

同伦支持向量机

姜进超 张 瑞

(山东理工大学理学院 淄博 255049)

摘 要 支持向量机(SVM)作为一种新颖的人工智能技术,已越来越广泛地运用于各个学科领域。同伦正则化方法也是近几年新兴的一种正则化方法,在数理方程反问题中得到了广泛的应用。将同伦正则化的思想应用到支持向量机中,建立了一种新的支持向量机模型,并对最常用的核函数——高斯核进行了修改。与传统的正则化方法相比,新模型最大的优点就是正则化参数的取值范围由无限区间变成了有限区间(0,1),从而大大缩短了正则化参数的优化时间。

关键词 支持向量机,同伦正则化,高斯核

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.022

Homotopy Support Vector Machine

JIANG Jin-chao ZHANG Rui

(School of Science, Shandong University of Technology, Zibo 255049, China)

Abstract As a novel artificial intelligence technology, support vector machines (SVM) has been more and more widely used in various fields. Homotopy regularization method is a regularization method emerging in recent years, which has been widely used in the inverse problem. This paper applied the ideas of the homotopy regularization to the support vector machine, and established a new SVM. Meanwhile, we modified the most used Gaussian kernel. Compared with conventional regularization method, the biggest advantage of the new model is that the regularization parameter values range changes from infinite interval to a limited range (0,1), thus it greatly shortens the time of regularization parameter optimization.

Keywords Support vector machine, Homotopy regularization, Gaussian kernel

1 引言

支持向量机由于其优良的泛化性能和强大的模拟非线性关系的能力,使得它在分类和回归等问题中得到了广泛的应用^[1-3]。然而模拟非线性关系的好坏,主要取决于模型中正则化参数的选取及所选择的核函数。近几年人们对支持向量机理论的研究主要集中在核函数的构造上,例如:基于正交多项式构造了多种核函数;N. Ye 等人基于第一类切比雪夫正交多项式,提出了第一类切比雪夫核函数^[4];Sedat Ozer 等人^[4]将提出的第一类切比雪夫核函数推广成更一般化的向量形式的核函数^[5];张瑞等人基于勒让德多项式、拉盖尔多项式、埃尔米特多项式等理论提出了几类半参数化核函数^[6-10]。

传统的支持向量机运用了数学中的正则化思想,但模型中正则化参数的取值范围是一个无穷区间(0, +∞),难以优化。同伦正则化方法是近几年新兴的一个正则化方法,在数理方程反问题等领域得到了广泛的应用^[11-13],并取得了非常理想的效果。与传统的正则化方法相比,其最大的优点就是正则化参数的取值范围由无限区间变成了有限区间(0,1),这样就使实验设计非常容易,而且能大大缩短正则化参数的优化时间。

本文将同伦正则化方法应用于支持向量机中,建立了新的支持向量机模型,除此之外,我们还将最常用的高斯核函数进行了修改,将高斯核的取值范围由原来的无穷区间(0, +∞)变成了有限区间(0,1),在 UCI 数据集的实验结果表明,其能大大缩短正则化参数与高斯核参数的优化时间,取得了比较理想的效果。本文第 2 节简单介绍支持向量机;第 3 节详细阐述同伦支持向量机;第 4 节给出实验结果,并对实验结果加以分析;最后对本文加以总结,并指出进一步的工作方向。

2 支持向量机简介

支持向量机(Support Vector Machine)是一种新型的机器学习方法,对于二值分类问题,算法如下:

设已知训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\}$, 其中 $x_i \in R^d$, $y_i \in \{-1, +1\}$ 。对于非线性可分的分类问题,引入从 R^d 到高维特征空间 H 的映射 Φ , 构造优化问题:

$$\begin{cases} \min & \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^l \xi_i \\ \text{s. t.} & y_i (\langle \omega, \Phi(x_i) \rangle + b) \geq 1 - \xi_i \quad i=1, \dots, l \\ & \xi_i \geq 0 \quad i=1, \dots, l \end{cases} \quad (1)$$

到稿日期:2015-05-07 返修日期:2015-06-24 本文受国家自然科学基金(11371231),山东理工大学博士基金(4041-410002)资助。

姜进超(1992-),男,硕士生,主要研究方向为支持向量机、数据挖掘等,E-mail:982857636@qq.com;张 瑞(1964-),男,博士,教授,硕士生导师,主要研究方向为深度学习、支持向量机、环境预测等,E-mail:zrlgz@sdu.edu.cn。

其对偶问题是:

$$\begin{cases} \min & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j a_i a_j K(x_i, x_j) - \sum_{i=1}^l a_i \\ \text{s. t.} & \sum_{i=1}^l y_i a_i = 0 \\ & 0 \leq a_i \leq C, i=1, \dots, l \end{cases} \quad (2)$$

其中, $K(x_i, x_j)$ 为对应于映射 Φ 的核函数: $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$, 其中 $\langle, \rangle$ 为向量的内积。通过求解上述对偶问题得最优解 $\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_l^*)^T$, 选取 α^* 的一个正分量 $0 < \alpha_j^* < C$, 计算阈值:

$$b^* = y_j - \sum_{i=1}^l y_i \alpha_i^* K(x_i, x_j)$$

最后构造决策函数:

$$f(x) = \text{sgn}(\sum_{i=1}^l \alpha_i^* y_i K(x_i, x) + b^*)$$

3 同伦支持向量机

所谓同伦, 笼统说就是对于两个函数 $f(x)$ 和 $g(x)$, 引入一个正则化参数 λ , 得到一个全新的函数: $L(x) = (1-\lambda)f(x) + \lambda g(x)$ 中, 其中参数 λ 的取值范围为 $(0, 1)$, 这样通过正则化参数, 就可以把两个函数联系起来。

在数学中, 人们通常应用同伦方法证明方程解的存在性、唯一性等定性性质^[12, 13]。而同伦正则化思想常被用于解决数理方程反问题, 其优点是正则化参数由原来的无穷区间 $(0, +\infty)$ 转化为有限区间 $(0, 1)$, 这样实验设计就变得很容易, 而且可以大大缩短正则化参数的优化时间。根据同伦正则化在数理方程反问题中的优点, 将该思想引入到支持向量机中。

利用同伦支持向量机解决如下优化问题:

$$\begin{cases} \min & \frac{1}{2} (1-\lambda) \|\omega\|^2 + \lambda \sum_{i=1}^l \xi_i \\ \text{s. t.} & y_i (\langle \omega, x_i \rangle + b) \geq 1 - \xi_i \quad i=1, \dots, l \\ & \xi_i \geq 0 \quad i=1, \dots, l \end{cases} \quad (3)$$

引入拉格朗日函数

$$L(\omega, b, \lambda, \zeta, \alpha, \beta) = \frac{1}{2} (1-\lambda) \|\omega\|^2 + \lambda \sum_{i=1}^l \xi_i - \sum_{i=1}^l \alpha_i \{y_i$$

$$[\langle \omega, x_i \rangle + b] - 1 + \xi_i\} - \sum_{i=1}^l \beta_i y_i$$

求这个函数关于原始变量的微分, 得到

$$\frac{\partial L}{\partial \omega} = (1-\lambda)\omega - \sum_{i=1}^l \alpha_i x_i y_i = 0 \Rightarrow$$

$$\omega = \frac{1}{1-\lambda} \sum_{i=1}^l \alpha_i x_i y_i \quad (1)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^l \alpha_i y_i = 0 \quad (2)$$

$$\frac{\partial L}{\partial \xi_i} = \lambda - \sum_{i=1}^l \alpha_i - \sum_{i=1}^l \beta_i = 0 \Rightarrow$$

$$\lambda = \sum_{i=1}^l \alpha_i + \sum_{i=1}^l \beta_i \quad (3)$$

最终得到其对偶目标函数:

$$\begin{cases} \min & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j \cdot K(x_i, x_j) - (1-\lambda) \sum_{i=1}^l \alpha_i \\ \text{s. t.} & \sum_{i=1}^l \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq \lambda \end{cases} \quad (4)$$

另外, 最常用的核函数是高斯核: $K(x, y_i) = e^{-\gamma \|x - y\|^2}$ ($\gamma > 0$), 由于其中参数 γ 的取值范围是 $(0, +\infty)$, 难以优化, 因此本文对其作如下修改: $K(x, y) = e^{-\frac{\gamma}{1-\gamma} \|x - y\|^2}$ ($\gamma \in (0,$

$1)$), 这样参数 γ 的取值范围变成了 $(0, 1)$, 这样做的最大好处就是实验设计变得容易, 而且能大大缩短优化时间。

4 实验结果分析

我们将同伦支持向量机与支持向量机在标准数据库 UCI 中的数据集中做了分类对比实验, 所用数据集如表 1 所列。实验中, 两种算法各参数的取值如下。

(1) 同伦支持向量机方法将同伦正则化参数 λ 与高斯核参数 γ 的优化区间替换为 $[\frac{1}{l}, 1 - \frac{1}{l}]$, 其中 l 为训练集中样本的个数。将该区间等分为 10 份, 共 11 个等分点。若最优点在端点取得, 则将以该点为中心、半径为 $\frac{1}{2l}$ 的闭区间再 10 等分; 若最优点在内部节点取得, 则将以该点为中心、 $\frac{1}{10}$ 为半径的闭区间再 10 等分; 最后取最优值为两参数所求的点, 实验结果如表 2 所列。

(2) 支持向量机算法中正则化参数 C 和高斯核参数 σ 的取值皆为 $[1, 10, 50, 100, 500, 1000, 5000, 10000, 15000, 20000, 30000]$, 若最优点在初始点 1 取得, 则将以该点为中心、0.9 为半径的闭区间再 10 等分; 若最优点在其他点取得, 则将以该点为中心、以左邻点对应的值为半径的闭区间再 10 等分; 最后取最优值作为两参数所求的点, 实验结果如表 3 所列。

表 1 实验中用到的 UCI 数据集

数据集	维数	训练集的个数	测试集的个数
banana	2	200	100
heart	13	170	100
thyroid	5	140	23
waveform	21	200	100
ringnorm	8	400	285
titanic	3	150	100

表 2 同伦支持向量机在标准 UCI 数据集上分类的最大精度及对应的参数

数据集	精度	同伦正则化参数	修改后的高斯核参数
banana	0.9200	0.3930	0.9900
heart	0.8300	0.5980	0.1080
thyroid	0.9867	0.5294	0.5784
waveform	0.9100	0.3612	0.2836
ringnorm	0.9469	0.4501	0.6098
titanic	0.7816	0.1080	0.3040

表 3 支持向量机在标准 UCI 数据集上分类的最大精度及对应的参数

数据集	精度	正则化参数	高斯核参数
banana	0.9100	19.2000	0.5421
heart	0.8300	8.0000	1.4571
thyroid	0.9467	42.0000	0.4600
waveform	0.9000	57.6000	2.9503
ringnorm	0.8727	1.5400	0.8200
titanic	0.7743	9.0000	9.0000

由表 2、表 3 对比可知, 同伦支持向量机方法与传统支持向量机方法相比, 具有实验设计简便, 而且容易求出最优点的优势。

结束语 本文主要是基于同伦正则化思想在数理方程反问题中的广泛应用, 取其优点, 将其应用到支持向量机中, 得到一种新的支持向量机模型。我们还对最常用的高斯核做了改进。与原模型相比较, 新模型将原模型的参数取值范围由无限区间变成了有限区间 $(0, 1)$, 这样在实验设计时会比较容

易,并且能大大缩短正则化参数以及高斯核参数的优化时间。同伦正则化思想还可以应用到深度学习等其他的研究领域中。

参考文献

- [1] Vapnik V. Statistical learning Theory[M]. New York: Wiley, 1998
- [2] Zhang Rui, Ma Yi-chen. Kernel Methods for Pattern Analysis [D]. Xi'an: Xi'an Jiaotong University, 2009 (in Chinese)
张瑞, 马逸尘. 模式分析的核方法[D]. 西安: 西安交通大学, 2009
- [3] Zhang R, Wang W J, Ma Y C. Least square transduction support vector machine[J]. Neural Processing Letters, 2009, 29(2): 133-142
- [4] Ye N, Sun R, Liu Y, et al. Support vector machine with orthogonal Chebyshev kernel[C]// Proceedings of the 18th international conference on Pattern Recognition. 2006
- [5] Sedat O, Chen C H, Hakan A C. A set of new Chebyshev kernel function for support vector machine pattern classification[J]. Pattern Recognition, 2011, 44(7): 1435-1447
- [6] Zhang R, Wang W J. Facilitating the Applications of Support machine by Using a new kernel[J]. Expert Systems with Application, 2011, 38: 14225-14230
- [7] Zhang Rui, Gao Hong, Zhang Li-wei. A New Set of Hermite Kernel Functions for Support Vector Machine[J]. Journal of Shanxi University (Natural Science Edition), 2012, 35(1): 38-42 (in Chinese)

- 张瑞, 高红, 张立伟. 一类新的支持向量机核函数-埃尔米特核函数[J]. 山西大学学报(自然科学版), 2012, 35(1): 38-42
- [8] Zhang Rui, Wang Wen-jian, Zhang Ya-dan, et al. Legendre Kernel Function for Support Vector Classification[J]. Computer Science, 2012, 39(7): 222-224 (in Chinese)
张瑞, 王文剑, 张亚丹, 等. 基于支持向量机分类问题的勒让德核函数[J]. 计算机科学, 2012, 39(7): 222-224
- [9] Zhang Rui, Wang Wen-jian, Wang Jia-qi, et al. Laguerre Kernel Functions for Support Vector Classification[J]. Computer Engineering and Applications, 2012, 48(36): 50-53 (in Chinese)
张瑞, 王文剑, 王嘉琦, 等. 一类新的基于拉盖尔正交多项式的核函数[J]. 计算机工程与应用, 2012, 48(36): 50-53
- [10] Zhang Rui, Yang Xiao, Tan Xiu-lin. New SVM Kernel Function Based on Gegenbauer Polynomial[J]. Journal of Shanxi University (Natural Science Edition), 2013, 36(1): 30-33 (in Chinese)
张瑞, 杨晓, 谭秀林. 基于盖根鲍尔多项式的 SVM 核函数[J]. 山西大学学报(自然科学版), 2013, 36(1): 30-33
- [11] Liao Shi-jun. Beyond the Perturbation: the basic idea and application of the Homotopy analysis method [J]. Journal of Mechanics, 2008, 38(1): 1-34
- [12] Agarwal R P, O'Regan D. Homotopy and existence of solutions for the nonlinear equation $Lx \in N_x$ [J]. Nonlinear Analysis, 2001, 44(4): 537-544
- [13] Soriano J M. On the existence of zero points of a continuous function[J]. Acta Mathematica Scientia, 2002, 22(2): 171-177

(上接第 82 页)

结束语 本文以微博消息为对象,研究微博热点事件中的话题演化问题。按照处理流程的不同,微博话题演化处理框架分为 5 个阶段,分别是数据采集、预处理、话题抽取、话题关联和结果展示。在话题抽取阶段,以 LDA 模型为基础,结合时间因素和微博转发特性,提取微博话题信息。在话题关联阶段,综合考虑话题的内容信息和话题下微博消息集间的转发关系,通过话题内容相关性和话题转发关联度两个指标来构建话题关联函数,重现话题演化过程。利用真实微博数据进行实验,结果表明 MTERR 算法可以有效地抽取话题信息,构建话题演化图,从而反映热点事件发展演化的过程。

参考文献

- [1] Hong Yu, Zhang Yu, Liu Ting, et al. Topic Detection and Tracking Review[J]. Journal of Chinese Information Processing, 2007, 21(6): 71-87 (in Chinese)
洪宇, 张宇, 刘挺, 等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报, 2007, 21(6): 71-87
- [2] Brants T, Chen F, Farahat A. A system for new event detection [C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada: ACM, 2003: 330-337
- [3] Kumaran G, Allan J. Text classification and named entities for new event detection [C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield: ACM, 2004: 297-304
- [4] Hofmann T. Probabilistic Latent Semantic Indexing [C]// Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

- Berkeley, CA, USA: ACM, 1999: 50-57
- [5] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022
- [6] Ha-Thuc V, Mejova Y, Harris C G, et al. A relevance-based topic model for news event tracking [C]// Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston, MA, USA: ACM, 2009: 764-765
- [7] Alsumait L, Barbar D, Domeniconi C. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking [C]// Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy: IEEE Computer Society, 2008: 3-12
- [8] Wang X, Zhang K, Jin X M, et al. Mining common topics from multiple asynchronous text streams [C]// Proceedings of the Second International Conference on Web Search and Web Data Mining. Barcelona, Spain: ACM, 2009: 192-201
- [9] Gohr A, Hinneburg A, Schult R, et al. Topic Evolution in a Stream of Documents [C]// Proceedings of the SIAM International Conference on Data Mining. Sparks, Nevada, USA: SIAM, 2009: 859-870
- [10] Jo Y, Hopcroft J E, Lagoze C. The web of topics: discovering the topology of topic evolution in a corpus [C]// Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India: ACM, 2011: 257-266
- [11] <http://nlp.stanford.edu/software/tmt/tmt-0.4>
- [12] Manning C D, Raghavan P, Schutze H. Introduction to Information Retrieval [M]. New York: Cambridge University Press, 2008: 142-143
- [13] MH370 [EB/OL]. (2014-03-08) [2014-10-30]. <http://baike.baidu.com/view/12368712.htm>