DOI:10.11721/cqnuj20150509

基于最小流形类内离散度支持向量机的 110 m 栏成绩预测方法研究*

张 雅 清
(太原学院 数学系,太原 030012)

摘要:针对支持向量机在成绩预测时面临的泛化能力不足问题,提出基于最小流形类内离散度支持向量机 M^2SVM。实验选取 2000－2009 年 59 次刘翔 110 m 栏成绩作为研究对象,首先将前 54 次成绩作为训练样本并对模型进行训练得到分类标准,然后将后 5 次成绩作为测试样本并依次输入模型,比较预测结果与实际结果之间的相似程度,从而说明所提方法的有效性。该方法对人才选拔、成绩提升和梯队建设等具有重要意义。

关键词:支持向量机;最小流形类内离散度;110 m 栏;成绩预测
中图分类号:TB114　　**文献标志码**:A　　**文章编号**:1672-6693(2015)05-0165-04

随着 2004 年雅典奥运会刘翔 110 m 栏夺冠,110 m 栏逐渐成为中国田径为数不多的亮点[1-3]。目前,众多青少年从事 110 m 栏专业训练,如何利用以往成绩对日后成绩进行预测成为田径界乃至体育界面临的重要课题之一,相关研究成果对于人才选拔、成绩提升和梯队建设具有重要意义。

近年来,支持向量机(Support vector machine, SVM)[4-7]作为一种经典的模式分类方法广泛应用于成绩预测领域。基本思想是在 Vapnik 建立的统计学习理论基础上提出结构风险最小化原则,通过最大化分类间隔,寻找一个最优分类面实现样本的有效分类。然而随着应用的深入 SVM 逐渐暴露出泛化能力有限的问题,其原因在于该方法在建立最优分类面时只考虑类间的绝对间隔而忽略各类的分布性状。鉴于此,在流形判别分析(Manifold-based discriminant analysis, MDA)[8]的基础上提出最小流形类内离散度支持向量机(Minimum manifold-based within-class scatter support vector machine, M^2SVM)。该方法在建立最优分类面时,充分利用样本的全局信息和局部信息,有效地提高了预测效率。

1 最小流形类内离散度支持向量机

设给定样本集 $\boldsymbol{X}=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_N,y_N)\}$,其中 $\boldsymbol{x}_i\in\mathbf{R}^d(1\leqslant i\leqslant N_1+N_2=N)$为样本,$y_i\in\{1,-1\}$为类别标签。当 $1\leqslant i\leqslant N_1$ 时,$y_i=1$;当 $N_1+1\leqslant i\leqslant N$ 时,$y_i=-1$。第一类含有 N_1 个样本$\{\boldsymbol{x}_i,y_i\}_{i=1}^{N_1}$,第二类含有 N_2 个样本$\{\boldsymbol{x}_j,y_j\}_{j=N_1+1}^{N}$。$\bar{\boldsymbol{x}}$ 表示所有样本均值,$\bar{\boldsymbol{x}}_1$ 和 $\bar{\boldsymbol{x}}_2$ 分别表示第一类和第二类样本均值。

1.1 SVM

设超平面方程为 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}+b=0$,分类间隔为 $2/\|\boldsymbol{W}\|$,该最优化问题可描述为:

$$\min_{\boldsymbol{W},b,\xi_i}\frac{1}{2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+C\sum_{i=1}^{N}\xi_i, \tag{1}$$

$$\text{s.t. } y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)\geqslant 1-\xi_i,\xi_i\geqslant 0\quad i=1,\cdots,N, \tag{2}$$

其中 C 为惩罚因子,它控制对错分样本的惩罚程度,通过引入松弛因子 ξ_i 允许错分样本的存在。

1.2 MDA

文献[8]提出流形判别分析 MDA,其中有两个重要概念:基于流形的类内离散度(Manifold-based within-class scatter,MWCS)$\boldsymbol{M}_W$ 和基于流形的类间离散度(Manifold-based between-class scatter, MBCS)$\boldsymbol{M}_B$。$\boldsymbol{M}_W$ 和 $\boldsymbol{M}_B$ 的定义如下:

$$\boldsymbol{M}_W=\mu\boldsymbol{S}_W+(1-\mu)\boldsymbol{S}_S, \tag{3}$$

$$\boldsymbol{M}_B=\lambda\boldsymbol{S}_B+(1-\lambda)\boldsymbol{S}_D, \tag{4}$$

其中 μ 和 λ 为平衡参数。$\boldsymbol{S}_W$ 和 $\boldsymbol{S}_B$ 分别表征样本的类内离散度和类间离散度;$\boldsymbol{S}_S$ 和 $\boldsymbol{S}_D$ 分别保持同类样本和异

* **收稿日期**:2014-11-26　**修回日期**:2015-05-26　**网络出版时间**:2015-05-15 12:44
资助项目:国家自然科学基金(No. 61202311);山西省社会经济统计科研课题(No. KJ[2014]036)
作者简介:张雅清,讲师,研究方向为概率论与数理统计,E-mail:zhyqwhwei@163.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.n.20150515.1244.022.html

类样本的局部流形结构。$\boldsymbol{S}_W$,$\boldsymbol{S}_B$,$\boldsymbol{S}_S$,$\boldsymbol{S}_D$ 定义如下:

$$\boldsymbol{S}_W=\sum_{i=1}^{N_1}(\boldsymbol{x}_i-\overline{\boldsymbol{x}}_1)(\boldsymbol{x}_i-\overline{\boldsymbol{x}}_1)^{\mathrm{T}}+\sum_{j=1}^{N_2}(\boldsymbol{x}_j-\overline{\boldsymbol{x}}_2)(\boldsymbol{x}_j-\overline{\boldsymbol{x}}_2)^{\mathrm{T}},\tag{5}$$

$$\boldsymbol{S}_B=N_1(\overline{\boldsymbol{x}}_1-\overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_1-\overline{\boldsymbol{x}})^{\mathrm{T}}+N_2(\overline{\boldsymbol{x}}_2-\overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_2-\overline{\boldsymbol{x}})^{\mathrm{T}},\tag{6}$$

$$\boldsymbol{S}_S=\boldsymbol{X}(\boldsymbol{S}'-\boldsymbol{S})\boldsymbol{X}^{\mathrm{T}},\tag{7}$$

其中 $\boldsymbol{S}'$ 为对角阵且 $\boldsymbol{S}'=\sum_j S_{ij}$,其中 S_{ij} 为同类权重函数:

$$S_{ij}=\begin{cases}\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2), & y_i=y_j,\\ 0, & y_i\neq y_j,\end{cases}\tag{8}$$

$$\boldsymbol{S}_D=\boldsymbol{X}(\boldsymbol{D}'-\boldsymbol{D})\boldsymbol{X}^{\mathrm{T}},\tag{9}$$

其中 $\boldsymbol{D}'$ 为对角阵且 $\boldsymbol{D}'=\sum_j D_{ij}$,其中 D_{ij} 为异类权重函数:

$$D_{ij}=\begin{cases}\exp(-1/\|x_i-x_j\|^2), & y_i\neq y_j,\\ 0, & y_i=y_j。\end{cases}\tag{10}$$

MDA 的基本思路是在 Fisher 准则的基础上通过最大化 MBCS 与 MWCS 之比获得最佳投影方向。上述思想可转化为如下优化问题:

$$J=\max_{w}\frac{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_B\boldsymbol{W}}{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_W\boldsymbol{W}}=\max_{w}\frac{\boldsymbol{W}^{\mathrm{T}}(\lambda\boldsymbol{S}_B+(1-\lambda)\boldsymbol{S}_D)\boldsymbol{W}}{\boldsymbol{W}^{\mathrm{T}}(\mu\boldsymbol{S}_W+(1-\mu)\boldsymbol{S}_S)\boldsymbol{W}}。\tag{11}$$

1.3 M^2SVM

1.3.1 最优化问题 为了解决 SVM 面临的泛化能力不足问题,提出最小 MWCS 支持向量机 M^2SVM,其最优化问题可描述为:

$$\min_{\boldsymbol{W}}\quad \boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_W\boldsymbol{W}+C\sum_{i=1}^{N}\xi_i,\tag{12}$$

$$\text{s.t. } y_i(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1-\xi_i,\xi_i\geqslant 0,\ i=1,\cdots,N。$$

由 Lagrangian 定理可得上述优化问题的对偶形式:

$$\max_{\alpha}\quad \alpha^{\mathrm{T}}\mathbf{1}-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\alpha},\tag{13}$$

$$\text{s.t.}\quad \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}=0,\mathbf{0}\leqslant\boldsymbol{\alpha}\leqslant\boldsymbol{C},$$

其中 $\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_N]^{\mathrm{T}}$,$\mathbf{1}=[1,\cdots,1]^{\mathrm{T}}$,$Y=[y_1,\cdots,y_N]^{\mathrm{T}}$,$\mathbf{0}=[0,\cdots,0]^{\mathrm{T}}$,$\boldsymbol{C}=[C,\cdots,C]^{\mathrm{T}}$,$\boldsymbol{P}=\left[\frac{1}{2}y_iy_j\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{M}_W^{-1}\boldsymbol{x}_j\right]$。当 $\boldsymbol{M}_W$ 奇异时,采用扰动法予以解决,即在矩阵 $\boldsymbol{M}_W$ 主对角线上加一个很小的正数来消除 $\boldsymbol{M}_W$ 奇异性。

1.3.2 决策函数 M^2SVM 的判别函数定义如下:

$$f(\boldsymbol{x})=\operatorname{sgn}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}+b),\tag{14}$$

其中

$$\boldsymbol{W}=\frac{1}{2}\boldsymbol{M}_W^{-1}\sum_{i=1}^{N}\alpha_iy_i\boldsymbol{x}_i,\tag{15}$$

$$b=y_i-\frac{1}{2}\sum_{j=1}^{N}\alpha_jy_jx_j{}^{\mathrm{T}}\boldsymbol{M}_W^{-1}\boldsymbol{x}_i(\boldsymbol{x}_i\text{ 为任意支持向量})。\tag{16}$$

2 实验结果与分析

2.1 实验环境及数据集

实验环境为 Windows XP 及 Matlab 7.0。以 2000—2009 年 59 次刘翔 110 m 栏成绩[9]作为实验数据集,如表 1 所示。为了方便处理,利用下式对成绩进行归一化处理:

$$x'=x/x_{\max},\tag{17}$$

其中 x' 表示归一化后的成绩,x 表示实际成绩,$x_{\max}$ 表示成绩最大值。

2.2 实验方法

选取实验数据集前 54 次成绩作为训练样本,剩余的 5 次成绩作为测试样本。首先利用训练样本对模型进行训练得到分类标准,然后将测试样本依次输入模型比较预测结果与实际结果之间的相似程度,从而说明所提方法的有效性。

表 1　实验数据集

编号	实际成绩/s	归一化成绩	编号	实际成绩/s	归一化成绩	编号	实际成绩/s	归一化成绩
1	13.87	0.999 999	21	13.20	0.951 694	41	13.12	0.945 926
2	13.32	0.960 346	22	13.06	0.941 601	42	13.05	0.940 880
3	13.42	0.967 556	23	13.40	0.966 114	43	13.08	0.943 043
4	13.33	0.961 067	24	13.25	0.955 299	44	13.10	0.944 484
5	13.36	0.963 230	25	13.11	0.945 205	45	13.21	0.952 415
6	13.76	0.992 069	26	13.06	0.941 601	46	13.22	0.953 136
7	13.12	0.945 926	27	13.27	0.956 741	47	13.21	0.952 415
8	13.56	0.977 650	28	13.26	0.956 020	48	13.19	0.950 973
9	13.27	0.956 741	29	13.18	0.950 252	49	12.88	0.928 623
10	13.50	0.973 324	30	12.91	0.930 786	50	13.15	0.948 089
11	13.27	0.956 741	31	13.59	0.979 813	51	13.14	0.947 368
12	13.51	0.974 045	32	13.23	0.953 857	52	12.92	0.931 507
13	13.45	0.969 719	33	13.12	0.945 926	53	13.23	0.953 857
14	13.22	0.953 136	34	13.06	0.941 601	54	13.15	0.948 089
15	13.20	0.951 694	35	13.11	0.945 205	55	13.01	0.937 996
16	13.75	0.991 348	36	13.21	0.952 415	56	13.23	0.953 857
17	13.23	0.953 857	37	13.06	0.941 601	57	12.95	0.933 670
18	13.19	0.950 973	38	13.05	0.940 880	58	13.21	0.952 415
19	13.27	0.956 741	39	13.24	0.954 578	59	13.20	0.951 694
20	13.31	0.959 625	40	13.08	0.943 043			

2.3 实验参数设置方法

本文参数通过网格搜索策略选择[10]。SVM、M²SVM 中,惩罚因子 C 在网格{0.01, 0.05, 0.1, 0.5, 1, 5, 10}中搜索选取;M²SVM 中,参数 μ 在网格{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8,0.9}中搜索选取。

2.4 预测模型评价标准

实验对预测效果的评价主要有两个指标:相对误差(RE)和平均绝对百分比误差(MAPE),其定义如下:

$$RE=\frac{x-\tilde{x}}{x}\times 100\%, \tag{18}$$

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\frac{|x-\tilde{x}|}{|x|}\times 100\%, \tag{19}$$

其中 x 表示实际成绩,$\tilde{x}$ 表示预测成绩。

2.5 实验结果

实验结果如表 2 所示。

表 2　SVM 与 M²SVM 预测结果比较

实验批次		SVM			M²SVM		
编号	实际成绩/s	预测成绩/s	绝对误差	$RE(\times 10^{-4})$	预测成绩/s	绝对误差	$RE(\times 10^{-4})$
55	13.01	13.088	−0.078	−59.954	13.012	−0.002	−1.537
56	13.23	13.112	0.118	89.191	13.238	−0.008	−6.047
57	12.95	13.068	−0.118	91.120	12.799	0.151	116.602
58	13.21	13.176	0.034	25.738	13.203	0.007	5.299
59	13.20	13.178	0.022	16.667	13.180	0.020	15.152
$MAPE(\times 10^{-4})$			56.534			28.927	

由表 2 可以看出,与 SVM 相比,M²SVM 对测试样本进行预测时大多较为理想。从整体平均绝对百分比误差看,SVM 的 MAPE 值为 $5.653\ 4\times 10^{-3}$,而 M²SVM 的 MAPE 仅为 $2.892\ 7\times 10^{-3}$,这表明与 SVM 相比,M²SVM 整体成绩测能力有了较大幅度的提升。M²SVM 在成绩预测方面的优势与其工作原理密切相关,即 M²SVM 在分类决策时同时考虑样本的全局信息和局部信息。

3 结论

支持向量机作为一种经典的模式分类方法广泛应用于成绩预测领域。随着应用的深入,该方法泛化能力不

足的问题日益凸显。为了进一步提高其泛化能力,在深入分析流形判别分析的基础上提出基于最小流形类内离散度支持向量机 M^2SVM。该方法在成绩预测时同时考虑样本的全局信息和局部信息,因此比 SVM 具有更优的预测能力。M^2SVM 对研究未来各类运动成绩发展的趋势具有一定的参考价值,并可作为人才选拔、成绩提升、梯队建设的重要辅助工具。

参考文献:

[1] 王斑斑,杜少武,廖蓉蓉,等. 2010 年亚运会刘翔 110 米栏竞争实力的最优组合预测与分析[J]. 天津体育学院学报,2010,25(2):167-170.
Wang T T,Du S W,Liao R R,et al. Prediction and analysis the optimum combination of competitiveness of Liu Xiang in 2010 Asian games[J]. Journal of Tianjin University of Sport,2010,25(2):167-170.

[2] 高云,许晓峰,汪丽华,等. 对我国优秀男子 110 米栏运动员第七栏腾空技术的运动学分析[J]. 武汉体育学院学报,2009,43(11):91-94.
Gao Y,Xu X F,Wang L H,et al. Seventh column techniques of 110 metre hurdles of Chinese elite male hurdlers [J]. Journal of Wuhan Institute of Physical Education ,2009,43(11):91-94.

[3] 郭雪奇,黄勇,王乐军,等. 刘翔 110 米栏分段时间与总成绩的相关关系研究[J]. 成都体育学院学报,2009,35(7):37-39.
Guo X Q,Huang Y,Wang L J,et al. Research on the correlation between the period time and total score in Liu Xiang's 110 meter hurdles[J]. Journal of Chengdu Sport University,2009,35(7):37-39.

[4] Khan N M,Ksantini R,Ahamad I S. A novel SVM+NDA model for classification with an application to face recognition[J]. Pattern Recognition,2012,45(1):66-79.

[5] 邬啸,魏延,吴瑕. 基于混合核函数的支持向量机[J]. 重庆理工大学学报:自然科学版,2011,25(10):66-70.
Wu X,Wei Y,Wu X. Support Vector Machine based on hybrid kernel function[J]. Journal of Chongqing University of Technology:Natural Science,2011,25(10):66-70.

[6] Vapnik V. The nature of statistical learning theory[M]. New York:Springer-Verlag,1995.

[7] 郭辉. 支持向量机选择及其在股票走势预测中的应用[J]. 重庆师范大学学报:自然科学版,2007,24(4):45-49.
Guo H. Application of online selection support vector classification in the prediction of ups and downs in stock market[J]. Journal of Chongqing Normal University: Natural Science Edition,2007,24(4):45-49.

[8] 刘忠宝,潘广贞,赵文娟. 流形判别分析[J]. 电子与信息学报,2013,35(9):204-205.
Liu Z B,Pan G Z,Zhao W J. Manifold-based discriminant analysis[J]. Journal of Electronics & Information Technology,2013,35(9):204-205.

[9] 龙斌. 基于支持向量机的刘翔 110 m 栏成绩预测[J]. 天津体育学院学报,2009,24(4):330-333.
Long B. SVM-based prediction of Liu Xiang's achievements in 110 m hurdle[J]. Journal of Tianjin University of Sport,2009,24(4):330-333.

[10] Muller K R,Mika S,Ratsch G,et al. An introduction to kernel-based learning algorithms[J]. IEEE Transactions on Neural Networks,2001,12(2):181-202.

Research on Prediction of Achievements in 110 m Hurdle based on Minimum Manifold-Based within-Class Scatter Support Vector Machine

ZHANG Yaqing

(Department of Mathematics, Taiyuan College, Taiyuan 030012, China)

Abstract: When dealing with the achievement prediction, SVM (Support Vector Machine) suffers from limitation of generalization capability. In view of this, Manifold-based within-class Scatter Support Vector Machine (M^2SVM) is proposed and is used in the achievement prediction of 110 m hurdle. 59 achievements of Liu Xiang from the year 2000 to 2009 are collected and construct the experimental dataset. Firstly, the first 54 achievements are used as training set and applied to build the prediction model; the last 5 achievements are taken as test set. The effectiveness of M^2SVM is verified by the similarity of the expected results and the actual results. The proposed method M^2SVM is important to talent selection, achievement improvement and echelon construction.

Key words: Support Vector Machine (SVM); minimum manifold-based within-class scatter Support Vector Machine (M^2SVM); 110 m hurdle; achievement prediction

(责任编辑 游中胜)