支持向量机和蚁群算法的网页分类研究

宋军涛[1],周 铜[2],杜庆灵[1]
SONG Jun-tao[1],ZHOU Tong[2],DU Qing-ling[1]

1.河南工业大学 信息科学与工程学院,郑州 450001
2.中州大学,郑州 450052
1.Department of Information Science and Engineering,Henan University of Technology,Zhengzhou 450001,China
2.ZhongZhou University,Zhengzhou 450052,China
E-mail:songjuntao1982@163.com

SONG Jun-tao,ZHOU Tong,DU Qing-ling.Study of categorization of Web-page of support vector machine and ant colony algorithm.Computer Engineering and Applications,2009,45(17):122-124.

Abstract: Web page categorization is the foundation and core problem of web data mining,it is a typical application based on technology of natural language processing and machine learning.It is imperative to find an effective and efficient method for web page categorization.In this paper,a new method is proposed for web page categorization based on ant colony optimization algorithm(ACOA) and support vector machines(SVMs).The experimental results show that the method is effective and robust,only to make up for the use of support vector machines for large sample training set less than the slow convergence with better precision and recall.

Key words: web page categorization;ant colony algorithm(ACA);support vector machine(SVM);recall;precision

摘 要:网页分类技术是Web数据挖掘的基础与核心,是基于自然语言处理技术和机器学习算法的一个典型的具体应用。基于统计学习理论和蚁群算法理论,提出了一种基于支持向量机和蚁群算法相结合的构造网页分类器的高效分类方法,实验结果证明了该方法的有效性和鲁棒性,弥补了仅利用支持向量机对于大样本训练集收敛慢的不足,具有较好的准确率和召回率。

关键词:网页分类;蚁群算法;支持向量机;召回率;准确率
DOI:10.3778/j.issn.1002-8331.2009.17.037 **文章编号**:1002-8331(2009)17-0122-03 **文献标识码**:A **中图分类号**:TP391

1 引言

随着计算机网络技术的快速发展和Web 2.0广泛深入的应用,同时互联网上的信息呈指数级增长,人们可利用的网络信息也越来越多,如何从这些海量的信息中高效地获得所需信息已成为当前必须解决的问题。bayes分类、关联规则、决策树分类、单纯的支持向量机算法[1]等在文本分类中得到广泛的应用并取得了很大的进步,与文本分类最大的不同在于网页属于半结构化数据且数量极大,先前的文本分类算法直接引用到网页分类中,其分类精度和效率受到一定的影响,特别是大规模样本适应性不强,收敛速度慢。支持向量机分类最大的优点在于构造一个最优超平面,且有强大的统计理论基础,分类精度高,因此在分类中得到广泛的应用,但对于大规模样本训练过程中,表现出适应性不强,收敛速度比较慢的不足,为此提出了基于蚁群算法和支持向量机的网页分类方法。

2 SVM基本原理和蚁群算法

2.1 SVM基本原理

支持向量机(Support Vector Machine,SVM)[2]是Vapnik等人于20世纪90年代中期提出的一类新型机器学习方法,其理论基础是统计学习理论。与基于经验风险最小化原理的传统的统计学习方法不同,SVM基于的是结构风险最小化原理。SVM不仅结构简单,而且各种技术性能尤其是推广能力比神经网络等方法有明显提高。标准的SVM的实现涉及求解线性约束的二次规划问题,该问题可以收敛到全局最优解。但当训练数据很多时——这是很多实际问题所碰到的情况,二次规划问题的求解受到存储器容量的限制,而且分类速度也得不到保证,从而限制了SVM在很多问题上的应用。常用的做法是将问题分解为若干个子问题,然后再对这些子问题进行逐一优化。这样做的缺陷是:结果可能只是一个次优解,并且分解的时候可能需要多次分解才能将问题的规模转化为能解决的程度。

基金项目:2007年公安部应用创新计划项目(the Project of Application Innovation of the Ministry of Public Security in 2007 No.2007YYCIZXHNST063)。
作者简介:宋军涛(1983-),男,硕士研究生,主要研究方向为Web数据挖掘;周铜(1962-),男,副教授,主要研究方向为计算机应用、智能信息处理;杜庆灵(1963-),男,教授,博士后,主要研究方向为网络信息安全、智能计算、模式识别。
收稿日期:2008-12-29 **修回日期**:2009-03-02

2.2　蚁群算法简介

蚁群算法[3](Ant Algorithm)由意大利人 M.Dorigo 等人首先提出的,蚁群算法是一个新的启发算法,蚁群算法固有的并发性和可扩充性,使它非常适合用于带约束的二次优化求解问题。在同一时间内,所有影响资源状态的因素都能由一个事情,即信息素描述。进而能够非常简单、快速地获得预测结果。

该算法按照启发式思想,通过信息传媒 Pheromone 的诱导作用,逐步收敛到问题的全局最优解,是求解适应性计算问题的一种有效方法。由于带约束的二次规划求解问题是众所周知的 NP 问题,而大量的实验表明,蚁群算法是一种有效的求解 NP 类问题的新型算法,具有较强的鲁棒性和内在的分布并行性,且易于和其他方法相结合。蚁群算法还有一个很大的优点就是具有可扩展性。所谓可扩展性指的是在原有一个规模 n 的问题上求出最优解后,再增加 m 个节点,可在原有解的基础上快速找到该问题的 $n+m$ 规模的最优解,本文将蚁群算法与支持向量机结合充分保证支持向量机分类精度的同时弥补其针对大规模训练样本收敛慢的不足。

3　ACA 及 SVM 在网页分类中的应用

3.1　网页文档预处理

网页中数据最大的特点是半结构化,数据呈现形式多样化,首先用向量空间模型(SVM)[4-5]表示 Web 页面数据信息,本文采用 TFIDF[6]向量表示方法,并非所有特征对分类都有用,况且不同的特征所起的作用也不一样,预处理阶段最重要的是关键特征提取及降维,直接影响到后续分类的精度。

3.2　SVM 训练及 ACA 优化求解流程

首先将样本集分为训练集(80%)和测试集(20%)两部分,训练集专门用于 SVM 训练,而测试集专门用于评价测试 SVM 分类性能,其中影响 SVM 分类性能的两个重要参数是折衷参数 C 和高斯核函数 $K(x_i,x_j)=\Phi(x_i)\Phi(x_j)$,训练过程中,不断地调整参数 C,同时利用上述蚁群算法对高斯核函数进行优化,最终选取合适的参数值,整个过程利用上述蚁群算法对 SVM 分类器函数优化求解,输出分类结果,整个分类过程如图 1。

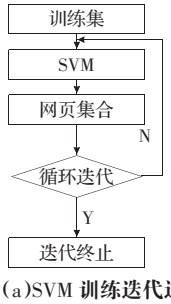


图 1(a)SVM 训练迭代过程

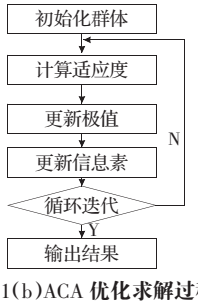


图 1(b)ACA 优化求解过程

3.3　SVM 训练算法描述

(1)假设待训练样本集为 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_i,y_i)\}$,其中 $x_i\in Rn,y_i\in\{1,2,\cdots M\},i=1,2,\cdots,n$。

(2)SVM[7-8]算法的出发点是利用核技巧在高维空间里进行有效的计算,找出支持向量及其系数构造最优分类面。而此最优分类面的构造问题实质上是在约束条件下求解一个二次优化问题,以得到一个最优的决策函数,最优分类超平面为:

$$w^t\Phi(x)+b=0$$

决策分类目标函数为:

$$F(x)=\mathrm{sgn}[w^t\Phi(x)+b]$$

(3)最优分类超平面问题描述为:

$$\min_{w,b,\xi_i}\frac{1}{2}w^t w+C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.}\quad (w^t\Phi(x)+b)\geqslant 1-\xi_i$$

$\xi_i\geqslant 0,i=1,\cdots,n,b$ 是一个阈值。

(4)其对偶最优化问题为:

$$\max_{\alpha}\left\{\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j)\right\}$$

$$\text{s.t.}\quad 0\leqslant\alpha_i\leqslant C,i=1,\cdots,n\qquad\sum_{i=1}^{l}\alpha_i y_i=0$$

α_i 为二次优化问题的最优解,其中,$K(x_i,x_j)=\Phi(x_i)\Phi(x_j)$称为高斯核函数,$n$ 为训练样本个数,C 为惩罚因子,它控制的是训练错误率与模型复杂度间的折衷。容易证明,该优化问题的解中只有一部分(通常是很少的部分)α_i 不为零,这些不为零 α_i 的对应的样本就是支持向量,只有支持向量影响最终的划分结果。

3.4　蚁群算法[9-10]求解步骤

算法优化求解开始时,训练集中的每一个样本点对应一个蚂蚁智能体,从 $t=0$ 时刻开始第一轮搜索,当 $t=1$ 时所有蚂蚁遍历所有样本点,完成了第一轮搜索。此时更新路径上的信息素,并将禁忌表清空。然后开始下一轮搜索,在达到用户设定的最大搜索轮数 NC 时,算法终止,此时找到的最优解作为问题的最优解。

(1)初始化:设定 $t=0$,循环轮次数 $NC=0$,对每条路径设置 $\zeta_i(0)=C$ 和 $\Delta\zeta_{ij}=0$,将 m 只蚂蚁均匀放置 n 个训练集中。

(2)设置 $s=1$(s 是禁忌表的索引),对 $k=1,\cdots,m$,将第 k 只蚂蚁所在的样本点的编号放入禁忌表中。

(3)在禁忌表索引 $s\leqslant n$ 时,重复如下操作:

①$s=s+i$;

②对 $k=1,\cdots,m$,做如下工作:依照公式

$$P_{ij}^{k}(t)=\frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{j\in allowed_k}j[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}$$

计算蚂蚁在样本点 i 选择第 j 个样本点作为下一站的概率,并按概率选择样本点,将蚂蚁移至该样本点,将其编号放入禁忌表中。

(4)对 $k=1,\cdots,m$,做如下工作:计算第 k 只蚂蚁走过的路径长度,记录当前找到的最优解,按公式

$$\Delta\tau^{ij}=\sum_{t}^{t+n}\Delta\tau_{ij}^{k}$$

计算每只蚂蚁的信息素增量。

(5)对所有的点(i,j)根据公式

$$\tau_{ij}(t+n)=\rho\times\tau_{ij}(t)+\Delta\tau_{ij}^{k}$$

更新路径上的信息素。

设置 $t=t+n$;$NC=NC+1$;设置所有的 $\Delta\zeta_{ij}=0$。

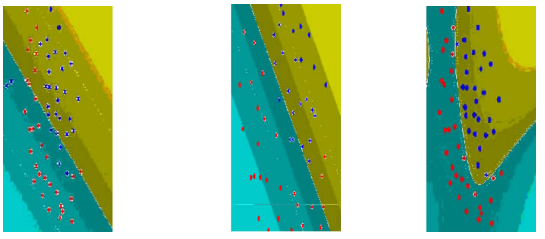
(6)如果 $NC\leqslant NC_{\max}$ 或者没有出现收敛现象,设置它们的禁忌表为空;转到第(2)步;否则,输出最优解,算法结束。

4　实验结果与分析

4.1　实验结果

实验在 MATLAB7.0 环境进行仿真,数据采用 10 000 篇网

页文档(根据网页主题内容共分 6 个大类别),通过预处理后从中随机抽取 20%用于测试评价分类器的性能,剩余 80%用于训练,图 2 是本文方法在训练过程中分别对不同的核函数(多层感知器核函数、多项式核函数图、高斯核函数)进行优化的分类效果图,结果表明本方法选用高斯核函数在网页分类(特别是高维空间训练集)中效率比较高,明显由于其他两种核函数。



(a)多层感知器核函数 (b)多项式核函数 (c)高斯核函数

图 2 对不同核函数进行优化的分类效果图

提出的方法与其他 SVM 分类方法相比,在训练样本规模较小时分类精度没有明显的区别,但随着样本规模的增大,本文的方法在保证分类精度的同时适应性有明显提高,弥补了 SVM 本身针对大样本适应性方面的不足。

4.2 分析评价比较

表 1 提出方法与其他两种方法在准确率、召唤率、宏平均和微平均方面的比较

	准确率	召回率	宏 F1	微 F1
SVM	0.926 83	0.918 27	0.926 52	0.928 41
ACA	0.911 74	0.926 59	0.919 43	0.921 26
ACA+SVM	0.938 36	0.938 28	0.942 62	0.937 53

本文采用传统的评价方法:正确率(P),召回率(R)和 $F1$ 值来衡量分类的效果,其中

$$P_i=\frac{l_i}{m_i}\times 100\% \quad R_i=\frac{l_i}{n_i}\times 100\% \quad F1_i=\frac{2\times P_i\times R_i}{P_i+R_i}$$

l_i,m_i,n_i 分别表示第 $i(i=1,\cdots,6)$类分类结果中正确的网页数目,实际包含的网页数目和结果中出现的数目。微平均计算方法如下公式:

$$P_{micro}=\frac{1}{m}\sum_{i=1}^{m}P_i \quad R_{micro}=\frac{1}{m}\sum_{i=1}^{m}R_i \quad F_{micro}=\frac{2\times P_{micro}\times R_{micro}}{P_{micro}+R_{micro}}$$

5 结束语

本文利用蚁群算法和支持向量机理论相结合,提出了基于蚁群算法和支持向量机的分类方法,对寻求构建高效的网页分类器进行了研究,并与传统的分类算法如 bayes 算法、决策树算法等进行分析比较,结果表明该方法应用在网页分类中在有效性、准确性方面有明显的改善,尤其在大规模训练集中适应性更强。该方法在提高搜索引擎的速度和质量方面有很大的应用空间,是进一步学习研究的重点。

参考文献:

[1] Rossi F,Villa N.Support vector machine for functional data classification[J].Neurocomputing,2006,69:730-742.

[2] Vapnik V.Universal learning technology:Support vector machines[J].NEC Journal of Advanced Technology,2005(2):137-144.

[3] 段海滨.蚁群算法原理及应用[M].北京:科学出版社,2005.

[4] Chen Pai-Hsuen,Lin Chih-Jen.A tutorial on V-support vector machines[J].Applied Stochastic Models in Business and Industry,2005,21(2):111-136.

[5] Chang Chin-chang,Lin Chi-hjen.LIBSVM:A library for support vector machines[J/OL].(2005-10).http://www.csie.ntu.tw/~cjlin/libsvm.

[6] Mlademic D,Grobelink M.Feature selection on hierarchy of Web documents[J].Decision Support System,2003,35(1):45-87.

[7] 贾洞,梁久祯.基于支持向量机的中文网页自动分类[J].计算机工程,2005(5):18-22.

[8] 牛强,王志晓.基 SVM 的中文网页分类方法的研究[J].计算机工程设计,2007(4):34-36.

[9] 朱刚,马良.TSP 问题的蚁群算法求解[J].计算机工程与应用,2007,43(10):79-80.

[10] 彭涛.基于粒子群优化算法的网页分类技术[J].计算机研究与发展,2006(6):33-38.

[11] 许建潮,胡明.中文 web 文本的特征获取与分类[J].计算机工程,2005(4):24-26.

(上接 118 页)

表示网络仿真结果,“+”表示专家评估结果。从图中看出二者拟合情况较好,具有较强的自适应能力。

6 结束语

本文提出了一种信息安全风险评估的模糊神经网络模型,该模型采用 BP 神经网络方法,对神经网络的输入进行了预处理,将模糊系统的输出作为神经网络的输入。该模型是一种非线性方法,不带有明显的主观成分和人为因素。人工神经网络经过训练,可以实时地估算风险因素的风险级别。

参考文献:

[1] GB/T 20984-2007 信息安全技术 信息系统的风险评估规范[S].北京:国家标准出版社,2007.

[2] Rainer R K.Risk analysis for information technology[J].Journal of Management Information Systems,1991,8(1):129-147.

[3] Tregear J.Risk assessment,Information Security Technical Report[R].2001-09.

[4] Patrick P.Minimization method for training feed forward neural network[J].Neural Network,1994,7:145-163.

[5] Hansen L K.Neural network ensembles[J].IEEE Transactions on Pattern Analysis and Machine Intelligence.1990,12(10):993-1001.

[6] 张青贵.人工神经网络[M].北京:中国水利水电出版社,2004.

[7] 赵振宇,徐用懋.模糊理论和神经网络的基础与应用[M].北京:清华大学出版社,1996.

[8] Zhao Dong-mei,Wang Jing-hong,Ma Jian-Feng.Fuzzy risk assessment of network security[C]//Proceedings of the 4th International Conference on Machine Learning and Cybernetics,Dalian,13-16 August 2006:4400-4405.