基于最大间隔最小体积超球支持向量机的多主题分类算法

艾 青[1] 赵 骥[1] 秦玉平[2]

(辽宁科技大学软件学院 鞍山 114051)[1] (渤海大学信息科学与工程学院 锦州 121000)[2]

摘 要 针对多主题分类,结合最大间隔最小体积超球支持向量机和模糊理论,提出一种多主题最大间隔最小体积超球支持向量机来实现多主题分类。该算法首先基于最大间隔最小体积超球支持向量机,采用1-a-r方法训练子分类器,通过子分类器得到待分类样本的隶属度向量,再依据隶属度向量判定该待分类样本所属类别。实验结果表明,该算法具有较好的准确率、召回率、F1值。

关键词 最大间隔最小体积超球支持向量机,隶属度,隶属度向量

中图法分类号 TP18 **文献标识码** A

Multi-subjects Classification Algorithm Based on Maximal-margin Minimal-volume Hypersphere Support Vector Machine

AI Qing[1] ZHAO Ji[1] QIN Yu-ping[2]

(College of Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China)[1]

(College of Information Science and Technology, Bohai University, Jinzhou 121000, China)[2]

Abstract For multi-subjects classification problem, a multi-subjects maximal-margin minimal-volume hypersphere support vector machine was proposed according to maximal-margin minimal-volume hypersphere support vector machine and fuzzy theory. The algorithm uses 1-a-r maximal-margin minimal-volume hypersphere support vector machine to train sub-classifiers, obtain membership vector of the sample that is classified according to the classifiers. At last it labels the subjects that the sample belongs to according to the membership vector. The experimental results show that the algorithm has higher performance on precision, recall and F1.

Keywords Maximal-margin minimal-volume hypersphere support vector machine, Membership, Membership vector

1 引言

支持向量机是一种建立在统计学习理论基础上的新的分类和回归技术[1,2]。它基于结构风险最小化原则,根据有限样本信息在模型的复杂度和学习能力之间寻求最佳的折衷,由于其出色的泛化性能,已广泛应用于文本分类[3,4]、人脸识别[5,6]等领域,并取得良好的效果。

但是支持向量机多分类算法[7-9]面临着一个问题,即分类算法要求训练样本为单一主题,并且只能为待分类样本标注一个类别。然而,实际应用中,大部分样本都是多主题的,例如,一篇关于911事件的新闻报道,它应该既属于军事类新闻,又属于政治类新闻。对于这种情况,支持向量机多分类算法是无能为力的。文献[10]提出了基于DAGSVM的兼类标注算法,该算法结合DAGSVM和模糊理论,采用模糊决策面替代分明决策面来实现兼类标注。文献[11]提出了一种基于超球支持向量机的多主题分类算法。该算法用超球支持向量机训练得到每个超球,计算待分类样本到每个超球球心的距离,依据距离得到隶属度向量,最后根据隶属度向量判定该样本所属的主题。文献[12]结合支持向量机类间最大分类间隔和支持向量数据描述类内最小体积思想,提出一种最大间隔最小体积球形支持向量机。该模型建立两个大小不一的同心超球,将正负类样本分别映射到小超球内和大超球外,模型目标函数最大化两超球间隔,实现正负类类间间隔的最大化和各类类内体积的最小化,进而提高了模型的泛化性能。本文结合最大间隔最小体积超球支持向量机和模糊理论,提出一种多主题最大间隔最小体积超球支持向量机来实现多主题分类。该算法首先基于最大间隔最小体积超球支持向量机,采用1-a-r方法训练子分类器,通过子分类器得到待分类样本的隶属度向量,再依据隶属度向量判定该待分类样本所属类别。

本文第2节单介绍了最大间隔最小体积超球支持向量机的数学模型;第3节详细阐述了多主题最大间隔最小体积超球支持向量机;第4节给出了在标准数据集上的比较结果;最后得出结论。

2 最大间隔最小体积超球支持向量机

设正负类样本$\{x_i, y_i\}$,$i=1,2,\cdots,n$,$x_i \in R^d$,$y_i \in \{-1, 1\}$,记正类训练样本为x_i^+,正类训练样本数为n^+,负类训练样本为x_j^-,负类训练样本数为n^-。设存在两个同心超球S_1

到稿日期:2011-09-19 返修日期:2011-11-23 本文受国家自然科学基金项目(60603023),辽宁省教育厅资助科研课题(2010076)资助。

艾 青(1980—),男,硕士,讲师,主要研究方向为机器学习和数据挖掘,E-mail:lyaiqing@gmail.com;赵 骥(1974—),男,博士生,副教授,主要研究方向为机器学习和图形图像处理;秦玉平(1965—),男,博士,教授,主要研究方向为机器学习和决策支持系统。

和 S_2，球心为 c。S_1 为小超球，半径为 R_1；S_2 为大超球，半径为 R_2，且 $R_2 \geqslant R_1$。S_1 将正类样本包裹其中，S_2 将负类样本排除其外。最大间隔最小体积超球支持向量机数学模型如下：

$$\max_{R_1,R_2,c,\xi_i,\xi_j} R_2^2 - MR_1^2 - C_+ \sum_i \xi_i^+ - C_- \sum_j \xi_j^- \tag{1}$$

$$\text{s.t.} \ \| x_i^+ - c \|^2 \leqslant R_1^2 + \xi_i^+ \tag{2}$$

$$\| x_j^- - c \|^2 \geqslant R_2^2 - \xi_j^- \tag{3}$$

$$R_2^2 - R_1^2 \geqslant 0, \xi_i \geqslant 0, \xi_j \geqslant 0, \forall i, \forall j \tag{4}$$

式中，ξ_i^+ 和 ξ_j^- 为松弛因子，分别用于约束正负类奇异点；C_+ 和 C_- 为惩罚因子，用于控制模型训练精度和泛化性能的平衡；参数 $M(M>0, M\neq 1)$ 用于 R_2^2 和 R_1^2 的折衷。

使用对偶理论，原始问题式(1)－式(4)的对偶问题为：

$$\begin{aligned}\min_{\alpha} & -\sum_i x_i^+ \cdot x_i^+ + \sum_j x_j^- \cdot x_j^- + \\ & \frac{1}{M-1}\sum_{i,l}\alpha_i^+\alpha_l^+(x_i^+ \cdot x_l^+) - \\ & \frac{2}{M-1}\sum_{i,j}\alpha_i^+\alpha_j^-(x_i^+ \cdot x_j^-) + \\ & \frac{1}{M-1}\sum_{j,k}\alpha_j^-\alpha_k^-(x_j^- \cdot x_k^-)\end{aligned} \tag{5}$$

$$\text{s.t.} \ \sum_i \alpha_i^+ - \sum_j \alpha_j^- = M-1 \tag{6}$$

$$\sum_i \alpha_i^+ \geqslant M, 0 \leqslant \alpha_i^+ \leqslant C_+, 0 \leqslant \alpha_j^- \leqslant C_- \tag{7}$$

求解对偶问题式(5)－式(7)可得最优解 α_i^+、α_j^-。当 $0<\alpha_i^+<C_+$ 时，对应的 x_i^+ 称为超球 S_1 的支持向量；当 $0<\alpha_j^-<C_-$ 时，对应的 x_j^- 称为超球 S_2 的支持向量；任取 S_1 支持向量和 S_2 支持向量，计算其到球心距离即可获得 R_1 和 R_2。其中，球心 $c=\frac{1}{M-1}(\sum_i \alpha_i^+ x_i^+ - \sum_j \alpha_j^- x_j^-)$。

3 多主题最大间隔最小体积超球支持向量机

设给定多主题样本集 $A=\{x_i, E_i\}_{i=1}^l$ 和核函数 $K(x_i, x_j)$，其中，$x_i \in R_n$，$E_i=\{y_{ij}\}_{j=1}^p$，$y_{ij} \in \{1,2,3,\cdots,q\}$，$q$ 是样本集 A 中含有的总类别数，$p(\leqslant q)$ 是样本 x_i 的兼类数；K 对应某特征空间 Z 中的内积，即 $K(x_i,x_j)=g(x_i)\cdot g(x_j)$，变换 $g: X \mapsto Z$ 是将样本从输入空间映射到特征空间。对于训练样本，用 1-a-r 方法构建 q 个 $H_i(c_i, R_i^+, R_-i)(i=1,2,\cdots,q)$，其中，$H_i$ 为以第 i 类为正类、其余类为负类训练最大间隔最小体积超球支持向量机所得的最小包围第 i 类样本并最大排除其余类样本的同心超球；c_i 为该同心超球的球心；R_i^+ 为最小包围第 i 类样本超球的半径；R_i^- 为最大排除其余类样本超球的半径。记 $H_i(c_i, R_i^+)$ 为 $H_i(c_i, R_i^+, R_i^-)$ 的内超球，$H_i(c_i, R_i^-)$ 为 $H_i(c_i, R_i^+, R_i^-)$ 的外超球。

定义 1 待分类样本 x 关于超球 $H_i(c_i, R_i^+)$、$H_i(c_i, R_i^-)$ 隶属度定义如下：

$$m_i^+(x) = 1 - d_i(x)/R_i^+ \tag{8}$$

$$m_i^-(x) = 1 - d_i(x)/R_i^- \tag{9}$$

式中，$d_i(x) = \| x - c_i \|$。

定义 2 待分类样本 x 关于超球 $H_i(c_i, R_i^+, R_i^-)$ 隶属度定义如下：

$$m_i(x) = 0.6 * m_i^+(x) + 0.4 * m_i^-(x) \tag{10}$$

由于存在 q 个超球对，因此可以得到如下的隶属度向量：

$$M(x) = \begin{bmatrix} m_1(x) \\ \vdots \\ m_q(x) \end{bmatrix} \tag{11}$$

式中，$m_i(x)(i=1,\cdots,q)$ 是待分类样本 x 关于超球 $H_i(c_i, R_i^+, R_i^-)$ 的隶属度。

经过上面的计算，已经得到了待分类样本 x 对于各类的模糊值。去模糊过程如下：如果存在 $m_i(x)>0$，则设阈值 $\kappa(0<\kappa\leqslant 1)$，当样本对于某一类别的模糊值大于等于 $\kappa \times \max(m_i(x))$ 时，用该类别标记样本。如果所有 $m_i(x)<0$，则待分类样本所属主题为距该样本最近的超球。

算法具体描述如下：

Step 1 给定待分类样本 x 和权值 κ；

Step 2 根据式(10)得到待分类样本 x 的隶属度矩阵 $M(x)$；

Step 3 如果 $\text{MAX}(m_i(x))>0$，则转 Step 4；否则，转 Step 5；

Step 4 如果 $m_i(x) \geqslant \kappa \times \text{MAX}(m_i(x))$，则样本属于主题 i；

Step 5 如果 $d_i(x) = \text{MIN}(d_j(x))$，则样本属于主题 i；

Step 6 结束。

4 实验结果与分析

仿真实验采用 MB04a 标准数据集 scene，该数据集包括 6 类($q=6$)，样本兼类数最大为 3($p=3$)共 2407 个样本，随机选取其中的 1211 个样本作为训练样本，其余的 1196 个样本作为测试样本。

实验中采用通用的准确率(AP)、召回率(AR)和平均 $F1$ 值(AF)作为评价指标。

$$\text{准确率(P)} = N_c / N_a \tag{12}$$

$$\text{召回率(R)} = N_c / N_r \tag{13}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{14}$$

式中，N_c 代表对每个测试样本测试后得到的正确主题数；N_a 代表对每个测试样本测试后得到的所有主题数；N_r 代表每个测试样本实际的主题数。

定义 3 平均准确率

$$\text{AP} = \frac{\sum P}{n} \tag{15}$$

定义 4 平均召回率

$$\text{AR} = \frac{\sum R}{n} \tag{16}$$

定义 5 平均 $F1$ 值

$$\text{AF} = \frac{\sum F1}{n} \tag{17}$$

式中，n 为测试样本总数。

图 1 针对不同的 κ 给出了平均准确率、平均召回率和平均 $F1$ 值的变化情况，其中，核函数为径向基函数 $K(x,y)=e^{-0.25\|x-y\|^2}$，系统参数 $C_+=C_-=100$，$M=32$。从图 1 可以看出，当 κ 较小时，平均准确率较低，平均召回率较高。随着 κ 的增加，平均准确率提高，平均召回率降低。当平均准确率与平均召回率最接近时，平均 $F1$ 值取得最大值，即在该点处系统的整体性能最佳。大量实验表明，κ 值一般在 0.75～0.95 之间为最佳。

表 1 给出了多主题最大间隔最小体积超球支持向量机与

(下转第 267 页)

[4] 宋树华,程承旗,关丽,等. 全球空间数据剖分模型分析[J]. 地理与地理信息科学,2008,24(4):11-15

[5] 程承旗,郭辉. 基于剖分数据模型的影像信息表达研究[J]. 测绘通报,2009(10):12-14,17

[6] 程承旗,宋树华,万元嵬,等. 基于全球剖分模型的空间信息编码模型初探[J]. 地理与地理信息科学,2009,25(4):8-11

[7] 袁文,程承旗,马蔼乃,等. 球面三角区域四叉树 L 空间填充曲线[J]. 中国科学 E 辑,2004,34(5):584-600

[8] Dutton G H. A hierarchical coordinate system for geoprocessing and cartography[J]. Lecture Notes in Earth Science, Berlin: Springer-Verlag,1999,79

[9] 袁文,马蔼乃,管晓静. 一种新的球面三角投影:等角比投影(EARP)[J]. 测绘学报,2005,34(1):78-84

[10] 袁文,庄大方,袁武,等. 基于等角比例投影的球面三角四叉树剖分模型[J]. 遥感学报,2009,13(1):103-111

[11] 李德仁. 论广义空间信息网格和狭义空间信息网格[J]. 遥感学报,2005,9(5):513-520

[12] 邓淑明,胡思仁. 地理信息网络服务与应用[M]. 曾杉,译. 北京:科学出版社,2004

[13] Shao Zhen-feng, Li De-ren. Design and implementation of service-oriented spatial information sharing framework in digital city [J]. Geo-Spatial Information Science,2009,12(2):104-109

[14] Li De-ren, Shen Xin. Geospatial information service based on digital measurable image-Take Image City Wuhan as an example [J]. Geo-Spatial Information Science,2010,13(2):79-84

[15] 苗放,叶成名,刘瑞,等. 新一代数字地球平台与"数字中国"技术体系架构探讨[J]. 测绘科学,2007,32(6):157-158,168

(上接第 238 页)

文献[10,11]所提多主题分类算法的比较结果。实验中使用的核函数为径向基函数 $K(x,y)=e^{-\gamma\|x-y\|^2}$。本文算法参数 $C_+=C_-=C$,M,文献[10]所提算法参数 C,文献[11]所提算法参数 v 以及径向基函数参数 γ 均通过交叉确认得出,其中折取 3;而本文算法经验参数 κ,文献[10]所提算法经验参数 θ,文献[11]所提算法经验参数 τ,则需通过大量实验得出,如表 2 所列。

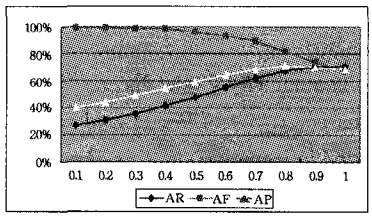

图 1 不同 κ 下 AP、AR、AF 的变化情况

表 1 本文所提多主题分类算法与其他算法的比较结果

算法	平均准确率	平均召回率	平均 F1 值
文献[10]多主题算法	43.13%	46.82%	43.62%
文献[11]多主题算法	49.45%	90.93%	58.30%
本文多主题算法	67.52%	80.94%	70.35%

表 2 参数设置

参数 (交叉确认范围,步长)	文献[4]多主题算法	文献[6]多主题算法	本文多主题算法
$C(2^0\sim2^5,2^1)$	2	——	16
$v(2^{-5}\sim2^{-1},2^1)$	——	0.0625	——
$\gamma(2^{-5}\sim2^5,2^1)$	0.03125	0.125	0.25
$M(2^1\sim2^5,2^1)$	——	——	32
θ	0.7	——	——
τ	——	0.7	——
κ	——	——	0.8

从表 1 可以看出,与文献[10,11]所提算法相比,本文所提多主题分类算法其平均准确率、平均召回率和平均 $F1$ 值都有明显的提高。

结束语 本文提出了一种多主题分类算法,它结合最大间隔最小体积超球支持向量机和模糊理论实现了多主题分类。实验结果表明,该算法与传统算法相比,具有更好的准确率、召回率和 $F1$ 值,是一种更为实用的处理多主题分类的算法。

参 考 文 献

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

[2] 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

[3] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Feature[C]// Proceedings of ECML-98,10th European Conference on Machine Learning. Berlin: Springer,1998:137-142

[4] 马金娜,田大钢. 基于支持向量机的中文文本自动分类研究[J]. 系统工程与电子技术,2007,29(3):475-478

[5] 崔国勤,高文. 基于双层虚拟视图和支持向量的人脸识别方法[J]. 计算机学报,2005,28(3):368-375

[6] 谢赛琴,沈福明,邱雪娜. 基于支持向量机的人脸识别方法[J]. 计算机工程,2009,35(16):186-188

[7] Krebel U G. Pairwise Classification and Support Vector Machines[C]// Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT press,1999:255-268

[8] Bennett K P. Combining Support Vector and Mathematical Programming Methods for Classification[C]// Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT press,1999:307-326

[9] Platt J C, Cristianini N, Shawe-Taylor J. Large Margin DAGs for multiclass classification[C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000: 547-553

[10] 王晔,黄上滕. 基于支持向量机的文本兼类标注[J]. 计算机工程与应用,2006,42(2):182-185

[11] 艾青,秦玉平,李迎春. 基于超球支持向量机的多主题文本分类算法[J]. 计算机工程与设计,2010,31(10):2273-2275

[12] 文传军,詹永照,陈长军. 最大间隔最小体积球形支持向量机[J]. 控制与决策,2010,25(1):79-83