就可以确定一种映射 $\varphi:D\times C\rightarrow\{T,F\}$,其中 $F=\{1,-1\}$。若文本 d_i 属于类别 C,则 F 的值为 1,否则为 -1。典型的基于机器学习的 Web 文本分类模型一般由测试文本集和训练文本集两部分组成,在进行必要的文本数据结构化后,在训练阶段,设定每个训练样本集的归属类别,然后通过训练文本集学习得到一个分类器,最后对测试集文档按照一定的算法用事先得到的分类器进行分类。在这个过程中建立学习模型准确率的评估反馈,以进一步提高分类器的性能。具体的 Web 文档分类模型如图 1 所示。

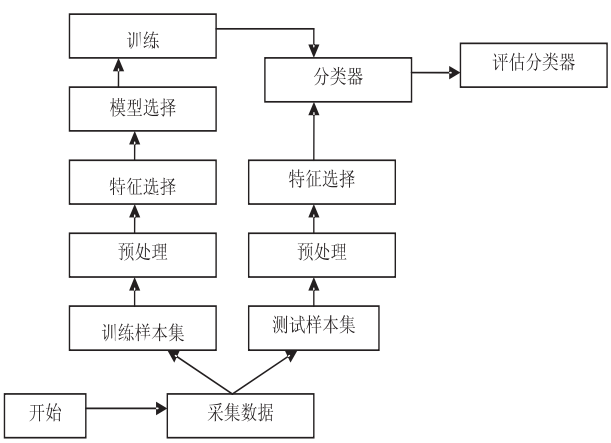

图 1　Web 文档分类模型

在进行分类之前,必须把以 HTML 网页格式形成的非结构化或半结构化文档进行预处理,抽取其中的特征信息,表示成结构化数据,以便于进行分类学习。向量空间模型(Vector Space Model)是目前常用到的并且较成熟的文本表示方法,在文中的应用中采用该模型来表示 Web 文本。首先,去除掉 HTML 的标记元素和对分类没有影响的虚词、感叹词等,根据特征词集进行分词。然后,假设文档中出现的词条没有先后顺序,用特征项 t_i 代表词条,为表示该特征项在文档中的重要性和特征表达能力,赋予一个权重 w_i,这样一篇文档就可以转换为由组成它的词条所对应的特征项和对应的权重来表示:$V=(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$。其中,权重可以表示为特征项在文档 d 中出现频率的函数即:$w_i(d)=\varphi(f_i(d))$,这里用 TF-IDF 函数表示[4]:

$$w(t,d)=\frac{f(t,d)\times\log(N/n_t+0.01)}{\sqrt{\sum[f(t,d)\times\log(N/n_t+0.01)]^2}}$$

其中,$w(t,d)$ 表示词条 t 在文本 d 中的权重;$f(t,d)$ 表示词 t 在文档 d 中的频率;N 为训练文本的总数;n_t 表示训练集中出现词条 t 的文本数。

2　基于多核学习 SVM 的 Web 文本分类

2.1　多核学习方法中参数的计算方法

核函数提供了在高维空间的线性算法的解决方法,其最大的特点就是对于像 SVM 这样总是以点积形式表示的线性算法,映射到高维特征空间实现点积运算,这样就不必关心数据样本在特征空间中的具体表现形式,从而得到非线性分类的基于核函数的 SVM。但是,当面对不同的应用问题时,就需要能从不同的角度来描述数据空间,而不同的核函数在对数据空间进行映射得到的特征空间上,得到的结果是迥然不同的。这就需要依靠模型设计人员依靠领域经验来选择不同的核函数。基于这样的问题,现在提出并发展了许多的多核学习方法[5],其目的就是希望能通过构造出具有多个核函数的模型,来更好地表达针对具体问题的特征空间。基核函数的凸组合模型是目前最常用的方法,如基于半定规划(Semi Definite Programming,SDP)的学习方法[6]、基于二次约束性二次规划(Quadratically Constrained Quadratic Program,QCQP)的学习方法[7]、基于半无限线性规划(Semi-Infinite Linear Program,SILP)的学习方法[8],以及简单多核学习(simple MKL)的方法[2]。在这些多核学习模型中,为了求解得到凸组合的系数,通常采用的都是基于最大化间隔准则的方法,而最大化间隔准则容易导致缩放问题和初始值问题[9]。所以在文中引入一种基于 SVM 泛化误差界的精确高效构造多核模型(GMKL)的方法[10],

建立更灵活、性能更优的核函数。概括来说,首先基于留一法(leave-one-out)错误界给出多核学习优化形式,然后推导出该目标函数对于组合系数微分的计算公式,最后应用标准的投影梯度算法来求解该优化问题。在文中采取如下的核函数形式:

$$K_s = \sum_{i=1}^{n} \theta_i K_s, \text{s.t. } \theta_i \geqslant 0, \sum_{i=1}^{n} \theta_i = 1$$

其中,K_s 为基核函数,n 为基核个数。

采用基于最小半径间隔界的方法。当阈值为 0,即 $b = 0$ 时,以下不等式成立:

$$L((x_1, y_1), \cdots, (x_l, y_l)) \leqslant \frac{R^2}{\gamma^2} =: T_{\text{RM}} \tag{1}$$

其中,R 为特种空间中包含所有训练数据点的最小半径;γ 为间隔,即

$$R^2 = \min_{t,c} t, \text{s.t. } t \geqslant \| \varphi(x_i) - c \|_2^2 \tag{2}$$

$$\gamma = \min_{(x_i \cdot y_i) \in D} \frac{y_i(w \cdot \varphi(x_i) + b)}{\| w \|_2} \tag{3}$$

其中,$w = \sum_{i=1}^{l} \alpha_i^0 \varphi(x_i)$。由于(2)式的对偶形式为

$$\max_{\beta} \sum_{i=1}^{l} \beta_i K_\theta(x_i, x_j) - \sum_{i,j=1}^{l} \beta_i \beta_j K_\theta(x_i, x_j)$$

$$\text{s.t } \sum_{i=1}^{l} \beta_i = 1, \beta_i \geqslant 0 \tag{4}$$

因而,可以通过求解上述优化问题计算 R^2。

$$\max_{\theta} T_M := \gamma^2, \text{s.t. } \theta_i \geqslant 0, \sum_{i=1}^{m} \theta_i = 1 \tag{5}$$

从上述公式(5)可以看出,基于最大化间隔准则在优化求解参数 θ 时,由于 R 和 θ 的相关性,使得当 γ^2 最大时,R 不一定最小,从而也就无法保证SVM的泛化性[11]。所以针对这一问题,考虑直接最小化留一法错误上界,即最小化 T_{RM} 来确定组合系数 θ。由此得到基于超球半径界的计算公式:$\min_{\theta} T_{\text{RM}} := \frac{R^2}{\gamma^2}, \text{s.t. } \theta_i \geqslant 0, \sum_{i=1}^{m} \theta_i = 1$。从该式可以看出该方法能保证SVM的泛化性能。最后,基于 T_θ 的微分,采用投影梯度算法确定梯度下降方向,并使用一维线性搜索确定步长,更新组合系数 $\theta^{t+1} = \theta^t + \gamma_t D_t$,其中,$D_t$ 为梯度下降方向;γ_t 为下降步长。

2.2 基于 SVM 的 Web 文本分类方法

典型的支持向量机是二分类器,所以为解决 Web 文本分类这样的多分类问题,采取结合决策树来实现多分类。在初始化过程中,选择 10 个高斯核函数:$k_G(x, x') = \exp(-\frac{\| x - x' \|_2^2}{2\sigma^2}), \sigma \in \{0.5, 1, 2, 5, 7, 10, 12, 15, 17, 20\}$ 作为基核,具体的算法步骤如下:

Step1:初始化过程。

Step2:训练过程。先初始化参数 θ,令迭代次数 $t = 0$,更新组合系数 $\theta^{t+1} = \theta^t + \gamma_t D_t$,直到满足迭代停止要求。再根据得到的多核模型 K_θ,用随机抽取的训练文本集对基于多核 K_θ 的 SVM 训练得到最优分类超平面的支持向量集,形成判别函数。

Step3:判决过程。将测试集输入多核 K_θ 的 SVM 得到二分类结果。

Step4:二叉决策树分类过程。在二类可分的基础上形成二叉决策树,利用分类判决函数得出多类可分的结果。

Step5:输出分类结果。

3 实验测试与分析

文中的实验数据是从互联网上下载的 2 100 篇中文网页,先人工分为体育、游戏、影视、新闻、音乐五个类别,然后任意抽取其中的 1 405 篇作为训练样本,其余的作为测试集,根据前述的 Web 文档信息预处理方法得到 27 869 个词条。为测试系统的分类质量,采用如下公式:测试值 $= \frac{\text{精确率} \times \text{召回率} \times 2}{\text{精确率} + \text{召回率}}$ 得到的结果作为考察标准,其中精确率和召回率能体现分类质量的两个不同方面。用上述数据和算法进行训练和测试,统计结果如表 1 所示。从实验结果可以看出采用多核学习的支持向量机的 Web 分类效果良好。

表 1 Web 文本分类文档测试结果

种类	游戏	影视	体育	新闻	音乐
训练总样本数	315	426	231	205	228
训练时间/s	6.4	7.3	6.6	5.4	5.9
精确率/%	95.2	94.4	96.4	96.7	95.4
召回率/%	92.2	91.6	93.1	93.4	94.5
测试值	92.1	92.6	93.2	93.5	94.1

4 结束语

多核方法已经成为当前核机器学习的热点,尤其是在解决一些复杂问题时,组合核函数比单核函数具有更优的性能。文中对一种优化的融合多核函数的算法进行了介绍和分析,进而提出了一种基于优化的多核学习的支持向量机,将其应用于 Web 文本分类,并对真实文本进行了分类测试,结果表明该方法是一种性能比较优秀的分类方法,能较好地满足 Web 应用上对数据挖掘的需求。

参考文献:

[1] Vapink V N. The Nature of Statistical Learning Theory[M]. New York:Springer Verlag,1995.

[2] Rakotomamonjy A, Bach F, Canu S. Simple MKL[J]. Journal of Machine Learning Research,2008(9):2491-2521.

(下转第 86 页)

图 2 k-均值和 LDA-KM 聚类结果

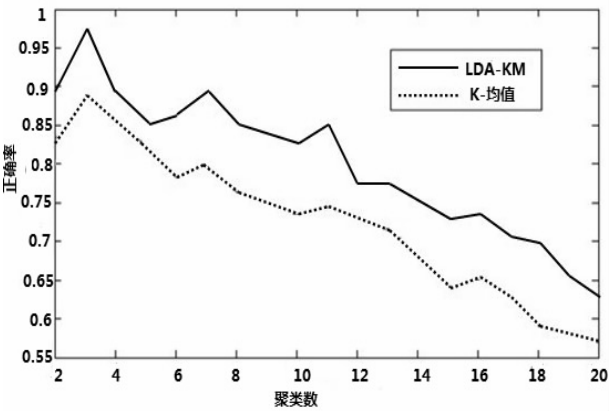

图 3 不同聚类数上,k-均值和 LDA-KM 聚类的正确率比较结果

应子空间选择方法。该方法采用反复迭代的方式,用无监督学习的方法找到具有最佳的可分离性的子空间和数据的类别,实验结果表明该方法的聚类效果优于传统的 k-均值聚类。然而,还有一些问题需要进一步的研究。例如,如何确定最佳的聚类数?如何选择合适的降维方法确定最初的低维子空间?在以后的工作中,将对这些问题进行研究。

参考文献:

[1] 贺 玲,吴玲达,蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究,2007(1):10-13.

[2] 于洪涛,段军义,杜照丰. 一种基于聚类技术的个性化信息检索方法[J]. 计算机工程与应用,2008,44(8):187-188.

[3] 李旭超,刘海宽,王 飞,等. 图像分割中的模糊聚类方法[J]. 中国图象图形学报,2012,17(4):447-458.

[4] Jolliffe T. Principal component analysis[M]. New York: Springer-Verlag,1986.

[5] Cox T,Cox M. Multidimensional scaling[M]. London: Chapman-Hall,1994.

[6] Roweis S T,Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science,2000,290:2323-2326.

[7] De la Torre F,Kanade T. Discriminative cluster analysis[C]//Proc of International Conference on Machine Learning. New York: ACM,2006:241-248.

[8] Ding C,Li T. Adaptive dimension reduction using discriminant analysis and k-means clustering[C]//Proc of International Conference on Machine Learning. New York: ACM,2007:521-528.

[9] Fukunaga K. Introduction to statistical pattern recognition[M]. Boston: Academic Press,1990.

[10] 周爱武,于亚飞. K-Means 聚类算法的研究[J]. 计算机技术与发展,2011,21(2):62-65.

[11] 黄 韬,刘胜辉,谭艳娜. 基于 k-means 聚类算法的研究[J]. 计算机技术与发展,2011,21(7):54-57.

[12] 黄移军. 基于局部线性嵌入法的流形学习[J]. 数学理论与应用,2009,29(4):38-42.

[13] Frank A,Asuncion A. UCI machine learning repository[D]. Irvine,CA: University of California,2010.

(上接第 82 页)

[3] Shawe-Taylor J,Cristianini N. Kernel Methodsfor Pattern Analysis[M]. Cambridge: Cambridge University Press,2004.

[4] 庞剑锋,卜东波,白 硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究,2001(9):23-26.

[5] 汪洪桥,孙富春,蔡艳宁,等. 多核学习方法[J]. 自动化学报,2010,36(8):1037-1050.

[6] Lanckriet G R,Cristianini N,Bartlett P. Learning the kernel matrix with semidefinite programming[J]. Journal of Machine Learning Research,2004(5):27-72.

[7] Bach F R,Lanckriet G R G,Jordan M I. Multiple kernel learning,conic duality,and the SMO algorithm[C]//Proceedings of the 21th International Conference on Machine Learning. New York: ACM Press,2004.

[8] Sonnenburg S,Ratsch G,Schafer C. A general and efficient multiple kernel learning algorithm[C]//Proc of Neural Information Processing Systems. Cambridge: MIT Press,2006.

[9] Gai K,Chen G,Zhang C. Learning kernels with radiuses of minimum enclosing balls[C]//Proc of 24th Annual Conference on Neural Information Processing System. Cambridge: MIT Press,2010.

[10] 刘 勇,廖士中. 基于支持向量机泛化误差界的多核学习方法[J]. 武汉大学学报(理学版),2012,58(2):149-156.

[11] Cortes C,Mohri M,Rostamizadeh A. Generalization bounds for learning kernels[C]//Proceedings of the 27th International Conference of Machine Learning. New York: ACM Press,2010.

基于优化的多核学习方法的Web文本分类的研究

作者： 江伟， 潘昊， JIANG Wei， PAN Hao

作者单位： 江伟,JIANG Wei(武汉理工大学 计算机科学与技术学院，湖北 武汉 430070; 武汉科技大学城市学院 信息工程学部，湖北 武汉 430083)， 潘昊,PAN Hao(武汉理工大学 计算机科学与技术学院,湖北 武汉,430070)

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2013(10)