

基于加权马氏距离的模糊多核支持向量机

戴小路, 汪廷华, 周慧颖

引用本文

戴小路, 汪廷华, 周慧颖. 基于加权马氏距离的模糊多核支持向量机[J]. 计算机科学, 2022, 49(11A): 210800216-5.

DAI Xiao-lu, WANG Ting-hua, ZHOU Hui-ying. [Fuzzy Multiple Kernel Support Vector Machine Based on Weighted Mahalanobis Distance](#) [J]. Computer Science, 2022, 49(11A): 210800216-5.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于FWA-PSO-MSVM的船舶区域配电电力系统故障诊断](#)

Fault Diagnosis of Shipboard Zonal Distribution Power System Based on FWA-PSO-MSVM
计算机科学, 2022, 49(11A): 210800209-5. <https://doi.org/10.11896/jsjcx.210800209>

[一种适于多分类问题的支持向量机加速方法](#)

Acceleration of SVM for Multi-class Classification
计算机科学, 2022, 49(6A): 297-300. <https://doi.org/10.11896/jsjcx.210400149>

[基于改进麻雀搜索优化支持向量机的渔船捕捞方式识别](#)

Fishing Type Identification of Marine Fishing Vessels Based on Support Vector Machine Optimized by Improved Sparrow Search Algorithm
计算机科学, 2022, 49(6A): 211-216. <https://doi.org/10.11896/jsjcx.220300216>

[一种基于支持向量机的主动度量学习算法](#)

Active Metric Learning Based on Support Vector Machines
计算机科学, 2022, 49(6A): 113-118. <https://doi.org/10.11896/jsjcx.210500034>

[基于SVM的类别增量人体活动识别方法](#)

Human Activity Recognition Method Based on Class Increment SVM
计算机科学, 2022, 49(5): 78-83. <https://doi.org/10.11896/jsjcx.210400024>

基于加权马氏距离的模糊多核支持向量机

戴小路 汪廷华 周慧颖

赣南师范大学数学与计算机科学学院 江西 赣州 341000

(1353119318@qq.com)

摘要 模糊支持向量机通过引入模糊隶属度有效区分不同样本的重要程度,降低了传统支持向量机对噪声数据的敏感性。针对基于欧氏距离设计的隶属度函数忽略了样本的总体分布,且未考虑样本特征重要性的区分,提出了一种基于加权马氏距离的模糊支持向量机方法。首先应用 Relief-F 算法计算样本特征权重,然后基于该权重计算样本距其类中心的加权马氏距离,最后根据该距离值度量样本隶属度。在此基础上,考虑到核函数及其核参数难以确定,将模糊支持向量机与多核学习方法相结合,提出基于加权马氏距离的模糊多核支持向量机,采用加权求和形式构建多核,并遵循中心核对齐原则确定每个核的权重。该方法不仅降低了弱相关特征对分类效果的影响,而且使数据表达更加全面准确。实验结果表明,基于加权马氏距离的模糊支持向量机的分类精度高于基于欧氏距离和基于马氏距离的模糊支持向量机,且基于加权马氏距离的模糊多核支持向量机的分类性能较单核模型更优。

关键词: 支持向量机;中心核对齐;加权马氏距离;多核学习;隶属度函数

中图分类号 TP181

Fuzzy Multiple Kernel Support Vector Machine Based on Weighted Mahalanobis Distance

DAI Xiao-lu, WANG Ting-hua and ZHOU Hui-ying

School of Mathematics and Computer Science, Gannan Normal University, Ganzhou, Jiangxi 341000, China

Abstract Fuzzy support vector machine (FSVM) effectively distinguishes the importance of different samples by introducing fuzzy memberships, which reduces the sensitivity of traditional support vector machines to noise data. The membership function designed based on Euclidean distance ignores the overall distribution of samples and does not consider the different importance of sample features. A fuzzy support vector machine method based on weighted Mahalanobis distance is proposed. This method first applies the Relief-F algorithm to estimate the weight of each feature. Then it utilizes the weight for calculating the weighted Mahalanobis distance between the sample and the center of its class. Finally, the fuzzy membership of the sample is calculated based on weighted Mahalanobis distance. Furthermore, considering the difficulty of determining the kernel function and its parameters, a fuzzy multi-kernel support vector machine (FMKSVM) based on weighted Mahalanobis distance is put forward, which combines FSVM with multiple kernel learning methods. The multi-kernel is constructed in the form of weighted sum, and the weight of each kernel is calculated according to the central kernel alignment method (CKA). The proposed method not only reduces the influence of weakly relevant features on classification results, but also enables a more adequate and accurate representation of the data. Experimental results show that, FSVM based on weighted Mahalanobis distance has higher classification accuracy than FSVM based on Euclidean distance and Mahalanobis distance, and the classification performance of FMKSVM based on weighted Mahalanobis distance is superior to that of the single-kernel model.

Keywords Support vector machine, Centered kernel alignment, Weighted Mahalanobis distance, Multiple kernel learning, Membership function

1 引言

支持向量机(Support Vector Machine, SVM)^[1]以结构风险最小化为基准,通过构建最佳分类超平面有效提升学习机的泛化性能,在小样本、非线性问题等方面表现出色,现已广泛应用于诸多实际领域^[2-4]。然而 SVM 对噪声和离群点十分敏感,对此 Lin 等提出了模糊支持向量机(Fuzzy SVM, FS-

VM)^[5],引入模糊隶属度实现样本的重要性区分,对降低异常值的影响卓有成效。大多数 FSVM 的隶属度函数都是基于欧氏距离设计的,未考虑到样本的总体分布, Jiao^[6]提出的基于马氏距离的 FSVM 探索了样本的全局信息,但忽略了样本各特征对训练的贡献不同这一特点,没有对特征的重要程度进行有效区分。基于此,本文提出一种基于加权马氏距离的模糊支持向量机方法。首先采用 Relief-F 算法对各样本

基金项目:国家自然科学基金(61966002);赣南师范大学研究生创新基金项目(YCX20A019)

This work was supported by the National Natural Science Foundation of China(61966002) and Graduate Student Innovation Fund Project of Gannan Normal University(YCX20A019).

通信作者:汪廷华(wthpku@163.com)

特征的重要程度进行评估,得到相应权重,并根据该权重求得各样本距其类中心的加权马氏距离,进而计算样本的模糊隶属度值。

另外,针对 SVM 核函数选择困难以及对多源异构数据解释性不足等缺陷,多核学习(Multiple Kernel Learning, MKL)^[7-8]运用多个或多种核函数的组合形式,使数据的相似性描述更加准确,模型灵活性更强、可解释性更高。模糊多核支持向量机(Fuzzy Multiple Kernel SVM, FMKSVM)^[9]将 FSVM 与 MKL 相结合,同时利用隶属度函数和多核组合形式更大程度地改善了模型性能,给 SVM 的研究提供更多元的设计思路^[10-14]。将基于加权马氏距离的模糊支持向量机方法拓展到多核维度,本文提出基于加权马氏距离的模糊多核支持向量机,以加权求和的方式构建多核组合函数,其中多核权重系数依据中心核对齐方法进行判定。通过实验对比,该方法的分类精度优于单核的 FSVM,并且基于加权马氏距离的 FSVM 比基于欧氏距离和基于马氏距离的 FSVM 分类准确率更高,说明了所提方法行之有效。

2 模糊支持向量机

以二分类问题为例,假设训练集:

$$T = \{(\mathbf{x}_1, y_1, s_1), \dots, (\mathbf{x}_l, y_l, s_l)\}, \mathbf{x}_i \in R^N$$

$$y_i \in \{-1, +1\}, s_i \in [\epsilon, 1], i = 1, \dots, l$$
(1)

其中, \mathbf{x}_i 代表样本点, y_i 表示样本的类别标签, 模糊隶属度 s_i 意味着样本 \mathbf{x}_i 归属于某一类的可靠程度, ϵ 为一个任意小的正数。与标准支持向量机类似, 模糊支持向量机算法的目标也是寻找一个最优分类超平面使分类间隔最大化, FSVM 求解最优分类面可以形式化为如下的优化问题:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l s_i \xi_i$$

$$\text{s. t. } y_i (\mathbf{w}^\top \phi(\mathbf{x}_i) + b) - 1 + \xi_i \geq 0$$

$$\xi_i \geq 0, i = 1, \dots, l$$
(2)

其中, \mathbf{w}, b 分别代表分类超平面的法向量和偏移量; 惩罚因子 C 为常数, 用于控制超平面间距最大化和误差最小化之间的平衡, C 值越大, 对误分类样本的约束就越大; ξ_i 表示分类误差项, 加权误差项 $s_i \xi_i$ 用于衡量权重不等的样本的误分类程度, 样本隶属度越小, 加权误差项 $s_i \xi_i$ 的作用就越小, 则样本分类贡献率越低。

与标准 SVM 求解方法类似, 通过构建拉格朗日函数和鞍点条件获得 FSVM 原问题的对偶形式:

$$\max \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s. t. } \sum_{i=1}^l \alpha_i y_i = 0, 0 \leq \alpha_i \leq s_i C$$
(3)

其中, $K(\mathbf{x}_i, \mathbf{x}_j)$ 为核函数, 故而得出对应最优分类面的判别函数为:

$$f(\mathbf{x}) = \text{sign}(\sum_{i=1}^l \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b)$$
(4)

3 加权马氏距离模糊隶属度

3.1 Relief-F 特征加权

Relief 算法是 Kira 等^[15]在 20 世纪 90 年代提出的一种过滤式特征选择算法, 最初主要针对二分类问题。该算法根据同类及异类近邻样本间的特征距离差异来判定其重要程度, 当某一特征在同类近邻样本间的距离很小, 而在不同类的

最近邻样本之间距离差异很大时, 说明该特征于分类有益, 则赋予该特征较大的权重; 反之使该特征权重减小。Relief 算法的时间复杂度只受样本的特征总数及抽样次数影响, 十分高效。Relief-F 算法^[16](具体见算法 1) 拓展了 Relief 算法的适用范围, 能够实现多类别样本的特征处理, 特征权重计算式如下:

$$W(A) = W(A) - \frac{\sum_{j=1}^q \text{diff}(A, \mathbf{x}, H_j)}{nq} + \dots +$$

$$\frac{\sum_{y \in \text{class}(\mathbf{x})} \left[\frac{p(y) \times \sum_{j=1}^q \text{diff}(A, \mathbf{x}, M_j(y))}{1 - p(\text{Class}(\mathbf{x}))} \right]}{nq}$$
(5)

其中, n 代表样本的抽样次数, q 为最近邻样本个数, $p(y)$ 表示样本属于类别 y 的概率, H_j 是类别相同的最近邻样本集, $M_j(y)$ 则为类 y 的异类近邻样本集合, $\text{diff}(A, \mathbf{x}_1, \mathbf{x}_2)$ 表示样本 \mathbf{x}_1 和样本 \mathbf{x}_2 在特征 A 上的差, 其计算式如下:

$$\text{diff}(A, \mathbf{x}_1, \mathbf{x}_2) = \begin{cases} \frac{|\mathbf{x}_1[A] - \mathbf{x}_2[A]|}{\max(A) - \min(A)}, & \text{if } A \text{ is continuous} \\ 0, & \text{if } A \text{ is discrete and } \mathbf{x}_1[A] = \mathbf{x}_2[A] \\ 1, & \text{if } A \text{ is discrete and } \mathbf{x}_1[A] \neq \mathbf{x}_2[A] \end{cases}$$
(6)

其中, $\mathbf{x}[A]$ 表示样本 \mathbf{x} 在特征 A 上的取值。

算法 1 Relief-F 算法

输入: 训练数据集 T , 样本抽样次数 n , 最近邻样本个数 q

输出: 特征权重 W

1. 将每个特征的权重初始化为 0
2. 从 T 中随机抽取任一样本 \mathbf{x} , 查找 q 个与 \mathbf{x} 类别相同的最近邻样本, 构建同类最近邻样本集 $H_j (j=1, \dots, q)$, 同理从 \mathbf{x} 的每个异类样本集中搜寻 k 个最近邻得到异类近邻样本集 $M_j(y)$;
3. 依据式(5)对特征权重进行更新;
4. 重复步骤 2 和步骤 3 n 次, 取 n 次计算结果的平均值作为最终特征权重。

3.2 加权马氏距离

大多隶属度构造都是基于样本间的欧氏距离进行判定, 这种方式忽略了样本属性之间的关系, 未考虑到样本总体分布对距离造成的影响。不同于欧氏距离, 马氏距离^[17]是基于样本的全局信息计算出来的, 代表数据的协方差距离, 可以有效解除属性间的关联性干扰, 且独立于样本数据的测量尺度, 消除了量纲性。马氏距离计算如下:

$$\text{dist}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$
(7)

其中, \mathbf{S} 为样本的总体协方差矩阵。

根据 Relief-F 算法求得的样本特征权重构造特征加权矩阵 \mathbf{P} :

$$\mathbf{P} = \begin{bmatrix} W_1 & & & \\ & W_2 & & \\ & & \ddots & \\ & & & W_N \end{bmatrix}$$
(8)

上述矩阵 \mathbf{P} 为对角矩阵, 其中 W_i 表示特征 i 的权重, N 代表特征的总个数。则样本与类中心的加权马氏距离计算式为:

$$\text{dist}^p(\mathbf{x}_i, \bar{\mathbf{x}}) = \sqrt{\sum_{j=1}^N (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)^\top (\mathbf{S} * \mathbf{P})^{-1} \sum_{j=1}^N (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)}$$
(9)

其中, $*$ 表示矩阵的乘积运算, 基于加权马氏距离的隶属度

函数根据样本点到其所属类中心的加权马氏距离确定样本重要性,距离越小,则赋予该样本越大的隶属度值。基于加权马氏距离的隶属度函数表达式如下:

$$s_i = 1 - \frac{\text{dist}^p(x_i, \bar{x})}{r} \quad (10)$$

其中, $r = \max \text{dist}^p(x_i, \bar{x})$ 表示类半径, \bar{x} 为样本类中心。

4 基于中心核对齐的多核构造

核函数的选取于 SVM 而言十分关键,但现阶段该方面尚无明确的理论依据,通常根据经验或采用穷举法选取分类性能最优的核函数。常用的核函数如表 1 所列。

表 1 常用核函数

Table 1 Common kernel functions

核函数	公式
线性核	$K(x_i, x_j) = (x_i \cdot x_j)$
多项式核	$K(x_i, x_j) = (x_i \cdot x_j + 1)^d, d \geq 1$
径向基核	$K(x_i, x_j) = \exp(- x_i - x_j ^2 / 2\sigma^2), \sigma > 0$
Sigmoid 核	$K(x_i, x_j) = \tanh(\kappa x_i^T x_j - \delta), \kappa > 0, \delta < 0$

考虑到核选择的困难性及单核学习的局限性,多核学习利用多个核函数不同的映射能力使数据表达更加准确,模型更加灵活,本文将基于加权马氏距离的模糊支持向量机扩展到多维特征空间,采用加权求和的方式^[18]构建多核:

$$K_{\mu}(x_i, x_j) = \sum_{m=1}^M \mu_m K_m(x_i, x_j) \quad (11)$$

s. t. $\sum_{m=1}^M \mu_m = 1, 0 \leq \mu_m \leq 1,$

其中, K_m 表示不同的核函数,总数为 M 个, μ 代表多核对应的权重向量。

本文采用启发式方法计算多核权重系数,提出基于中心核对齐(Centered Kernel Alignment,CKA)^[19]的多核学习方法。利用每个核与理想核之间 CKA 的大小来确定其相应的核权重系数,当某一核相对于理想核的中心核对齐值越大时,说明其对分类贡献率越大,则为该核分配较大的权重。核对齐(Kernel Target Alignment,KTA)作为两个核矩阵之间相似度的度量准则,定义如下:

$$KTA(K_1, K_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \quad (12)$$

其中, K_1 与 K_2 分别表示核函数 K_1 和 K_2 映射后的核矩阵, $\langle \cdot, \cdot \rangle_F$ 为两矩阵间的 Frobenius 内积:

$$\langle K_1, K_2 \rangle_F = \sum_{i=1}^l \sum_{j=1}^l K_1(x_i, x_j) K_2(x_i, x_j) \quad (13)$$

CKA 在 KTA 的基础上增加了对核矩阵的中心化操作,可有效解决样本在特征空间距原点较远及类分布不平衡等问题,使数据得到更加充分的表达。核矩阵 K 的中心化矩阵定义如下:

$$\bar{K} = HKH \quad (14)$$

其中,矩阵 $H = \mathbf{I} - \mathbf{e}\mathbf{e}^T/l, \mathbf{I}$ 表示 l 阶单位矩阵, \mathbf{e} 为 l 维的全 1 列向量。则 K_1 与 K_2 之间的中心核对齐定义为:

$$CKA(K_1, K_2) = \frac{\langle \bar{K}_1, \bar{K}_2 \rangle_F}{\sqrt{\langle \bar{K}_1, \bar{K}_1 \rangle_F \langle \bar{K}_2, \bar{K}_2 \rangle_F}} \quad (15)$$

根据核函数与理想核之间的中心核对齐值判定其对应的核权重:

$$\mu_i = \frac{CKA(K_i, Y)}{\sum_{m=1}^M CKA(K_m, Y)} \quad (16)$$

其中,标签矩阵 $Y = \mathbf{y}\mathbf{y}^T$ 是基于数据集 T 的理想核矩阵, $\mathbf{y} = [y_1, \dots, y_l]^T$ 。将求得的多核权重系数 $\mu = [\mu_1, \dots, \mu_M]^T$ 代入式(11)可得组合核函数。将组合核函数引入模糊支持向量机中即得基于加权马氏距离的模糊多核支持向量机,其优化问题变为:

$$\min \frac{1}{2} \left(\sum_{m=1}^M \mu_m \|w_m\|_2 \right)^2 + C \sum_{i=1}^l s_i \xi_i$$

s. t. $y_i \left(\sum_{m=1}^M w_m^T \phi_m(x_i) + b \right) - 1 + \xi_i \geq 0$ (17)

$$\sum_{m=1}^M \mu_m = 1, \mu_m \geq 0, \xi_i \geq 0$$

其中, μ_m 表示第 m 个核函数的权重,映射函数 $\phi_m(x)$ 将样本映射到 M 个高维特征空间中:

$$f_m: x \rightarrow \phi_m(x), m = 1, \dots, M \quad (18)$$

决策函数相应如下:

$$f(x) = \text{sign} \left(\sum_{i=1}^l \alpha_i y_i K_{\mu}(x_i, x) + b \right) \quad (19)$$

其中, $K_{\mu}(x_i, x_j)$ 为多核组合函数。

基于加权马氏距离的模糊多核支持向量机算法的主要算法框架如下:

- (1)数据预处理,划分训练集与测试集;
- (2)基于 Relief-F 方法确定特征权重,构建特征加权矩阵 \mathbf{P} ;
- (3)构造特征加权马氏距离隶属度函数,计算样本隶属度值 s ;
- (4)选择基础核函数,计算中心核对齐值,得到核权重 μ ,构建加权求和的多核组合形式;
- (5)通过交叉验证法寻找参数 C 的最优值;
- (6)型的训练与预测,相关性评估。

5 实验结果与分析

为验证所提方法的有效性,从 UCI 机器学习数据库^[20]中选取了 9 个数据集,数据信息如表 2 所列。

表 2 数据集信息

Table 2 Data set information

序号	数据集	样本总数	特征数	类别数
1	WBC	683	9	2
2	Vertebral	310	6	2
3	Pima	768	8	2
4	Ionosphere	351	33	2
5	Heart	270	13	2
6	Climate	540	18	2
7	Australian	690	15	2
8	Seeds	210	7	3
9	Balance-scale	625	4	3

实验开发环境为 Matlab R2019a,计算机配置为 i5 处理器、CPU 频率 2.4GHz、内存(RAM)4GB、操作系统为 Windows7。在上述 9 个数据集上分别对本文提出的基于加权马氏距离的模糊支持向量机(WM-FSVM)与传统支持向量机(SVM)、基于欧氏距离的 FSVM(E-FSVM)、基于马氏距离的 FSVM(M-FSVM)的分类性能进行对比实验,同时将基于加权马氏距离的模糊多核支持向量机(WM-FMKSVM)与基于欧氏距离的 FMKSVM(E-FMKSVM)、基于马氏距离的 FMKSVM(M-FMKSVM)进行比较,以分类精度作为性能评价指标。选用多个不同核参数的径向基核与多项式核组建多核函数,其中高斯核参数设置为 $\{2^{0.5}, 2^1, 2^2, 2^5, 2^7\}$,

$2^{10}, 2^{12}, 2^{15}, 2^{17}, 2^{20}$ }, 多项式核参数为 $\{1, 2, 3\}$ 。惩罚参数 C 的选取由交叉验证法确定, 其取值范围为 $C = \{2^{-5}, 2^{-3}, \dots, 2^{13}, 2^{15}\}$ 。实验中所有数据都经过标准化处理, 训

练集与测试集分别按 3:7, 5:5, 7:3 的比例进行基于类别比例的随机划分, 实验结果以 10 次实验的均值为准, 结果分别如表 3—表 5 所列。

表 3 分类准确率对比(3:7)

Table 3 Comparison of classification accuracy(3:7)

(单位:%)

Algorithm	SVM	E-FSVM	M-FSVM	WM-FSVM	E-FMK SVM	M-FMK SVM	WM-FMK SVM
data 1	93.89	94.26	95.41	96.53	96.03	96.66	97.70
data 2	82.49	82.95	83.87	84.79	84.41	85.54	86.02
data 3	74.42	74.35	75.46	76.21	76.32	77.15	77.69
data 4	91.24	92.06	92.87	93.68	93.50	94.22	95.03
data 5	81.42	81.48	82.54	83.60	83.13	84.19	85.24
data 6	91.94	92.37	93.19	94.07	93.65	94.24	95.03
data 7	83.82	83.65	84.17	85.30	85.26	86.14	86.82
data 8	91.47	92.16	92.83	93.51	93.45	94.12	94.87
data 9	92.95	93.47	94.24	95.05	94.91	95.46	96.29

表 4 分类准确率对比(5:5)

Table 4 Comparison of classification accuracy(5:5)

(单位:%)

Algorithm	SVM	E-FSVM	M-FSVM	WM-FSVM	E-FMK SVM	M-FMK SVM	WM-FMK SVM
data 1	94.44	95.32	95.91	96.84	96.57	97.35	98.54
data 2	83.87	84.16	85.30	86.45	86.03	86.84	87.69
data 3	75.48	75.26	76.30	77.49	77.63	78.56	79.21
data 4	92.48	92.97	93.81	94.89	94.63	95.59	96.14
data 5	82.72	83.13	84.24	85.19	84.97	85.52	86.63
data 6	92.59	93.33	94.04	95.19	94.93	96.13	97.04
data 7	84.76	84.54	85.35	86.99	86.51	87.44	88.31
data 8	92.43	92.83	93.33	94.29	94.24	95.19	96.14
data 9	93.81	94.59	95.87	96.51	96.49	97.76	98.72

表 5 分类准确率对比(7:3)

Table 5 Comparison of classification accuracy(7:3)

(单位:%)

Algorithm	SVM	E-FSVM	M-FSVM	WM-FSVM	E-FMK SVM	M-FMK SVM	WM-FMK SVM
data 1	95.12	95.61	96.29	97.07	96.96	98.05	99.02
data 2	84.74	85.25	86.02	87.16	87.07	88.46	89.24
data 3	76.76	76.54	77.09	78.57	78.32	79.26	80.12
data 4	93.45	93.80	94.29	95.32	95.17	96.34	97.21
data 5	83.95	84.11	85.32	86.30	86.49	87.12	88.36
data 6	93.21	94.71	95.06	96.01	95.83	96.72	98.15
data 7	85.62	85.76	86.97	87.64	87.42	88.38	89.35
data 8	93.06	93.66	94.56	95.24	95.60	96.28	97.41
data 9	96.41	97.30	97.88	98.93	98.49	99.01	100.00

从实验结果可以看出, 本文提出的基于加权马氏距离的 FSVM 方法在选定的 9 个数据集上的分类准确率均高于传统 SVM, 也优于基于欧氏距离和基于马氏距离的 FSVM 方法。同时, 多核的加权马氏距离 FSVM 方法的分类性能较单核 FSVM 而言更为优异, 也比基于欧氏距离和马氏距离的 FMKSVM 分类效果更好。纵向对比整体实验, 算法的分类精度随着训练集占比的增多而明显提升, 这说明, 就同一数据集而言, 划分的训练集越多, 模型训练越充分, 学习效果越好。但不排除随着训练集的增多会出现过拟合现象, 导致模型泛化性能变差, 因此在模型训练中要合理设置训练集的大小。上述结果进一步证实了 WM-FMK SVM 的有效性, 它进一步提升了支持向量机算法的分类精度, 使模型分类性能更优, 更具优势。

结束语 针对传统支持向量机对噪声的敏感性及核选择的困难性, 本文通过分析支持向量机的特性, 将 Relief-F 算法与中心核对齐方法应用于 SVM 中, 提出了一种基于加权

马氏距离的模糊多核支持向量机方法, 以降低弱相关或不相关特征对分类效果的影响, 减弱噪声的影响并避免核选择难题, 使数据表达更充分。相关实验验证了改进算法在一定程度上提升了分类性能, 说明了其可行性与有效性。目前关于隶属度构造及多核学习尚未出现标准化的准则, 现有方法各具利弊。因此, 进一步改善模糊隶属度计算方法以及探索更好的多核组合形式是我们今后努力的研究方向。同时, 鉴于本文主要集中于 SVM 同类策略算法间的横向比较, 在算法多元性上略有欠缺, 后续也将着眼于最新非 SVM 算法的探索与研究, 以纵向对比增强相关改进算法的分类性能。

参考文献

- [1] CORTES C, VAPNIK V N. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.
- [2] WANG M W, WAN Y C, YE Z W, et al. Remote sensing image classification based on the optimal support vector machine and

- modified binary coded ant colony optimization algorithm[J]. Information Sciences, 2017, 402(9): 50-68.
- [3] DAGHER I, DAHDAH E, SHAKIK M A. Facial expression recognition using three-stage support vector machines[J]. Visual Computing for Industry Biomedicine and Art, 2019, 2(1): 24.
- [4] HOSSEINZADEH M, RAHMANI A M, VO B, et al. Improving security using SVM-based anomaly detection: issues and challenges[J]. Soft Computing, 2021, 25(10): 3195-3223.
- [5] LIN C F, WANG S D. Fuzzy support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471.
- [6] JIAO Y P. Application of FSVM based on mahalanobis distance for user behavior recognition[J]. Science & Technology Vision, 2015, 81(21): 113-115.
- [7] BACH F R, LANCKRIET G R G, JORDAN M I. Multiple kernel learning, conic duality, and the SMO algorithm[C] // Proceedings of the Twenty-First International Conference on Machine Learning. New York: Association for Computing Machinery, 2004: 41-48.
- [8] WANG H Q, SUN F C, CAI Y N, et al. On multiple kernel learning methods[J]. Acta Automatica Sinica, 2010, 36(8): 1037-1050.
- [9] DAI X L, WANG T H, HU Z W. Research progress of fuzzy multiple kernel support vector machine [J]. Application Research of Computers, 2021, 38(10): 2896-2903.
- [10] WANG T H, LU J, ZHANG G Q. Two-stage fuzzy multiple kernel learning based on hilbert-schmidt independence criterion [J]. IEEE Transactions on Fuzzy Systems, 2018, 26(6): 3703-3714.
- [11] HE Q, ZHANG J Y. Kernel-target alignment multi-kernel fuzzy support vector machine[J]. CAAI Transactions on Intelligent Systems, 2019, 14(6): 1163-1169.
- [12] WANG Z, WANG B L, CHENG Y, et al. Cost-sensitive fuzzy multiple kernel learning for imbalanced problem[J]. Neurocomputing, 2019, 366(13): 178-193.
- [13] HE Q, ZHANG Q S, WANG H Y, et al. Local similarity-based fuzzy multiple kernel one-class support vector machine[J]. Complexity, 2020, 2020(10): 8853277.
- [14] ZHANG Q S, HE Q, ZHANG C L, et al. Fuzzy multiple kernel one-class support vector machine[J]. Journal of Beijing University of Civil Engineering and Architecture, 2020, 36(1): 82-90.
- [15] KIRA K, RENDELL L A. A practical approach to feature selection[J]. Machine Learning Proceedings, 1992, 48(1): 249-256.
- [16] ZUO Y H, JIA L Y, YOU J G, et al. Classification algorithm of fuzzy support vector machine based on Relief-F feature weighting[J]. Control and Instruments in Chemical Industry, 2019, 46(10): 834-838, 864.
- [17] MA L L, GUO K J, WANG J Z. A fault diagnosis method of vehicle transmission system based on improved SVM[J]. Transactions of Beijing Institute of Technology, 2020, 40(8): 856-860.
- [18] ZHOU Q, WU J W, FAN Y W, et al. Weighted linear multiple kernel learning for saliency detection[C] // Proceedings of the 2nd EAI International Conference on Robotic Sensor Networks. Cham: Kitakyushu, 2020: 201-213.
- [19] WANG T H, QIU Y Z, HUA J L. Centered kernel alignment inspired fuzzy support vector machine[J]. Fuzzy Sets and Systems, 2020, 394(9): 110-123.
- [20] DUA D, GRAFF C. UCI machine learning repository[EB/OL]. <http://archive.ics.uci.edu/ml>.



DAI Xiao-lu, born in 1997, postgraduate. Her main research interests include machine learning and data mining.



WANG Ting-hua, born in 1977, Ph.D. professor, is a member of China Computer Federation. His main research interests include artificial intelligence and machine learning.