软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(10):4735−4752 [doi: 10.13328/j.cnki.jos.007306] [CSTR: 32375.14.jos.007306]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向集值数据的孪生支持函数机*

梁志贞, 闵玉寒, 丁世飞

(中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116)

通信作者: 梁志贞, E-mail: liang@cumt.edu.cn

摘　要: 孪生支持向量机 (twin support vector machine, TSVM) 能有效地处理交叉或异或等类型的数据. 然而, 当处理集值数据时, TSVM 通常利用集值对象的均值、中值等统计信息. 不同于 TSVM, 提出能直接处理集值数据的孪生支持函数机 (twin support function machine, TSFM). 依据集值对象定义的支持函数, TSFM 在巴拿赫空间取得非平行的超平面. 为了抑制集值数据中的离群点, TSFM 采用了弹球损失函数并引入了集值对象的权重. 考虑到 TSFM 是无穷维空间的优化问题, 测度采用狄拉克测度的线性组合的形式, 这构建有限维空间的优化模型. 为了有效地求解优化模型, 利用采样策略将模型转化成二次规划 (quadratic programming, QP) 问题并推导出二次规划问题的对偶形式, 这为判断哪些采样点是支持向量提供了理论基础. 为了分类集值数据, 定义集值对象到巴拿赫空间的超平面的距离并由此得出判别规则. 也考虑支持函数的核化以便取得数据的非线性特征, 这使得提出的模型可用于不定核函数. 实验结果表明 TSFM 能获取交叉类型的集值数据的内在结构并且在离群点或集值对象包含少量高维事例的情况下取得了良好的分类性能.

关键词: 支持函数; 采样策略; 核函数; 判决规则; 集值数据

中图法分类号: TP18

中文引用格式: 梁志贞, 闵玉寒, 丁世飞. 面向集值数据的孪生支持函数机. 软件学报, 2025, 36(10): 4735–4752. http://www.jos.org.cn/1000-9825/7306.htm

英文引用格式: Liang ZZ, Min YH, Ding SF. Twin Support Function Machine for Set-valued Data. Ruan Jian Xue Bao/Journal of Software, 2025, 36(10): 4735–4752 (in Chinese). http://www.jos.org.cn/1000-9825/7306.htm

Twin Support Function Machine for Set-valued Data

LIANG Zhi-Zhen, MIN Yu-Han, DING Shi-Fei

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: Twin support vector machine (TSVM) can effectively tackle data such as cross or XOR data. However, when set-valued data are handled, TSVM usually makes use of statistical information of set-valued objects such as the mean and the median. Unlike TSVM, this study proposes twin support function machine (TSFM) that can directly deal with set-valued data. In terms of support functions defined for set-valued objects, TSFM obtains nonparallel hyperplanes in a Banach space. To suppress outliers in set-valued data, TSFM adopts the pinball loss function and introduce the weights of set-valued objects. Considering that TSFM involves optimization problems in the infinite-dimensional space, the measure is taken in the form of a linear combination of Dirac measures. Thus the optimization model in the finite-dimensional space is constructed. To solve the optimization model effectively, this study employs the sampling strategy to transform the model into quadratic programming (QP) problems. The dual formulations of the QP problems are derived, which provides theoretical foundations for determining which sampling points are support vectors. To classify set-valued data, the distance from the set-valued object to the hyperplane in a Banach space is defined, and the decision rule is derived therefrom. This study also considers the kernelization of support functions to capture the nonlinear features of data, which makes the proposed model available for indefinite kernels. Experimental results demonstrate that TSFM can capture the intrinsic structure of cross-plane set-valued data and obtain good classification performance in the case of outliers or set-valued objects containing a few high-dimensional examples.

Key words: support function; sampling strategy; kernel function; decision rule; set-valued data

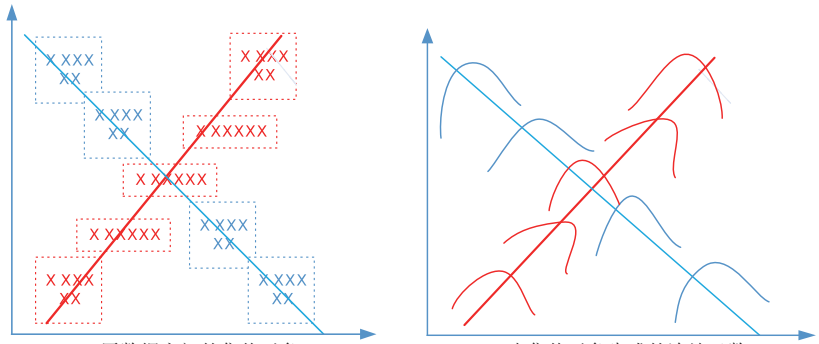
* 收稿时间: 2024-03-17; 修改时间: 2024-07-18; 采用时间: 2024-10-12; jos 在线出版时间: 2025-02-26
CNKI 网络首发时间: 2025-02-27

支持向量机 (support vector machine, SVM) 是一种监督的学习模型, 它主要用于数据分类任务[1−3]. 支持向量机在结构风险最小化原则下搜索最优的分类超平面. 归因于支持向量机模型的凸优化性质和良好的泛化性能, 它已被成功应用于视频分析[4]、异常点检测[5]、不平衡数据[6]等多个领域.

与支持向量机不同, 孪生支持向量机[7]寻找两个非平行的超平面, 它能有效地处理交叉或异或等类型的数据. 对于孪生支持向量机, 每类的样本逼近其中一个超平面, 并且远离另一个超平面. 在 SVM 和 TSVM 的基础上, 许多模型[8−10]被提出以增强原有模型的性能. 非平行的支持向量机[11]不仅利用不敏感损失函数来维持样本的稀疏性, 而且能避免矩阵逆运算. 弹球损失函数的 SVM 能减弱异常点的影响, 而基于分位数距离的 TSVM[12]同样能抑制数据中的异常点. 弹球损失函数的 TSVM 被推广到更加广义的损失函数[13], 这使得弹球损失函数的参数对不同类的样本是不同的. 为了维持样本的稀疏性, 具有不敏感区的孪生支持向量机[14]被提出. 与 TSVM 相比, 它不仅维持样本的稀疏性, 而且对噪声不敏感. 为了充分利用样本的几何性质, 基于弹性网的非平行的支持向量机[15]被提出. 为了有效地抑制数据中的异常点, 直觉模糊孪生支持向量机[16]被提出. 直觉模糊孪生支持向量机为每类样本构建了直觉模糊集并使用直觉模糊集定义样本的权重. 为了处理数据漂移问题, 聚合算子被融合到模糊支持向量机[17]. 非对称的对偶回归模型[18] 利用 TSVM 和可能性回归分析构建了新颖的超平面学习模型.

在实际应用中, 人们可能在不同的粒度下探索同一对象, 即同一对象的属性在不同的粒度下具有不同的值. 例如某人的头街有多个、可能掌握多种外语、爱好有多个、物体的深度特征包含多个神经网络层的输出等. 这样许多对象的属性是用多个值来描述的, 即采用集合的概念描述对象的属性, 从而形成集值数据[19−22]. 集值数据是由集值对象构成, 而一个集值对象通常包含多个数据点或事例, 如图 1(a) 中每个虚线框内的数据点构成一个集值对象. 集值数据已在数据挖掘[20]和决策系统[21,22]中被探索. 在集值数据分析中, 目前已经提出许多集值数据分类的方法[23−26]. 二阶锥规划 (second-order cone programming, SOCP) 方法[23]把集值对象建模为具有二阶矩的随机向量. 高斯分布的支持向量机 (support vector machine with Gaussian distribution, SVMG)[24]把集值对象建模为高斯分布的随机向量. 不确定感知的孪生支持向量机 (uncertainty-aware TSVM, UTSVM)[25]利用二个超平面处理二元分类问题. 具有分布输入的模糊孪生支持向量机 (fuzzy TSVM with distribution input, FTSVMD)[26]通过建模输入为高斯分布的随机向量来处理集值数据. 支持测度机 (support measure machine, SMM)[27]首先利用概率分布建模集值对象并定义集值对象之间的相似度, 接着利用支持向量机分类数据. 如果采用 SOCP, SVMG, UTSVM, FTSVMD 和 SMM 方法来处理集值对象, 需要对集值对象进行概率建模. 如果集值对象包含少量数据点, 那么对集值对象进行概率建模可能是不可靠的或不准确的, 而且通常集值对象中事例的概率分布未知. 不同于概率建模的方式, 支持函数机 (SFM)[28,29]利用集合的支持函数将集值对象转化成连续函数. 由连续函数空间构成巴拿赫 (Banach) 空间[30], 在此基础上构建了基于铰链损失函数的超平面学习模型. 为了探索输出为三角模糊数的形式, 可能性测度被用来推导出模糊支持函数机 (fuzzy SFM, FSFM)[31], FSFM 考虑了模糊类样本的隶属度. FSFM 和 SFM 的主要区别在于前者将样本的标签建模为三角模糊数而后者的标签是标量形式. FSFM 也不需要为集值对象进行概率建模.

然而, 现实世界可能存在图 1 所示的交叉类型的集值数据. 图 1 表示了交叉类型的集值数据及其生成的连续函数, 每一种颜色表示一类集值对象. 从图 1 可观测到两类集值数据位于两条交叉线附近. 图 1(a) 中的每个虚线框内的数据点构成一个集值对象, 图 1(b) 中的每条曲线表示一个连续函数, 这样一个集值对象被转化成巴拿赫空间的连续函数. 从图 1 可知需要两个超平面拟合这种类型的集值数据. 受孪生支持向量机和支持函数机的启发, 本文提出了一种新颖的超平面学习模型, 即孪生支持函数机 (twin support function machine, TSFM), 它能直接处理集值数据. TSFM 在二元分类问题上可取得两个非平行的超平面. 当实施 TSFM 时, 需要将集值对象转化成连续函数. 与 SFM 不同, TSFM 采用了弹球损失函数并利用两个非平行的超平面拟合集值数据. 原模型是无穷维空间的优化问题. 为了解决这个问题, 通过将测度空间限定为由狄拉克测度的线性组合形成的空间来取得有限维空间的优化模型. 通过采样策略将模型转化成二次规划问题, 并利用二次规划的优化算法求解转化后的模型, 这样本文的算法不同于求解 SFM 的算法. 为了对 SVM, SFM, TSVM 和 TSFM 进行比较, 表 1 列出了它们的区别与联系. 从表 1 可看出, TSFM 需要解决二次规划问题, 而 SFM 处理线性规划问题. 由于 TSFM 采用了测度的总变分作为正则化项, 所以它对采样点保持稀疏性. 简言之, 本文的主要贡献如下.



(a) 原数据空间的集值对象 (b) 由集值对象生成的连续函数

图 1 二维空间的两类集值数据和对应的连续函数

表 1 SVM, SFM, TSVM 和 TSFM 的特性

分类器	样本属性	支持向量的稀疏性	特征空间	目标函数	超平面的个数
SVM	向量形式	稀疏	希尔伯特空间	二次规划问题	1
SFM	集合形式	稀疏	巴拿赫空间	线性规划问题	1
TSVM	向量形式	半稀疏	希尔伯特空间	二次规划问题	2
TSFM	集合形式	稀疏	巴拿赫空间	二次规划问题	2

(1) 提出了一种新颖的超平面学习模型对集值数据进行分类, 它利用了弹球损失函数并考虑了集值对象的权重;

(2) 将测度空间限定为狄拉克测度的线性组合构成的空间, 这使得无穷维空间的优化问题被转化成有限维空间的优化问题, 并利用采样策略将模型转化成二次规划问题, 同时推导出二次规划问题的对偶形式以及讨论了基于核函数的支持函数;

(3) 在合成的集值数据和真实数据集上执行了一些实验, 实验结果表明了 TSFM 能有效地分类集值数据.

本文第 1 节简要回顾支持函数机. 第 2 节首先引入 TSFM 的优化模型并讨论 TSFM 的一些性质, 随后将其转化为易处理的模型并将支持函数扩展到核函数. 第 3 节在许多数据集上评估 TSFM 的性能. 最后一节给出结论和展望.

1 支持函数机

对二类集值数据, 令 n 个集值对象表示为 $\{(A_i,y_i),i=1,\ldots,n\}$, 其中 A_i 是 R^m 中包含多个元素的子集, $y_i\in\{-1,1\}$ 是集值对象的标签. 当 A_i 只包含一个数据点时, 集值数据退化为向量值数据. 为了有效地处理集值数据, 支持函数机[28]被提出. 与向量值数据不同, 集值数据的特征是用集合的概念来描述的. 对于支持函数机, 需要下面的定义和定理.

定义 1. 设 A 是 R^m 中的非空闭集. 与集合 A 相关的支持函数 $\sigma_A(\boldsymbol{x})$ 被定义为:

$$\sigma_A(\boldsymbol{x})=\sup\{<\boldsymbol{x},\omega>,\omega\in A\} \tag{1}$$

其中, $<\boldsymbol{x},\omega>$ 表示 $\boldsymbol{x}$ 和 ω 的内积运算. 支持函数也被称为支撑函数, 为了概念的一致性, 本文采用了支持函数的概念. 支持函数 $\sigma_A(\boldsymbol{x})$ 是凸的、齐次的和次加性的函数. 利用支持函数的定义可知 $\sigma_A(\boldsymbol{x})=\sigma_{co(A)}(\boldsymbol{x})$, 其中 $co(A)$ 表示 A 的凸包. 定义 1 实际上构建了集合和函数之间的关系. 从支持函数的定义知 $\sigma_A(\boldsymbol{x})$ 是关于 $\boldsymbol{x}$ 的连续函数. 考虑由集合 X 上的连续函数构成的函数空间, 表示为 $C(X)$. 通过定义连续函数的范数, $C(X)$ 形成了一个巴拿赫 (Banach) 空间[30]. 巴拿赫空间是一个完备的赋范线性空间, 也就是一个具有范数和度量完备性的线性空间. 在这个空间中, 任何一个柯西序列都有一个极限, 且该极限也在这个空间中. 与希尔伯特空间不同, 巴拿赫空间缺乏内积运算, 但在巴拿赫空间可构造线性泛函来获得它的对偶空间. 以下定理表明了连续函数空间 $C(X)$ 和对偶空间的关系.

定理 1[28,30]. 设 X 是局部紧豪斯多夫 (Hausdorff) 空间, $C(X)$ 表示在 X 边界上为 0 的连续函数空间. 对于 $C(X)$ 中的任意连续函数 $\sigma(\boldsymbol{x})$ 都存在正则的博雷尔 (Borel) 测度 μ 使得公式 (2) 成立:

$$\phi(\sigma)=\int_X\sigma\mathrm{d}\mu(\boldsymbol{x}),\ \sigma(\boldsymbol{x})\in C(X) \tag{2}$$

线性泛函 ϕ 的范数表示为 $\|\phi\|=|\mu|(X)=\|\mu\|$, 其中, $|\mu|(\boldsymbol{x})=\sup\sum_{i=1}^{s}|\mu(A_i)|$, A_i 是 X 的一个划分, $i=1,\ldots,s$.

利用定理 1 可定义巴拿赫空间的超平面, 超平面由定义 2 给出.

定义 2. 巴拿赫空间的超平面具有以下形式:

$$M=\left\{\sigma\in C(\boldsymbol{x}),\int_X\sigma(\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x})+b=0,b\in R\right\} \tag{3}$$

利用定义 1 将每个集值对象 A_i 转化成连续函数, 表示为 $\sigma_i(\boldsymbol{x})=\sigma_{A_i}(\boldsymbol{x})$. 在这种情况下, 根据集值对象和支持函数的定义可构建一个基于函数表示的训练集, 即为 $\{(\sigma_i(\boldsymbol{x}),y_i),i=1,\ldots,n\}$. 遵循着最大间隔的思想, 下面优化模型[28]被用来处理集值数据.

$$\begin{cases}\min\limits_{\mu,b,\xi_i} & \|\mu\|+C\sum\limits_{i=1}^{n}\xi_i\\ \text{s.t.} & \begin{cases}y_i\left(\int_X\sigma_i(\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x})+b\right)\geqslant 1-\xi_i,i=1,\ \ldots,n\\ \xi_i\geqslant 0,\ i=1,\ldots,n\end{cases}\end{cases} \tag{4}$$

其中, $\|\mu\|$ 是博雷尔测度 $\mu(\boldsymbol{x})$ 的总变分. 公式 (4) 是巴拿赫空间的凸优化问题. 因为公式 (4) 涉及巴拿赫空间的积分, 所以直接求解公式 (4) 是不可行的. 为了求解优化问题公式 (4), Chen 等人[28]将整个空间 R^m 划分为 s 个不同区域, 然后计算每个区域的拉东 (Radon) 测度.

2 孪生支持函数机

本节首先在巴拿赫空间为集值数据引入超平面学习模型, 它包含两个优化问题. 随后讨论了模型的性质并构建了基于采样策略的优化算法. 最后将支持函数推广到基于核函数的方法.

2.1 提出的模型

利用支持函数的定义把集值对象 A_i $(i=1,\ldots,n)$ 转化成连续函数 $\sigma_i(\boldsymbol{x})=\sigma_{A_i}(\boldsymbol{x})(i=1,\ldots,n)$, 这形成了连续函数空间. 在这种情况下, 需要对连续函数进行分类. 通常, 设计基于函数的分类算法比设计基于向量的分类算法更具挑战性, 这是因为前者涉及无穷维的函数空间, 并且类与类之间由于连续函数的存在使得边界更复杂. 为了便于描述, 假定前 l 个集值对象来自正类, 其余来自负类. 正像图 1 所示的那样, 有时需要两个超平面拟合两类的连续函数. 这样我们期望在巴拿赫空间寻找两个非平行的超平面, 这两个超平面表示为 $\int_X\sigma(\boldsymbol{x})\mathrm{d}\mu_1(\boldsymbol{x})+b_1=0$ 和 $\int_X\sigma(\boldsymbol{x})\mathrm{d}\mu_2(\boldsymbol{x})+b_2=0$. 为此构建了下面优化模型来取得巴拿赫空间的两个非平行的超平面:

$$\begin{cases}\min\limits_{\mu_1,b_1,\xi_i}\dfrac{1}{2}\sum_{i=1}^{l}w_i\left(\int_X\sigma_i(\boldsymbol{x})\,\mathrm{d}\mu_1(\boldsymbol{x})+b_1\right)^2+c_1\sum_{i=l+1}^{n}w_i\xi_i+c_2\|\mu_1\|\\ \text{s.t. } 1-\xi_i\leqslant y_i\left(\int_X\sigma_i(\boldsymbol{x})\mathrm{d}\mu_1(\boldsymbol{x})+b_1\right)\leqslant 1+\dfrac{\xi_i}{\tau},\ i=l+1,\ldots,n\end{cases} \tag{5}$$

$$\begin{cases}\min\limits_{\mu_2,b_2,\xi_i}\dfrac{1}{2}\sum_{i=l+1}^{n}w_i\left(\int_X\sigma_i(\boldsymbol{x})\,\mathrm{d}\mu_2(\boldsymbol{x})+b_2\right)^2+c_3\sum_{i=1}^{l}w_i\xi_i+c_4\|\mu_2\|\\ \text{s.t. } 1-\xi_i\leqslant y_i\left(\int_X\sigma_i(\boldsymbol{x})\mathrm{d}\mu_2(\boldsymbol{x})+b_2\right)\leqslant 1+\dfrac{\xi_i}{\tau},\ i=1,\ldots,l\end{cases} \tag{6}$$

其中, c_1,c_2,c_3,c_4 为非负参数, τ 在区间 $(0,1]$ 内取值, w_i 表示第 i 个集值对象的权重. 公式 (5) 和公式 (6) 采用了弹球损失函数[13,16]. 当 τ 趋向 0 时, 弹球损失函数逼近铰链损失函数. 当 $\tau=0$ 时, 公式 (5) 和公式 (6) 的第 2 个约束变成了 $\xi_i\geqslant 0$. 公式 (5) 使得正类的连续函数 $\sigma_i(\boldsymbol{x})(i=1,\ldots,l)$ 逼近超平面 $\int_X\sigma(\boldsymbol{x})\mathrm{d}\mu_1(\boldsymbol{x})+b_1=0$ 并且使得负类的连续函数 $\sigma_i(\boldsymbol{x})(i=l+1,\ldots,n)$ 远离这个超平面. 对于公式 (6), 负类的连续函数逼近超平面 $\int_X\sigma(\boldsymbol{x})\mathrm{d}\mu_2(\boldsymbol{x})+b_2=0$, 而正类的连续函数远离该超平面. 在公式 (5) 和公式 (6) 中, μ_1 和 μ_2 通常取拉东测度, 拉东测度是一类博雷尔测度, 但它没有明确的表达式. 我们引入了测度 μ_1 和 μ_2 的总变分来控制模型的复杂度. 由于公式 (5) 和公式 (6) 的优化变

量包含测度, 所以它们是无穷维空间的优化问题.

2.2 模型的性质和优化

尽管公式 (5) 和公式 (6) 是无穷维空间的优化问题, 但下面的定理表示了它们是凸优化问题.

定理 2. 公式 (5) 和公式 (6) 是凸优化问题.

证明: 公式 (5) 的目标函数的第 1 项是关于 μ_1 和 b_1 的二次函数的和, 第 2 项是松弛变量 ξ_i $(i=l+1,\ldots,n)$ 的和, 第 3 项是测度的总变分, 所以公式 (5) 的目标函数是凸的. 公式 (5) 的约束是关于 μ_1, b_1 和 ξ 的线性不等式. 因此公式 (5) 是一个凸优化问题. 同样地可知公式 (6) 是一个凸优化问题.

与 SFM 不同, 公式 (5) 的目标函数采用了弹球损失函数并考虑了集值对象的权重. 如果没有测度 μ_1 和 μ_2 的相关信息, 那么无法求解优化问题公式 (5) 和公式 (6). 例如, 是否这两个测度关于其他测度是绝对连续的? 因此需要探索公式 (5) 和公式 (6) 的易求解的优化模型, 即把它们转化成有限维空间的优化问题. 根据泛函理论的相关知识, 狄拉克测度在 $C(X)$ 的对偶空间是弱闭的[32]. 这提示我们探索狄拉克测度的线性组合. 在这种情况下, 假设测度 μ_1 和 μ_2 有下面表示形式:

$$\mu_1=\left\{\sum_{k=1}^{s}\alpha_k\delta_{\boldsymbol{x}_k},\boldsymbol{x}_k\in X,\alpha_k\in R\right\} \tag{7}$$

$$\mu_2=\left\{\sum_{k=1}^{s}\beta_k\delta_{\bar{\boldsymbol{x}}_k},\bar{\boldsymbol{x}}_k\in X,\beta_k\in R\right\} \tag{8}$$

其中, $\delta_{\boldsymbol{x}_k}$ 和 $\delta_{\bar{\boldsymbol{x}}_k}$ 分别表示点 $\boldsymbol{x}_k$ 和 $\bar{\boldsymbol{x}}_k$ 处的狄拉克测度. μ_1 和 μ_2 是由狄拉克测度构建的两个测度, 其支集包含 s 个数据点. 根据公式 (7) 和公式 (8) 中测度的表示形式, 利用下面定理取得博雷尔测度 μ_1 和 μ_2 的总变分.

定理 3. 令测度 μ_1 和 μ_2 如公式 (7) 和公式 (8) 所示, 那么下面等式成立:

$$\|\mu_1\|=\left\|\sum_{k=1}^{s}\alpha_k\delta_{\boldsymbol{x}_k}\right\|=\sum_{k=1}^{s}|\alpha_k| \tag{9}$$

$$\|\mu_2\|=\left\|\sum_{k=1}^{t}\beta_k\delta_{\bar{\boldsymbol{x}}_k}\right\|=\sum_{k=1}^{s}|\beta_k| \tag{10}$$

证明: 根据测度总变分的概念可取得 $\|\mu_1\|=\left\|\sum_{k=1}^{s}\alpha_k\delta_{\boldsymbol{x}_k}\right\|=\sup\dfrac{\left|\sum_{k=1}^{s}\alpha_k f(\boldsymbol{x}_k)\right|}{\|f\|_\infty}=\sum_{k=1}^{s}|\alpha_k|$, 其中, $\|f\|_\infty$ 表示连续函数 $f(x)$ 的最大范数. 相似地可取得 μ_2 的总变分.

利用 公式 (7)–公式 (10), 公式 (5) 和公式 (6) 被改写成下面形式:

$$\begin{cases}\min\limits_{\alpha_k,\boldsymbol{x}_k,b_1,\xi_i}\dfrac{1}{2}\sum\limits_{i=1}^{l}w_i\left(\sum\limits_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)\alpha_k+b_1\right)^2+c_1\sum\limits_{i=l+1}^{n}w_i\xi_i+c_2\sum\limits_{k=1}^{s}|\alpha_k|\\ \text{s.t. }1-\xi_i\leqslant y_i\left(\sum\limits_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)\alpha_k+b_1\right)\leqslant 1+\dfrac{\xi_i}{\tau},\ i=l+1,\ldots,n\end{cases} \tag{11}$$

$$\begin{cases}\min\limits_{\beta_k,\bar{\boldsymbol{x}}_k,b_2,\xi_i}\dfrac{1}{2}\sum\limits_{i=l+1}^{n}w_i\left(\sum\limits_{k=1}^{s}\sigma_i(\bar{\boldsymbol{x}}_k)\beta_k+b_2\right)^2+c_3\sum\limits_{i=1}^{l}w_i\xi_i+c_4\sum\limits_{k=1}^{s}|\beta_k|\\ \text{s.t. }1-\xi_i\leqslant y_i\left(\sum\limits_{k=1}^{s}\sigma_i(\bar{\boldsymbol{x}}_k)\beta_k+b_2\right)\leqslant 1+\dfrac{\xi_i}{\tau},\ i=1,\ldots,l\end{cases} \tag{12}$$

这样, 无穷维空间的优化问题公式 (5) 和公式 (6) 被转化成有限维空间的优化问题公式 (11) 和公式 (12). 但优化问题公式 (11) 和公式 (12) 关于它们的某些优化变量是非线性的. 这是因为通过狄拉克测度引入了一些数据点并且 $\sigma_i(\boldsymbol{x}_k)$ 是 $\boldsymbol{x}_k$ 的非线性函数. 因为 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_s$ 和 $\bar{\boldsymbol{x}}_1,\ldots,\bar{\boldsymbol{x}}_s$ 是优化变量, 所以公式 (11) 和公式 (12) 是非凸优化问题. 如果 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_s$ 是已知的, 那么公式 (11) 仅包含优化变量 (α_k,b_1,ξ_i). 为了有效地求解公式 (11) 和公式 (12), 需要探索 X 中的适当离散点. 不同于求解 SFM 的算法, 本文利用采样策略取得 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_s$ 和 $\bar{\boldsymbol{x}}_1,\ldots,\bar{\boldsymbol{x}}_s$. 常用的采样方式是

从集值对象 A_i $(i=1,\dots,n)$ 中采样 s 个数据点. 在后面的实验部分, 通过实验验证了这种采样策略可取得令人满意的结果. 因为 α 和 β 采用了 L1 范数的约束, 所以 α 和 β 中的元素是稀疏的. 当 s 个采样点给定时, 本文采用下面的优化问题取得巴拿赫空间的超平面:

$$\begin{cases} \min\limits_{\alpha_k,b_1,\xi_i} \dfrac{1}{2}\displaystyle\sum_{i=1}^{l} w_i\left(\sum_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)\alpha_k+b_1\right)^2+c_1\sum_{i=l+1}^{n}w_i\xi_i+c_2\sum_{k=1}^{s}|\alpha_k| \\ \text{s.t. } 1-\xi_i\leqslant y_i\left(\displaystyle\sum_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)\alpha_k+b_1\right)\leqslant 1+\dfrac{\xi_i}{\tau},\ i=l+1,\dots,n \end{cases} \tag{13}$$

$$\begin{cases} \min\limits_{\beta_k,b_2,\xi_i} \dfrac{1}{2}\displaystyle\sum_{i=l+1}^{n} w_i\left(\sum_{k=1}^{s}\sigma_i(\bar{\boldsymbol{x}}_k)\beta_k+b_2\right)^2+c_3\sum_{i=1}^{l}w_i\xi_i+c_4\sum_{k=1}^{s}|\beta_k| \\ \text{s.t. } 1-\xi_i\leqslant y_i\left(\displaystyle\sum_{k=1}^{s}\sigma_i(\bar{\boldsymbol{x}}_k)\beta_k+b_2\right)\leqslant 1+\dfrac{\xi_i}{\tau},\ i=1,\dots,l \end{cases} \tag{14}$$

从公式 (13) 和公式 (14) 可知它们是凸优化模型. 如果 $\alpha_k\neq 0$, 将相应的采样点 $\boldsymbol{x}_k$ 称为公式 (13) 的支持向量. 如果 $\beta_k\neq 0$, 将相应的采样点 $\bar{\boldsymbol{x}}_k$ 称为公式 (14) 的支持向量. 与 TSVM 不同, TSFM 的支持向量取决于采样点. 由于公式 (13) 和公式 (14) 有相似的形式, 下面主要讨论公式 (13) 的优化问题. 为了处理公式 (13) 中的绝对值符号, 这里引入了非负优化变量 $\bar{\alpha}_k$ 和 $\hat{\alpha}_k$. 设 $\bar{\alpha}_k+\hat{\alpha}_k=|\alpha_k|$ 和 $\bar{\alpha}_k-\hat{\alpha}_k=\alpha_k$, 其中, $\bar{\alpha}_k\geqslant 0,\hat{\alpha}_k\geqslant 0$, $k=1,\dots,s$. 这样公式 (13) 被转化成下面形式:

$$\begin{cases} \min\limits_{\bar{\alpha}_k,\hat{\alpha}_k,b_1,\xi_i} \dfrac{1}{2}\displaystyle\sum_{i=1}^{l} w_i\left(\sum_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)(\bar{\alpha}_k-\hat{\alpha}_k)+b_1\right)^2+c_1\sum_{i=l+1}^{n}w_i\xi_i+c_2\sum_{k=1}^{s}(\bar{\alpha}_k+\hat{\alpha}_k) \\ \text{s.t. } 1-\xi_i\leqslant y_i\left(\displaystyle\sum_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)(\bar{\alpha}_k-\hat{\alpha}_k)+b_1\right)\leqslant 1+\dfrac{\xi_i}{\tau},\ i=l+1,\dots,n \end{cases} \tag{15}$$

公式 (15) 是一个二次规划问题. 利用二次规划的优化算法可求解凸优化问题公式 (15). 对于凸优化问题, 可通过探索对偶解来验证原问题解的最优性, 这需要推导出 公式 (15) 的对偶形式. 首先公式 (15) 被转化成下面形式:

$$\begin{cases} \min\limits_{\bar{\alpha}_k,\hat{\alpha}_k,b_1,\xi_i,u_i} \dfrac{1}{2}\displaystyle\sum_{i=1}^{l} w_i(u_i)^2+c_1\sum_{i=l+1}^{n}w_i\xi_i+c_2\sum_{k=1}^{s}(\bar{\alpha}_k+\hat{\alpha}_k) \\ \text{s.t.} \begin{cases} u_i=\displaystyle\sum_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)(\bar{\alpha}_k-\hat{\alpha}_k)+b_1 \\ 1-\xi_i\leqslant y_i\left(\displaystyle\sum_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)(\bar{\alpha}_k-\hat{\alpha}_k)+b_1\right)\leqslant 1+\dfrac{\xi_i}{\tau},\ i=l+1,\dots,n \\ \bar{\alpha}_k\geqslant 0,\hat{\alpha}_k\geqslant 0 \end{cases} \end{cases} \tag{16}$$

根据公式 (16) 定义下面的拉格朗日函数:

$$\begin{aligned} L(\bar{\alpha}_k,\hat{\alpha}_k,b_1,\xi_i,u_i)=&\frac{1}{2}\sum_{i=1}^{l}w_i(u_i)^2+c_1\sum_{i=l+1}^{n}w_i\xi_i+c_2\sum_{k=1}^{s}(\bar{\alpha}_k+\hat{\alpha}_k)+\sum_{i=l+1}^{n}\bar{\gamma}_i\left(1-\xi_i-y_i\left(\sum_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)(\bar{\alpha}_k-\hat{\alpha}_k)+b_1\right)\right) \\ &+\sum_{i=1}^{l}\gamma_i\left(u_i-\sum_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)(\bar{\alpha}_k-\hat{\alpha}_k)+b_1\right)+\sum_{i=l+1}^{n}\hat{\gamma}_i\left(y_i\left(\sum_{k=1}^{s}\sigma_i(\boldsymbol{x}_k)(\bar{\alpha}_k-\hat{\alpha}_k)+b_1\right)-1-\frac{\xi_i}{\tau}\right)-\sum_{k=1}^{s}\hat{\theta}_k\hat{\alpha}_k-\sum_{k=1}^{s}\bar{\theta}_k\bar{\alpha}_k \end{aligned} \tag{17}$$

其中, γ_i $(i=1,\dots,l)$, $\bar{\gamma}_i,\hat{\gamma}_i$ $(i=l+1,\dots,n)$, $\bar{\theta},\hat{\theta}_k$ $(k=1,\dots,s)$ 是拉格朗日乘子. 将公式 (17) 的拉格朗日函数关于 $\bar{\alpha}_k,\hat{\alpha}_k$, ξ_i,u_i 和 b_1 的导数设定为 0, 那么可取得:

$$\frac{\partial L}{\partial u_i}=w_iu_i+\gamma_i=0 \tag{18}$$

$$\frac{\partial L}{\partial \bar{\alpha}_k} = c_2 - \sum_{i=l+1}^{n} (\bar{\gamma}_i - \hat{\gamma}_i) y_i \sigma_i(\boldsymbol{x}_k) - \sum_{i=1}^{l} \sigma_i(\boldsymbol{x}_k) \gamma_i - \bar{\theta}_k = 0 \tag{19}$$

$$\frac{\partial L}{\partial \hat{\alpha}_k} = c_2 + \sum_{i=l+1}^{n} (\bar{\gamma}_i - \hat{\gamma}_i) y_i \sigma_i(\boldsymbol{x}_k) + \sum_{i=1}^{l} \sigma_i(\boldsymbol{x}_k) \gamma_i - \hat{\theta}_k = 0 \tag{20}$$

$$\frac{\partial L}{\partial \xi_i} = w_i c_1 - \bar{\gamma}_i - \frac{\hat{\gamma}_i}{\tau} = 0,\ i = l+1, \ldots, n \tag{21}$$

$$\frac{\partial L}{\partial b_1} = -\sum_{i=1}^{l} \gamma_i - \sum_{i=l+1}^{n} (\bar{\gamma}_i - \hat{\gamma}_i) y_i = 0 \tag{22}$$

从公式 (17)–公式 (22) 可取得公式 (15) 的对偶优化问题, 表示为:

$$\begin{cases} \max\limits_{\gamma_i, \bar{\gamma}_i, \hat{\gamma}_i} -\dfrac{1}{2} \sum\limits_{i=1}^{l} \dfrac{(\gamma_i)^2}{w_i} + \sum\limits_{i=l+1}^{n} (\bar{\gamma}_i - \hat{\gamma}_i) y_i \\ \text{s.t.} \begin{cases} c_2 - \sum\limits_{i=l+1}^{n} y_i (\bar{\gamma}_i - \hat{\gamma}_i) \sigma_i(\boldsymbol{x}_k) - \sum\limits_{i=1}^{l} \gamma_i \sigma_i(\boldsymbol{x}_k) \geqslant 0,\ k = 1, \ldots, s \\ c_2 + \sum\limits_{i=l+1}^{n} y_i (\bar{\gamma}_i - \hat{\gamma}_i) \sigma_i(\boldsymbol{x}_k) + \sum\limits_{i=1}^{l} \gamma_i \sigma_i(\boldsymbol{x}_k) \geqslant 0,\ k = 1, \ldots, s \\ \sum\limits_{i=1}^{l} \gamma_i + \sum\limits_{i=l+1}^{n} (\bar{\gamma}_i - \hat{\gamma}_i) y_i = 0 \\ \tau w_i c_1 - \tau \bar{\gamma}_i - \hat{\gamma}_i = 0,\ i = l+1, \ldots, n \end{cases} \end{cases} \tag{23}$$

类似地可取得公式 (14) 的对偶优化问题, 表示为:

$$\begin{cases} \max\limits_{\eta_i, \bar{\eta}_i, \hat{\eta}_i} -\dfrac{1}{2} \sum\limits_{i=l+1}^{n} \dfrac{(\eta_i)^2}{w_i} + \sum\limits_{i=1}^{l} (\bar{\eta}_i - \hat{\eta}_i) y_i \\ \text{s.t.} \begin{cases} c_4 - \sum\limits_{i=1}^{l} y_i (\bar{\eta}_i - \hat{\eta}_i) \sigma_i(\boldsymbol{x}_k) - \sum\limits_{i=l+1}^{n} \eta_i \sigma_i(\boldsymbol{x}_k) \geqslant 0,\ k = 1, \ldots, s \\ c_4 + \sum\limits_{i=1}^{l} y_i (\bar{\eta}_i - \hat{\eta}_i) \sigma_i(\boldsymbol{x}_k) + \sum\limits_{i=l+1}^{n} \eta_i \sigma_i(\boldsymbol{x}_k) \geqslant 0,\ k = 1, \ldots, s \\ \sum\limits_{i=l+1}^{n} \eta_i + \sum\limits_{i=1}^{l} (\bar{\eta}_i - \hat{\eta}_i) y_i = 0 \\ \tau w_i c_3 - \tau \bar{\eta}_i - \hat{\eta}_i = 0,\ i = 1, \ldots, l \end{cases} \end{cases} \tag{24}$$

公式 (23) 和公式 (24) 是凸优化问题. 它们的约束个数随着采样点的增加而增加. 当采样点比较少时, 可借助公式 (23) 和公式 (24) 取得对偶解. 从公式 (13) 和公式 (14) 可知支持向量取决于采样点, 但非支持向量不影响它们的目标函数. 如果采样点不是支持向量, 则无需考虑此采样点. 当存在大量采样点时, 删除非支持向量的采样点能降低优化模型的复杂度. 因此关键的问题是如何判断哪些采样点是支持向量. 下面的定理给出了答案.

定理 4. 假定求解公式 (23) 取得最优解 $\gamma_i\ (i = 1, \ldots, l)$, $\bar{\gamma}_i, \hat{\gamma}_i\ (i = l+1, \ldots, n)$. 如果采样点 x_k 满足条件:

$$\left| \sum_{i=l+1}^{n} y_i (\bar{\gamma}_i - \hat{\gamma}_i) \sigma_i(\boldsymbol{x}_k) + \sum_{i=1}^{l} \gamma_i \sigma_i(\boldsymbol{x}_k) \right| < c_2 \tag{25}$$

那么 $\boldsymbol{x}_k$ 不是一个支持向量.

证明: 从拉格朗日函数 (17) 的定义可知 $\bar{\theta}$ 和 $\hat{\theta}_k$ 非负的. 从 Karush-Kuhn-Tucker 最优条件可知: 如果 $c_2 + \sum\limits_{i=l+l}^{n} y_i (\bar{\gamma}_i - \hat{\gamma}_i) \sigma_i(\boldsymbol{x}_k) - \sum\limits_{i=1}^{l} \gamma_i \sigma_i(\boldsymbol{x}_k) > 0$ 和 $c_2 + \sum\limits_{i=l+l}^{n} y_i (\bar{\gamma}_i - \hat{\gamma}_i) \sigma_i(\boldsymbol{x}_k) + \sum\limits_{i=1}^{l} \gamma_i \sigma_i(\boldsymbol{x}_k) > 0$, 那么从公式 (19) 和公式 (20) 知 $\bar{\theta} \neq 0$ 和 $\hat{\theta}_k \neq 0$. 利用互补条件得到 $\bar{\alpha}_k = 0$ 和 $\hat{\alpha}_k = 0$, 从而取得 $\alpha_k = 0$, 这表明 $\boldsymbol{x}_k$ 不是支持向量.

定理 5. 假定求解公式 (24) 取得最优解 $\eta_i\ (i = l+1, \ldots, n)$, $\bar{\eta}_i, \hat{\eta}_i\ (i = 1, \ldots, l)$ 如果采样点 $\bar{\boldsymbol{x}}_k$ 满足:

$$\left|\sum_{i=1}^{l} y_i(\bar{\eta}_i-\hat{\eta}_i)\sigma_i(\bar{\boldsymbol{x}}_k)+\sum_{t=1+1}^{n}\eta_i\sigma_i(\bar{\boldsymbol{x}}_k)\right|<c_4 \tag{26}$$

那么 $\bar{\boldsymbol{x}}_k$ 不是一个支持向量.

定理 5 的证明相似于定理 4 的证明. 定理 4 和定理 5 表示了哪些采样点不影响公式 (13) 和公式 (14) 的解, 这说明借助对偶解可确定原问题的最优解的哪些分量是 0. 公式 (23) 和公式 (24) 的目标函数不依赖于采样点, 这为设计有效的策略来处理采样点提供了一些提示. 例如, 当存在大量采样点时, 可首先选择一小部分采样点, 并假定选取 s_1 (<< s) 个采样点, 利用这 s_1 个采样点, 求解公式 (23) 和公式 (24), 然后利用定理 4 和定理 5 删除那些非支持向量的采样点. 在这种情况下, 从采样点获得一批支持向量, 并利用这些选定的支持向量训练公式 (13) 和公式 (14). 利用这种策略得到 (α,b_1) 和 (β,b_2) 并取得两个非平行的超平面. 不同于 SFM, 为了分类集值数据, 需要定义集值对象 A 到超平面的距离.

定义 3. 从集值对象 A 到巴拿赫空间的超平面 $\int_X \sigma_A(\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x})+b=0$ 的距离表示为:

$$D(A,(\mu,b))=\frac{\left|\int_X \sigma_A(\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x})+b\right|}{\|\mu\|} \tag{27}$$

从公式 (27) 可知计算距离 $D(A,(\mu,b))$ 需要积分运算, 直接求解积分取得距离是不可行的. 在实际实施时需要借助采样点来离散化距离并取得近似距离. 基于定义 3, 本文采用下面判决规则来取得集值对象 A 的标签:

$$\operatorname{argmin}\left\{\frac{\left|\int_X \sigma_A(\boldsymbol{x})\mathrm{d}\mu_1(\boldsymbol{x})+b_1\right|}{\mu_1},\frac{\left|\int_X \sigma_A(\boldsymbol{x})\mathrm{d}\mu_2(\boldsymbol{x})+b_2\right|}{\mu_2}\right\} \tag{28}$$

从 公式 (28) 可知判决规则中的距离函数需要利用采样点进行离散化. 从公式 (13) 和公式 (14) 可知 α 和 β 采用了 L1 范数的约束, 因此采样点是稀疏的. 为了便于理解 TSFM, 算法 1 列出了使用 TSFM 对集值数据进行分类的伪代码. 从算法 1 可知, 由于 公式 (11) 和 (12) 对应不同的优化问题, 算法 1 的步骤 3 和 4 可选择不同的采样点. 但在实际实施过程中通常采用相同的采样点.

算法 1. TSFM 的伪代码.

1. 设定参数 c_1,c_2,c_3,c_4;
2. 利用支持函数取得连续函数 $\sigma_i(\boldsymbol{x})=\sigma_{A_i}(\boldsymbol{x})(i=1,\ldots,n)$;
3. 对于公式 (11), 从集值对象 A_i $(i=1,\ldots,n)$ 采样 s 个数据点, 表示为 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_s$;
4. 对于公式 (12), 从集值对象 A_i $(i=1,\ldots,n)$ 采样 s 个数据点, 表示为 $\bar{\boldsymbol{x}}_1,\ldots,\bar{\boldsymbol{x}}_s$;
5. 通过求解公式 (13) 和公式 (14) 取得 (α,b_1) 和 (β,b_2);
6. 利用公式 (28) 分类集值对象 A.

2.3 核函数空间的支持函数

为了探索数据的非线性特征, 基于核函数的学习方法[33]利用非线性映射把原特征映射到高维空间. 令 ϕ 表示非线性映射, 由内积定义的核函数表示为 $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=<\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)>$. 如果给定核函数, 则核方法不需要非线性映射的显式表示. 高斯核和多项式核是广泛使用的核函数, 它们通常是正半定的. 使用正定核函数是确保支持向量机在对偶空间的目标函数是凸函数. 然而, 对于孪生支持函数机, 模型的优化变量为测度, 测度依赖于采样点的选择, 采样点影响着支持函数的取值. 这样需要考虑采样点 $\boldsymbol{x}$ 和集值对象 A 的非线性变换 $\phi(\boldsymbol{x})$ 和 $\phi(A)$. 利用这些变换可定义支持函数的核化. 下面定义给出了支持函数的核化.

定义 4. 令 $\kappa(\cdot,\cdot)$ 表示核函数以及 ϕ 表示对应的非线性映射. 支持函数的核化被表示为:

$$\sigma_{\phi(A)}(\phi(\boldsymbol{x}))=\sup\{\langle\phi(\boldsymbol{x}),\phi(\boldsymbol{\omega})\rangle,\boldsymbol{\omega}\in A\}=\sup\{\kappa(\boldsymbol{x},\boldsymbol{\omega}),\boldsymbol{\omega}\in A\} \tag{29}$$

利用定义 4 可取得支持函数的核化. 当确定的 ϕ 给定时, 公式 (29) 也可用于显式的内积运算. 核化过程实际上

对支持函数进行了预处理, 这样核化不影响公式 (13) 和公式 (14) 的凸性. 这也表明即使支持函数采用不定核函数进行核化, 公式 (13) 和公式 (14) 也是凸优化模型. 因此 TSFM 为不同类型的核函数和显式的内积运算提供了更加灵活的选择.

3 实验结果

本节在模拟的集值数据和一些真实世界的数据集上执行了一系列的实验来验证 TSFM 的有效性. 与 TSVM 一样, TSFM 包含多个超参数, 这些超参数会影响模型的性能. 为了减少模型的超参数, 令 $c_1=c_3$ 和 $c_2=c_4$, 这些设定相似于孪生支持向量机的超参数设定. 从集合 $\{10^i, i=-3,-2,\ldots,2,3\}$ 中选取最优超参数 c_1 和 c_2. τ 从 0 和 1 区间以间隔 0.1 选择最优超参数. 对于模型的权重 w_i, 它度量了每个集值对象的重要性. 为了容易取得权重, 本文首先取得每个集值对象的均值, 这样许多已有的权重方法[6,16,17]可被用来取得集值对象的权重, 从而利用均值取得的权重作为集值对象的权重. 本文采用模糊权重的方法[16]取得每个集值对象的权重. 对于支持函数的核化过程, 我们测试了高斯核 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2/\sigma\right)$, 其中, σ 从集合 $\{10^i, i=-3,-2,\ldots,2,3\}$ 中取值. 在实际实施时, 为了简单性, 我们没有构建集值对象中事例的凸包, 而是使用集值对象中的事例计算出连续函数的值. 对于采样点, 在下面的具体实验中会提到相应的采样方法. 所有实验均在配备 i7 处理器和 16 GB RAM 的计算机上完成, 采用 Matlab R2020b 编程语言实现了相关算法.

3.1 模拟的交叉类型的集值数据

本节利用交叉类型的集值数据来表明 TSFM 能获取集值数据的内在结构并通过调整参数抑制数据中的离群点. 对于实验, 首先为二元问题的每个类生成 100 个数据点. 正类的样本采用 $(z,z)^{\mathrm{T}}$ 的形式, 其中 z 取自区间 $[-5,5]$ 的均匀分布. 负类的样本采用 $(z,-z)^{\mathrm{T}}$ 的形式. 为了生成集值数据, 在每个数据点的基础上, 产生对应的高斯分布, 其均值来自每个数据点, 协方差矩阵设定为 $(0.1,0;0,0.1)$. 根据高斯分布, 为每个集值对象生成 5 个事例, 这实际上每个集值对象包含 5 个数据点. 这样从原始数据点构造了集值数据. 图 2 表示了 TSFM 采用线性核并在不同超参数 τ 下的实验结果, 其中, $c_1=c_3=1$ 和 $c_2=c_4=100$. 从集值数据中采样数据点, 利用 TSFM 取得了如图 2(a) 所示的交叉线. 从图 2 可看出, 在没有离群点的情况下, 超参数 τ 对超平面的影响不是很大, 在不同的 τ 下 TSFM 能获取集值数据的内在结构. 为了验证 TSFM 能否抑制离群点, 在原数据集的基础上, 人为增加了一些离群点. 图 3 表示了包含离群点的集值数据. 从图 3 可知, 一些数据点位于另一类对应的直线附近. 从图 3 可知, 当数据包含离群点时, τ 的变化明显影响了 TSFM 的性能. 当 $\tau=0$ 和 $\tau=0.8$ 时, 利用 TSFM 取得的交叉线偏离实际交叉线. 当 $\tau=0.5$ 时, 利用 TSFM 取得的交叉线和实际交叉线相差不多, 这说明在离群点的情况下, 可调整参数 τ 获取集值数据的内在结构. 实验结果表明: 弹球损失函数的孪生支持函数机通过调整超参数 τ 可捕捉集值数据的内在结构.

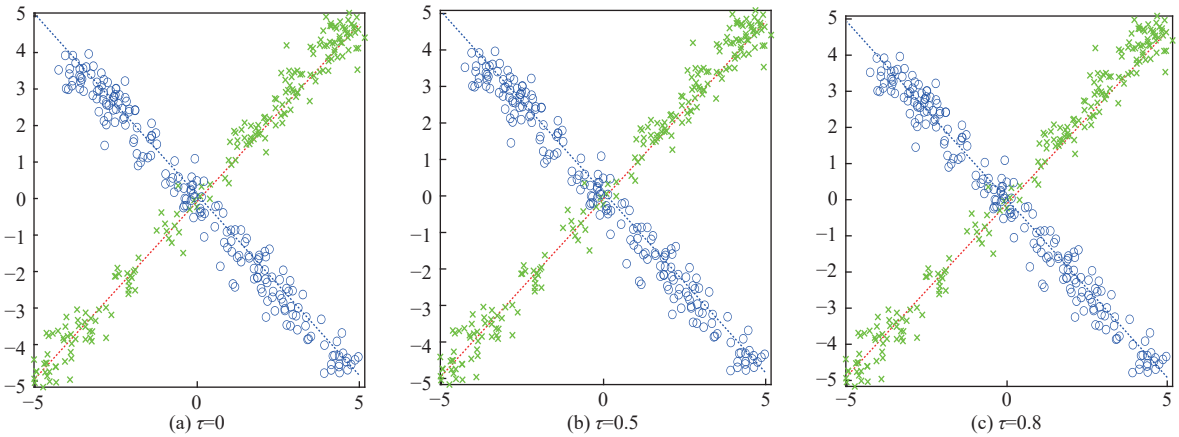


图 2 在交叉类型的集值数据上由 TSFM 取得的超平面

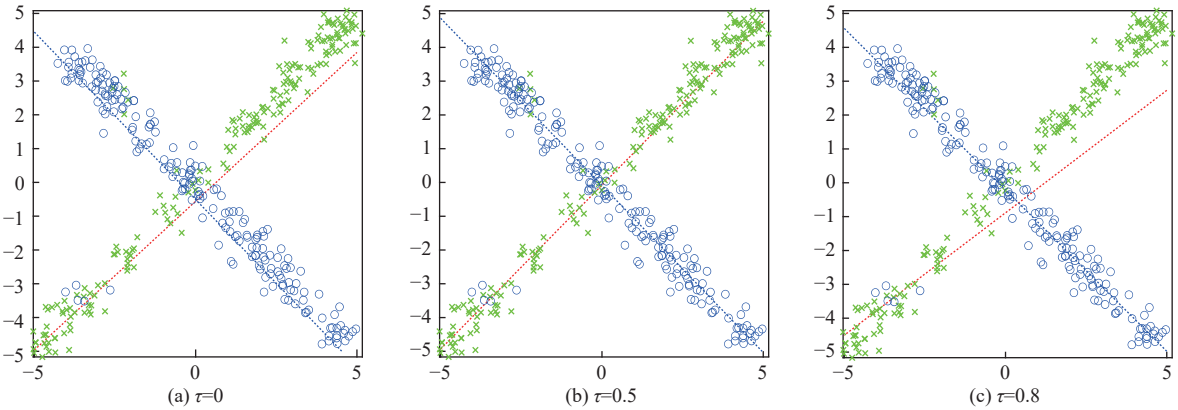


图 3　在包含离群点的交叉类型的集值数据上由 TSFM 取得的超平面

3.2　UCI 数据集上的实验

本节利用非高斯分布的集值对象来测试 TSFM 的分类性能. 从 UCI 数据仓库中选取一些数据集进行了实验. 所选用的数据集如表 2 所示, 这些数据集通常被用来评估分类器的性能. 对于这些数据集, 样本提供了向量的描述方式而不是集合的描述方式, 这意味着原始样本没有集值表示. 遵循着集值数据的生成方法[34], 我们首先计算出所有样本的第 i 个属性的标准差, 表示为 σ_i, 然后构造一个区间 $[x_i-\sqrt{3}\sigma_i, x_i+\sqrt{3}\sigma_i]$, 其中, x_i 是样本 x 的第 i 个属性的值. 对于集值数据, 根据均匀分布从这些区间生成 10 个事例. 因此每个集值对象包含 10 个事例. 通过这种方式为每个数据集构造出集值数据. 在实验中, 以集值对象所包含事例的均值作为采样点. 如果训练集的样本数超过 500, 则采样点数设定为 500. 实验中采用一对多的策略来处理多分类问题.

表 2　数据集的统计信息

数据集	样本数	特征数	类别数
Australian	690	14	2
Breast	683	9	2
Heart	270	13	2
Ionosphere	351	34	2
PlaningRelax	182	12	2
Diabetes	768	8	2
India	583	10	2
Sonar	208	60	2
Wireless	2 000	7	4
Segmentation	2 100	19	7
Drug	1 885	12	7
Statlog	2 310	19	7
Cardio	2 126	23	10
Satellite	4 430	36	6

为了比较, 本文也实施了几种集值数据的分类方法, 如二阶锥规划 (second-order cone programming, SOCP) 方法[23], 不确定感知的孪生支持向量机 (uncertainty-aware TSVM, UTSVM)[25]支持测度机 (support measure machine, SMM)[27], 支持函数机 (SFM)[28], 稀疏近似最近点 (sparse approximation nearest point, SANP) 方法[35]以及正则化协同表示分类 (regularized collaborative representation classification, RCRC) 方法[36]. 对于 SOCP, 它包含超参数 C, γ_i $(i=1,\ldots,n)$ 以及核超参数 σ. 为了减少超参数的数目, 遵循文献 [23] 中的策略, 令 $\gamma=\gamma_i$ $(i=1,\ldots,n)$ 且 $\gamma=\sqrt{\kappa/(1-\kappa)}$, 超参数 κ 取值于集合 $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, 超参数 C 取值于集合 $\{10^i, i=-3,-2,\ldots,2,3\}$. 对于 SANP, 它包含 3 个超参数 λ_1, λ_2 和 λ_3, 本文根据文献 [35] 中的方案调整这些超参数. 对于 RCRC, 它包含超参数

λ_1, λ_2, λ_3, γ 以及核超参数 σ, 超参数设定为 $\lambda_1 = 0.001, \lambda_2 = 0.001, \lambda_3 = 2.5/n_1$, 其中 n_1 是一个集值对象包含的数据点数目. 对于 SMMs, 它包含超参数 C 和核超参数 σ, 超参数 C 取值于集合 $\{10^i, i = -3, -2, \ldots, 2, 3\}$. 对于 UTSVM, 它包含超参数 $c_1, c_2, c_3, c_4, \epsilon$ 以及核超参数 σ, 超参数设定为 $c_1 = c_3, c_2 = c_4, \epsilon = 0.5$, 超参数 c_1 和 c_2 取值于集合 $\{10^i, i = -3, -2, \ldots, 2, 3\}$. 所有方法中使用的核函数是高斯核函数且核超参数 σ 取值于集合 $\{10^i, i = -3, -2, \ldots, 2, 3\}$.

本文使用五折交叉验证来评估各种算法的性能. 为了选择各种模型的适当超参数, 对训练集进行了额外的五折交叉验证, 以取得各种方法的最优超参数. 表 3 表示了各种方法在数据集上的实验结果, 实验结果为平均错误率和标准偏差, 最佳结果以黑体显示. 为了评价 TSFM 的稳健性, 把训练集的每一类标签以一定的比例替换成其他类的标签, 替换的比例为 5%, 这模拟了标签噪声且在标签噪声下产生离群点. 表 4 表示了各种方法在包含标签噪声的数据集上的实验结果.

表 3　不同方法在 UCI 数据集上的错误率和标准偏差 (%)

数据集	SANP	RCRC	SOCP	SFM	SMM	UTSVM	TSFM
Australian	16.72±3.55	15.39±3.71	14.51±3.85	14.13±3.46	14.01±2.98	14.62±3.19	**13.6±2.32**
Breast	4.21±1.22	3.41±1.09	2.89±1.52	**2.61±1.29**	3.02±1.37	3.57±1.26	2.61±0.73
Heart	30.31±4.78	20.12±4.12	16.70±4.03	16.81±3.41	18.12±3.34	24.25±3.50	**16.2±3.46**
Ionosphere	4.38±2.03	4.12±3.04	3.79±1.28	3.69±1.32	3.31±1.64	3.95±1.09	**3.17±1.26**
PlanningRelax	30.12±1.75	29.35±1.85	28.79±1.73	**28.3±1.89**	28.62±2.11	29.54±1.58	28.56±1.21
Diabetes	27.32±3.51	25.35±3.02	24.32±2.61	23.50±2.71	**22.3±2.49**	25.68±2.29	22.46±2.35
India	29.45±4.02	28.09±3.42	27.63±3.24	26.71±3.89	26.89±4.02	27.32±3.62	**26.1±3.47**
Sonar	22.13±4.05	20.78±4.76	20.12±5.15	19.05±5.23	19.78±4.23	20.05±4.24	**18.2±4.40**
Wireless	3.26±1.21	3.15±1.21	2.92±1.24	2.73±0.89	2.52±2.26	3.03±1.31	**2.02±1.32**
Segmentation	18.22±4.21	16.72±3.92	15.27±4.16	**13.8±3.91**	15.66±3.92	16.48±3.49	13.86±3.45
Drug	20.22±3.25	18.21±3.51	16.28±3.41	15.08±3.42	15.22±2.90	18.08±3.53	**14.3±3.46**
Statlog	10.50±3.80	11.42±3.25	9.78±2.91	9.22±4.01	8.67±3.52	10.51±2.84	**8.08±2.45**
Cardio	30.27±2.26	27.80±2.45	25.34±4.05	18.23±4.53	18.30±2.81	26.12±2.29	**17.5±2.47**
Satellite	9.56±1.27	8.47±1.05	7.25±1.89	6.27±1.63	6.34±1.62	7.89±1.86	**5.24±2.47**

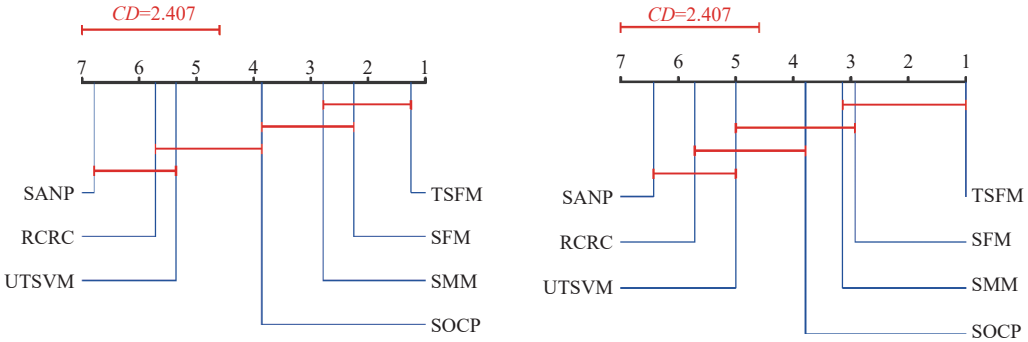
表 4　不同方法在标签噪声为 5% 的数据集上的错误率和标准偏差 (%)

数据集	SANP	RCRC	SOCP	SFM	SMM	UTSVM	TSFM
Australian	18.94±3.23	17.82±3.45	17.32±3.46	17.46±3.35	17.49±2.27	17.45±3.60	**16.77±2.47**
Breast	7.05±1.46	6.89±1.21	5.62±2.35	5.69±1.47	6.34±1.75	6.80±1.46	**4.89±1.27**
Heart	34.42±4.26	25.39±4.74	19.52±4.20	18.44±3.72	20.26±3.47	26.73±3.50	**18.21±3.63**
Ionosphere	6.52±2.42	6.59±3.56	6.89±1.65	6.74±1.59	6.83±1.99	7.06±1.82	**6.45±2.42**
PlanningRelax	32.25±1.95	31.98±1.42	30.42±1.90	30.56±1.96	30.89±2.42	34.24±1.77	**29.92±1.63**
Diabetes	28.45±3.64	27.67±3.41	26.83±2.88	26.48±2.89	24.84±2.97	26.45±2.46	**24.32±2.86**
India	31.55±4.23	30.14±3.67	29.92±3.68	28.34±3.90	28.68±4.34	29.54±4.04	**27.42±3.80**
Sonar	24.36±4.21	23.05±4.91	22.30±5.62	21.87±5.46	21.62±3.89	23.92±4.92	**20.85±4.34**
Wireless	5.42±1.33	5.74±1.46	5.08±1.41	5.13±1.24	4.93±2.42	5.23±2.45	**4.02±1.64**
Segmentation	20.56±4.46	19.26±4.55	17.83±4.42	16.52±3.72	17.39±4.41	18.56±3.97	**16.39±3.95**
Drug	23.53±3.70	20.54±3.69	19.46±3.55	18.34±3.69	18.59±3.49	20.30±3.46	**16.05±3.41**
Statlog	13.39±3.93	14.23±3.68	12.25±3.04	12.02±4.24	11.79±3.77	12.56±2.63	**10.53±2.96**
Cardio	33.46±2.56	31.67±2.34	28.56±4.24	20.45±4.46	22.65±3.08	28.94±2.56	**20.36±2.89**
Satellite	12.49±1.45	11.43±2.41	10.50±1.88	9.46±1.98	8.60±1.92	9.57±1.95	**8.45±2.15**

从表 3 可看出, SMM 在 Diabetes 数据集上取得了最佳性能, 而 SFM 在 Breast, PlanningRelax 和 Segmentation 数据集上的错误率最低. TSFM 在大多数数据集上取得了最佳的分类性能. SANP 和 RCRC 没有获得令人满意的结果, 这是因为我们探索了随机生成的集值对象. 因为集值对象并不服从高斯分布, 所以 UTSVM 并没有在这些数据集上取得好的分类性能. TSFM 在多类问题上一般优于其他模型, 这是因为 TSFM 采用了弹球损失函数和考虑了样本的权重. 基于最大间隔的方法如 SFM 和 SMM 会受到交界处数据点的影响, 而 TSFM 侧重关注接近超平面的样本

点. 从表 4 可看出, 当存在标签噪声时, 各种方法的性能存在一定程度的下降. 由于 TSFM 采用了样本的权重和弹球损失函数, TSFM 比其他方法取得更好的分类性能. 这说明 TSFM 能抑制标签噪声.

从表 3 和表 4 可看到在多个数据集上测试了多个分类器. 因此有必要根据它们在多个数据集上的性能对它们进行总体评价. 已有结果表明, 弗里德曼检验 (Friedman test) 与事后检验[37]是评价多种分类模型的有效工具. 对于弗里德曼检验, 原假设是所有分类器都有相等的秩. 在这个假设下, 可取得弗里德曼统计量, 该统计量为 $k-1$ 自由度的分布 χ_F^2, 其中, k 是分类器的个数. 弗里德曼检验考虑了分类器的平均秩. 如果两个分类器的平均秩至少存在 $CD=q_\alpha\sqrt{k(k+1)/6N}$ 值的差别[37], 则它们性能明显不同, 其中, N 是数据集的个数, q_α 是通过统计量取得的数值. 图 4(a) 表示了在 14 个数据集上 7 种方法对比的 CD 图, 图 4(b) 表示了在标签噪声的数据集上各种方法对比的 CD 图. 在置信度 $\alpha=0.05$ 的情况下, CD 的值为 2.407. 从图 4 可看出, 由于 TSFM 的平均秩最小, 因此 TSFM 比其他方法取得更好的性能.

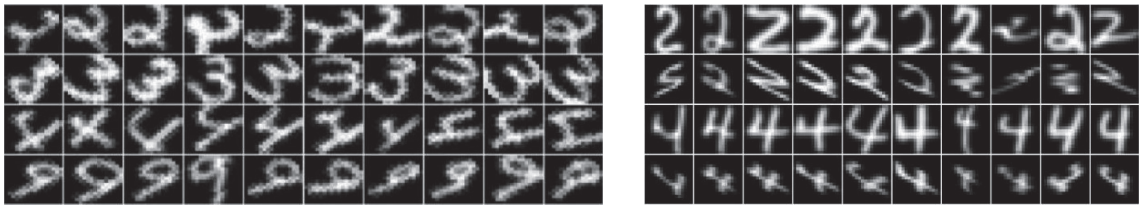


(a) 基于表 3 的结果不同方法的对比　　(b) 基于表 4 的结果不同方法的对比

图 4　依据 CD 图比较不同方法的性能

3.3　手写体数字集上的实验

本节在 USPS 数据集上测试了 TSFM 的稳健性. 为了生成集值数据, 我们考虑了两种几何变换: 旋转变换和错切变换. 对每幅 16×16 的手写体数字图像, 从高斯分布 $N(40,180)$ 取得旋转角度 θ, 并从高斯分布 $N(1,1)$ 取得错切因子 s_y. 利用这两种几何变换生成一系列图像. 这些图像相对原图像存在扭曲和变形. 为了测试二元分类任务, 我们将一些数字形成数字对, 即数字 2 和数字 3, 数字 4 和数字 9 构成数字对. 对每种情况, 从每类中随机抽取 100 幅图像, 然后利用几何变换为每幅图像生成 10, 20, 30 幅图像, 从而构建集值数据. 图 5 表示了几何变换后的一些图像. 为了生成用于测试的集值数据, 从每类图像中随机选择 100 幅图像, 并采用与训练集相似的策略生成集值数据. 训练集和测试集中的集值对象包含几何变换的事例, 这些几何变换的事例导致事例的波动范围比较大, 从而可能产生离群点. 我们取 20 次运行的平均作为实验结果, 并额外执行 5 次来选择各种模型的超参数. 实验结果如图 6 所示.



(a) 旋转的手写体数字图像　　(b) 错切的手写体数字图像

图 5　两种几何变换的手写体数字图像

从图 6 可看出, 随着集值对象包含的图像越多, 各种方法的性能在大多数情况下得到改善. 错切变换比旋转变换导致更大的错误率. 在这些图像数据上, SMM 的性能并不好于 SFM 的性能, 这是因为 SMM 利用了所有几何变

换的图像. 从图 6 可知, SANP 并不优于其他方法, 这是因为变换后的图像包含扭曲的图像. 由于 TSFM 采用了弹球损失函数和测度的总变分, 所以它能取得好的实验结果. 实验结果表明在处理具有不确定性 (扭曲和变形) 的集值数据上 TSFM 是有效的.

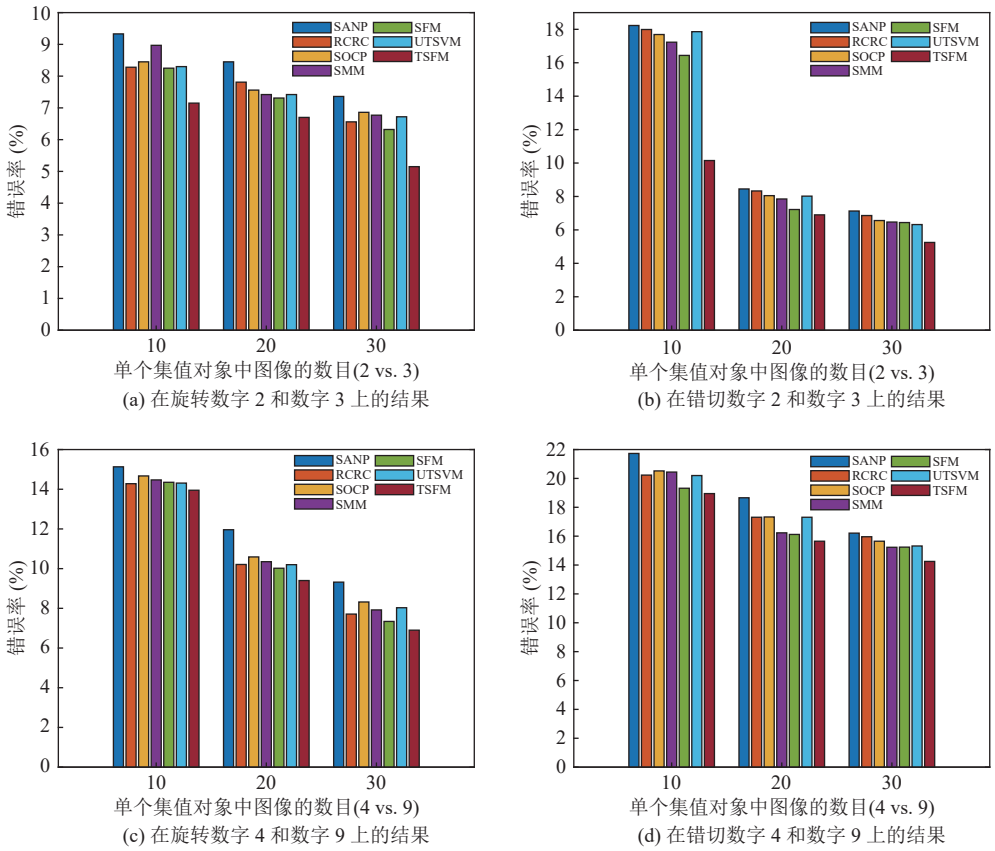


图 6 不同方法在手写体数字图像上的错误率

3.4 医学图像集上的实验

本节在 4 个医学图像集[38]上构建了集值数据并测试了 TSFM 在集值对象包含少量高维事例情况下的分类性能. 所用的数据集是 780 幅图像的乳腺图像集 (Breast), 4 708 幅图像的肺炎图像集 (Pneumonia), 1 633 幅图像的结节数据集 (Nodule) 以及 1 759 幅图像的 Synapse 数据集. 在图像数据集上使用深度学习模型取得的抽象特征而不是原始特征通常能改善模型的性能. 为此, 我们利用预训练的卷积神经网络提取图像的抽象特征, 即利用 ResNet18 网络的 RES5B-RELU 层作为输出结果. 这样每幅图像的特征表示为 $7\times7\times512$ 的张量形式. 为了减少计算量, 对张量表示的特征沿第 1 个轴执行平均运算, 沿着第 3 个轴对特征进行下采样. 这样处理之后张量的尺寸为 $1\times7\times512/4$, 可简化为 7×128 的矩阵形式. 这样把每幅图像看作 7 个 128 维的事例从而形成集值数据. 我们随机选择 70% 的样本构成训练集, 其他图像作为测试集. 对肺炎数据集, 我们随机选择 300 幅图像来训练公式 (23) 和公式 (24) 的模型, 并应用定理 4 和定理 5 来删除非支持向量. 随后利用 TSFM 的支持向量对应的采样点来训练公式 (13) 和公式 (14) 的模型. 表 5 表示了在 4 个医学图像集上的实验结果. 实验结果取自 10 次运行的平均值.

从表 5 可看出 SFM 并不优于 TSFM, 这是因为 SFM 通常比 TSFM 提供更稀疏的支持向量. 从表 5 可知 TSFM 在这些数据集上产生了最佳性能. 在这些方法中, SFM 和 TSFM 是基于采样的方法. SANP、RCRC、SMM 和 UTSVM 考虑了图像的所有表示形式. 如果图像的表示包含冗余特征, 这些冗余特征将会影响分类器的性能.

SFM 和 TSFM 借助支持函数来选择有效的特征表示. 实验结果表明了对图像的深度特征进行集值对象的建模以及采用两个非平行的超平面分类图像是合理的.

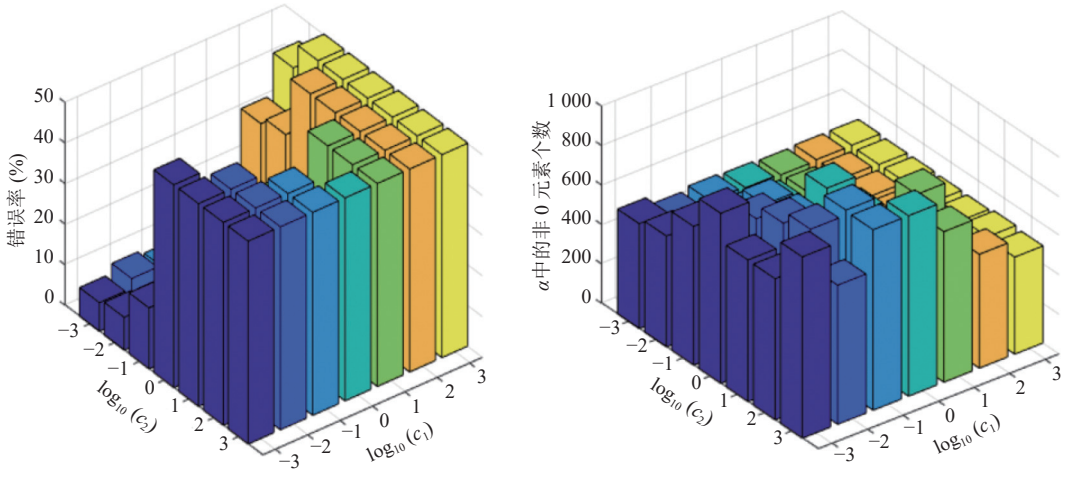
表 5　在医学图像数据集上的实验结果 (%)

数据集	SANP	RCRC	SOCP	SFM	SMM	UTSVM	TSFM
Breast	22.59±5.42	21.76±7.32	21.33±4.09	20.25±6.03	21.2±6.42	21.50±6.49	**19.6±5.46**
Pneumonia	8.52±3.88	7.32±3.25	7.15±3.02	6.37±2.89	6.65±3.21	7.28±3.02	**6.03±2.05**
Nodule	25.33±2.67	26.22±3.25	24.22±3.13	24.52±3.26	23.5±4.05	24.62±3.29	**22.4±3.12**
Synapse	20.45±4.82	19.55±5.02	17.52±4.85	17.26±4.92	16.8±4.92	19.28±4.56	**16.1±5.35**

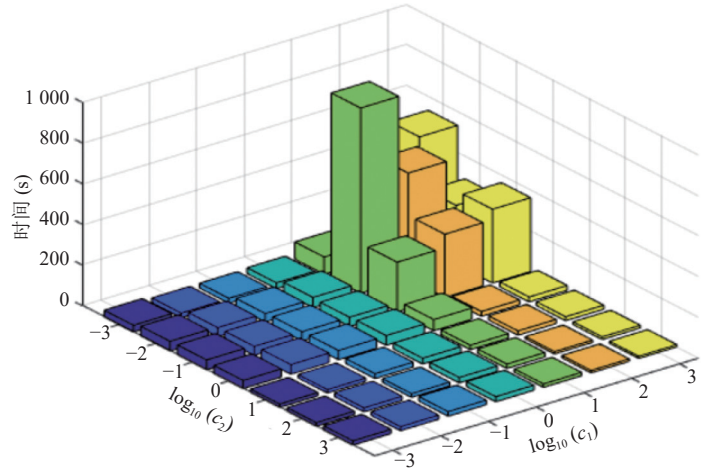
3.5　模型超参数的敏感度分析

TSFM 包含多个待确定的超参数. 与 TSVM 的超参数相类似, TSFM 的超参数将影响模型的性能. 当 TSFM 采用高斯核函数时, 核超参数 σ 会影响原数据的嵌入空间. 本节在 svmguide 数据集上进行了实验, 通过实验研究 TSFM 的超参数对其性能的影响. 该数据集由 3 089 个样本构成, 其中每个样本包含 4 个属性值和 1 个标签. 为了测试 TSFM 的超参数对其性能的影响, 随机选择一半的样本来训练 TSFM, 其余的样本用作测试集. 与 UCI 实验部分生成集值数据相类似, 利用均匀分布生成集值数据. 每个集值对象由 10 个事例组成. 为了在三维空间中可视化实验结果, 首先通过固定 $\sigma=1$ 和 $\tau=0.5$ 来研究超参数 c_1 和 c_2 对模型的影响. 实验结果如图 7 所示, 实验结果为 TSFM 的错误率, α 的非 0 元素的个数和 TSFM 的运行时间. 从图 7(a) 可看出, TSFM 的分类性能随超参数 c_1 和 c_2 的变化而变化. 当超参数 c_1 和 c_2 取较小的值时, TSFM 会取得较低的错误率. 从图 7(b) 可看出, α 的稀疏性受到超参数 c_1 和 c_2 的影响. 当 c_1 取较大值时, 模型有更少的支持向量. 在这种情况下, TSFM 的错误率非常高. 从图 7(c) 可观测到当 c_1 取较大值时和 c_2 取较小值时, 训练 TSFM 花费了较多的时间.

接下来通过固定 $c_1=1$ 和 $c_2=100$ 来研究超参数 σ 和 τ 对模型的影响. 图 8 表示了 TSFM 的实验结果, 实验结果为 TSFM 的错误率, α 中的非 0 元素的个数和 TSFM 的运行时间. 从图 8(a) 可看出, 超参数 σ 和 τ 会影响 TSFM 的分类性能. 如果超参数 σ 取太小的值, 则很难区分嵌入空间的样本, 这是因为嵌入空间的样本是相似的. 当超参数 σ 取太大值时, 很难挖掘样本之间的关系, 这是因为变换后的样本几乎是不相关的. 从图 8(a) 可知, 这两种情况导致了大的错误率. 从图 8(b) 可看出, α 的非 0 元素的个数随着 σ 和 τ 的变化而变化. 超参数 σ 取较小值会产生较多的支持向量. 从图 8(c) 可看出, TSFM 的运行时间受到超参数 σ 的影响. 当 σ 取较小值和 τ 为非 0 时, TSFM 一般需要更长的运行时间. 总的来说, 为了获得更好的分类结果, 应该注意超参数的选择问题, 通常对所有超参数的一系列值进行交叉验证从而选取对应良好性能的超参数.

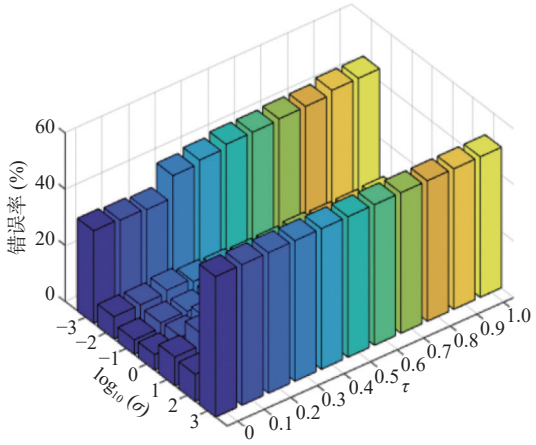


(a) TSFM 的错误率　　(b) α 中非 0 元素的个数

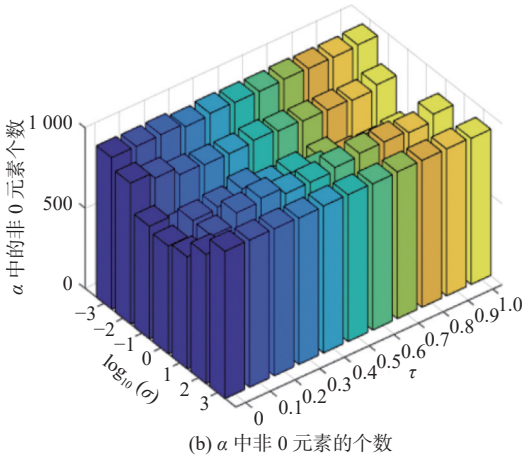


(c) TSFM 的运行时间

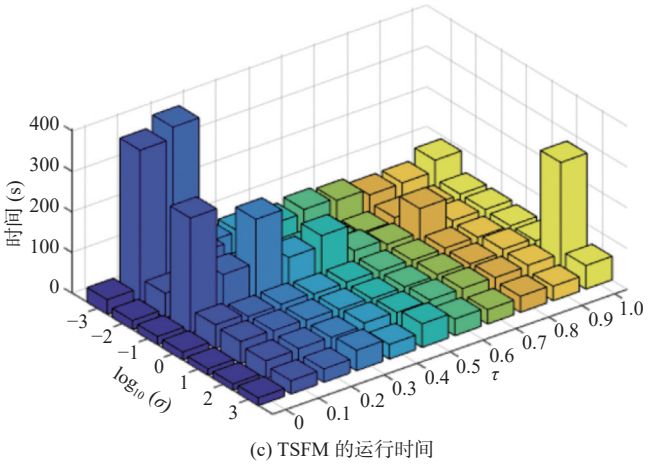
图 7　TSFM 的超参数 c_1 和 c_2 的敏感度分析



(a) TSFM 的错误率



(b) α 中非 0 元素的个数



(c) TSFM 的运行时间

图 8　TSFM 的超参数 σ 和 τ 的敏感度分析

4 结论和展望

本文利用弹球损失函数提出了新颖的超平面学习模型并用来处理集值数据. 不同于以前的模型, 提出的模型能取得巴拿赫空间的非平行的超平面. 通过研究模型的一些性质取得了易处理的优化模型, 并利用采样策略把有限维空间的优化问题转化成二次规划问题. 值得指出的是利用支持函数的核化取得的模型仍然是凸优化问题, 这使得模型可直接用于不定核函数. 在一系列的数据集上做了许多实验, 从实验可知: (1) TSFM 能获取交叉类型的集值数据的内在结构; (2) 当集值对象的事例是非高斯分布时, TSFM 比基于高斯建模的分类方法取得更好的分类性能; (3) TSFM 能有效地抑制集值数据的离群点; (4) 当集值对象包含少量高维事例时, TSFM 比基于概率建模的方法 (UTSVM, SMM, SOCP) 取得更好的分类性能; (5) TSFM 包含超参数, 这些超参数是数据依赖的, 即不同的数据集对应不同的最优超参数, 通常采用交叉验证来选择最优超参数. 尽管本文利用狄拉克测度的线性组合构成的测度空间取得了有限维空间的优化模型, 但是否存在连续测度是模型的解是值得探索的问题.

References:

[1] Panja R, Pal NR. MS-SVM: Minimally spanned support vector machine. Applied Soft Computing, 2018, 64: 356–365. [doi: 10.1016/j.asoc.2017.12.017]

[2] Wang XM, Wang ST, Huang ZX, Du YJ. Condensing the solution of support vector machines via radius-margin bound. Applied Soft Computing, 2021, 101: 107071. [doi: 10.1016/j.asoc.2020.107071]

[3] Zhai Z, Gu B, Li X, Huang H. Safe sample screening for robust support vector machine. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 6981–6988. [doi: 10.1609/aaai.v34i04.6182]

[4] Wang J, Cherian A. Discriminative video representation learning using support vector classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 43(2): 420–433. [doi: 10.1109/tpami.2019.2937292]

[5] Qiao Y, Wu K, Jin P. Efficient anomaly detection for high-dimensional sensing data with one-class support vector machine. IEEE Trans. on Knowledge and Data Engineering, 2023, 35(1): 404–417. [doi: 10.1109/tkde.2021.3077046]

[6] Tao XM, Li Q, Guo WJ, Ren C, Li CX, Liu R, Zou JR. Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. Information Sciences, 2019, 487: 31–56. [doi: 10.1016/j.ins.2019.02.062]

[7] Ding SF, Zhang J, Zhang XK, An YX. Survey on multi class twin support vector machines. Ruan Jian Xue Bao/Journal of Software, 2018, 29(1): 89–108 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5319.htm [doi: 10.13328/j.cnki.jos.005319]

[8] Zhang JH, Lai ZH, Kong H, Shen LL. Robust twin bounded support vector classifier with manifold regularization. IEEE Trans. on Cybernetics, 2023, 53(8): 5135–5150. [doi: 10.1109/tcyb.2022.3160013]

[9] Sun SL, Xie XJ, Dong C. Multiview learning with generalized eigenvalue proximal support vector machines. IEEE Trans. on Cybernetics, 2019, 49(2): 688–697. [doi: 10.1109/TCYB.2017.2786719]

[10] Rezvani S, Wu JH. Handling multi-class problem by intuitionistic fuzzy twin support vector machines based on relative density information. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14653–14664. [doi: 10.1109/tpami.2023.3310908]

[11] Tian YJ, Qi ZQ, Ju XC, Shi Y, Liu XH. Nonparallel support vector machines for pattern classification. IEEE Trans. on Cybernetics, 2014, 44(7): 1067–1079. [doi: 10.1109/tcyb.2013.2279167]

[12] Xu YT, Yang ZJ, Pan XL. A novel twin support-vector machine with pinball loss. IEEE Trans. on Neural Networks and Learning Systems, 2017, 28(2): 359–370. [doi: 10.1109/TNNLS.2015.2513006]

[13] Tanveer M, Sharma A, Suganthan PN. General twin support vector machine with pinball loss function. Information Sciences, 2019, 494: 311–327. [doi: 10.1016/j.ins.2019.04.032]

[14] Tanveer M, Tiwari A, Choudhary R, Jalan S. Sparse pinball twin support vector machines. Applied Soft Computing, 2019, 78: 164–175. [doi: 10.1016/j.asoc.2019.02.022]

[15] Qi K, Yang H. Elastic net nonparallel hyperplane support vector machine and its geometrical rationality. IEEE Trans. on Neural Networks and Learning Systems, 2022, 33(12): 7199–7209. [doi: 10.1109/TNNLS.2021.3084404]

[16] Liang ZZ, Zhang L. Intuitionistic fuzzy twin support vector machines with the insensitive pinball loss. Applied Soft Computing, 2022, 115: 108231. [doi: 10.1016/j.asoc.2021.108231]

[17] Maldonado S, López J, Vairetti C. Time-weighted fuzzy support vector machines for classification in changing environments. Information Sciences, 2021, 559: 97–110. [doi: 10.1016/j.ins.2021.01.070]

[18] Hao PY, Chiang JH, Chen YD. Possibilistic classification by support vector networks. Neural Networks, 2022, 149: 40–56. [doi: 10.1016/

j.neunet.2022.02.007]
[19] Zhu YW, Cao YR, Xue, Q, Wu QH, Zhang YS. Heavy hitter identification over large-domain set-valued data with local differential privacy. IEEE Trans. on Information Forensics and Security, 2024, 19: 414–426. [doi: 10.1109/TIFS.2023.3324726]
[20] Cao FY, Huang JZ, Liang JY, Zhao XW, Meng YF, Feng K, Qian YH. An algorithm for clustering categorical data with set-valued features. IEEE Trans. on Neural Networks and Learning Systems, 2018, 29(10): 4593–4606. [doi: 10.1109/TNNLS.2017.2770167]
[21] Chen YS, Li JJ, Lin RD, Chen DX, Huang ZH. Multi-scale set value decision information system. Control and Decision, 2022, 37(2): 455–463 (in Chinese with English abstract). [doi: 10.13195/j.kzyjc.2020.0882]
[22] Hu J, Chen Y, Zhang QH, Wang GY. Optimal scale selection for generalized multi-scale set-valued decision systems. Journal of Computer Research and Development, 2022, 59(9): 2027–2038 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.20210196]
[23] Shivaswamy PK, Bhattacharyya C, Smola AJ. Second order cone programming approaches for handling missing and uncertain data. Journal of Machine Learning Research, 2006, 7: 1283–1314.
[24] Tzelepis C, Mezaris V, Patras I. Linear maximum margin classifier for learning from uncertain data. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2948–2962. [doi: 10.1109/tpami.2017.2772235]
[25] Liang ZZ, Zhang L. Uncertainty-aware twin support vector machines. Pattern Recognition, 2022, 129: 108706. [doi: 10.1016/j.patcog.2022.108706]
[26] Liang ZZ, Ding SF. Fuzzy twin support vector machines with distribution inputs. IEEE Trans. on Fuzzy Systems, 2024, 32(1): 240–254. [doi: 10.1109/tfuzz.2023.3296503]
[27] Muandet K, Fukumizu K, Dinuzzo F, Schölkopf B. Learning from distributions via support measure machines. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012, 10–18. [doi: 10.5555/2999134.2999136]
[28] Chen JQ, Hu QH, Xue XP, Ha MH, Ma LT. Support function machine for set-based classification with application to water quality evaluation. Information Sciences, 2017, 388–389: 48–61. [doi: 10.1016/j.ins.2017.01.001]
[29] Chen JQ, Xue XP, Ma LT, Ha MH. Separability of set-valued data sets and existence of support hyperplanes in the support function machine. Information Sciences, 2018, 430–431: 432–443. [doi: 10.1016/j.ins.2017.11.057]
[30] Royden HL, Fitzpatrick PM. Real Analysis. 5th ed., London: Pearson, 2022.
[31] Chen JQ, Hu QH, Xue XP, Ha MH, Ma LT, Zhang XC, Yu ZP. Possibility measure based fuzzy support function machine for set-based fuzzy classifications. Information Sciences, 2019, 483: 192–205. [doi: 10.1016/j.ins.2019.01.022]
[32] Komornik V. Spaces of continuous functions. In: Komornik V, ed. Lectures on Functional Analysis and the Lebesgue Integral. London: Springer, 2016. 257–304. [doi: 10.1007/978-1-4471-6811-9_8]
[33] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. London: Cambridge University Press, 2000
[34] Aggarwal CC. Yu PS. Outlier detection with uncertain data. In: Proc. of the 2008 SIAM Int'l Conf. on Data Mining. Atlanta: SIAM, 2008. 483–493. [doi: 10.1137/1.9781611972788.44]
[35] Hu YQ, Mian AS, Owens R. Sparse approximated nearest points for image set classification. In: Proc. of the 2011 Conf. on Computer Vision and Pattern Recognition. Colorado: IEEE, 2011. 121–128. [doi: 10.1109/cvpr.2011.5995500]
[36] Zhu PF, Zuo WM, Zhang L, Shiu SCK, Zhang D. Image set-based collaborative representation for face recognition. IEEE Trans. on Information Forensics and Security, 2014, 9(7): 1120–1132. [doi: 10.1109/TIFS.2014.2324277]
[37] Demšar J. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 2006, 7: 1–30.
[38] Yang JC, Shi R, Ni BB. MedMNIST classification Decathlon: A lightweight AutoML benchmark for medical image analysis. In: Proc. of the 18th IEEE Int'l Symp. on Biomedical Imaging. Nice: IEEE, 2021. 191–195. [doi: 10.1109/isbi48211.2021.9434062]

附中文参考文献:

[7] 丁世飞, 张健, 张谢锴, 安悦瑄. 多分类孪生支持向量机研究进展. 软件学报, 2018, 29(1): 89–108. http://www.jos.org.cn/1000-9825/5319.htm [doi: 10.13328/j.cnki.jos.005319]
[21] 陈应生, 李进金, 林荣德, 陈东晓, 黄哲煌. 多尺度集值决策信息系统. 控制与决策, 2022, 37(2): 455–463. [doi: 10.13195/j.kzyjc.2020.0882]
[22] 胡军, 陈艳, 张清华, 王国胤. 广义多尺度集值决策系统最优尺度选择. 计算机研究与发展, 2022, 59(9): 2027–2038. [doi: 10.7544/issn1000-1239.20210196]

梁志贞(1976－), 男, 博士, 副教授, 主要研究领域为机器学习, 模式识别, 数据挖掘.

闵玉寒(2000－), 男, 硕士生, 主要研究领域为模式识别, 机器学习.

丁世飞(1963－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为人工智能与模式识别, 机器学习, 数据挖掘, 大数据智能分析.