

DOI:10.13364/j.issn.1672-6510.20220114

基于最优间隔分布的最小二乘支持向量机

张传雷¹, 武孟艳¹, 可 婷¹, 王 波², 王 伟²

(1. 天津科技大学人工智能学院, 天津 300457; 2. 思腾合力(天津)科技有限公司, 天津 301799)

摘 要: 支持向量机(support vector machine, SVM)通过结构风险最小化构建模型,模型简单易理解,但求解过程较复杂;而最小二乘支持向量机(least squares support vector machine, LSSVM)求解的是一组线性方程组,其求解效率远高于 SVM. 然而,传统 LSSVM 忽略了样本点到超平面分布的影响. 为了充分考虑数据的分布,本研究从 LSSVM 出发,加入最小化距离方差与最大化距离均值正则项,构建全新的 LSSVM 模型(LSSVM_{rv} 和 LSSVM_{rm_rv} 模型). 该模型的优势在于:依据数据的分布特征构建模型,分类效果更好;对偶问题仍是一组线性方程组,计算成本较小. 大量的数值实验进一步验证上述模型的优越性,与传统 SVM 和 LSSVM 相比,在 6 个数据集上的分类准确率均有所提高,并且节省了大量的训练时间.

关键词: 分类; 最小二乘支持向量机; 概率分布; 泛化能力

中图分类号: TP391

文献标志码: A

文章编号: 1672-6510(2023)02-0056-07

Least Squares Support Vector Machine Based on Optimal Margin Distribution

ZHANG Chuanlei¹, WU Mengyan¹, KE Ting¹, WANG Bo², WANG Wei²

(1. College of Artificial Intelligence, Tianjin University of Science & Technology, Tianjin 300457, China;

2. Sitonholly (Tianjin) Technology Co., Ltd., Tianjin 301799, China)

Abstract: Support vector machine (SVM) constructs a model by minimizing structural risk. The model is simple and easy to understand, but its solution process is complex. By contrast, the least squares support vector machine (LSSVM) solves a group of linear equations, and its efficiency is much higher than that of SVM. However, the traditional LSSVM ignores the influence of sample points to hyperplane distribution. In order to fully consider the distribution of data, this study starts from LSSVM and adds the regularization term of minimizing distance variance and maximizing distance mean to construct a new LSSVM model (LSSVM_{rv} model and LSSVM_{rm_rv} model). The advantage of this model is that it is built according to the distribution characteristics of data, and the classification effect is better. The dual problem is still a group of linear equations, but the computational cost is small. A large number of numerical experiments have further proven the superiority of the new model. Compared with traditional SVM and LSSVM, the classification accuracy on the six data sets has been improved and has saved a lot of training time.

Key words: classification; least squares support vector machine; probability distribution; generalization ability

随着信息化以及人工智能的发展,许多工作被机器所替代,计算机的能力不断提升并且可以进行繁琐的工作,提高了工作效率,使人们的生活更加便捷、高效. 分类任务^[1]是计算机可以实现的任务之一,利用计算机进行分类在日常生活中已经屡见不

鲜,如常见的医疗诊断^[2]、垃圾邮件分类^[3]、骚扰电话识别^[4]等都属于计算机分类在生活中的应用. 计算机进行分类的基本流程为:首先通过数据进行知识提取,利用提取的知识生成分类预测模型;当有新的数据需要分类时,利用搭建好的预测模型对数据进行分

收稿日期: 2022-05-03; 修回日期: 2022-09-05

基金项目: 天津市自然科学基金资助项目(18JCZDJC32100)

作者简介: 张传雷(1973—),男,山东沂源人,教授;通信作者: 可 婷,讲师, keting@tust.edu.cn

类和输出,从而实现数据的分类预测。

随着人工智能的发展,大量的分类模型被提出,基于深度学习的方法为分类工作的发展做出了巨大的贡献,不断涌现的神经网络算法^[5-8]也极大地提高了模型的分类准确率。深度学习模型需要大量数据进行训练才能取得较好的学习效果,但现实生活中的数据往往不充足。基于机器学习的分类模型可以避免使用大量的训练数据,利用小规模训练数据也可使模型获得较好的预测精度。随着机器学习的不断发展,基于机器学习的分类算法也在不断完善。支持向量机(support vector machine, SVM)因其强大的分类能力以及数据稀疏性等优点,在机器学习算法中具有重要地位。SVM 通过寻找一个超平面进行数据划分,其基本思想较容易理解,模型简单易懂^[9]。SVM 具有强大的泛化能力,现已成为重要的机器算法之一。

传统的 SVM 是寻找一个超平面对样本进行分割,其分割的原则是间隔最大化,因而可以很好地解决线性可分问题。对于非线性可分问题, SVM 模型可以引入核技巧,通过核映射将原始数据映射到高维空间,从而实现数据的线性可分。传统 SVM 模型是一个凸二次规划问题求解,求解过程复杂。为了更好地求解模型,Platt^[10]提出了序列最小优化算法(sequential minimal optimization, SMO),该算法每次迭代只改变两个样本的拉格朗日乘子,将模型划分为若干个小的子模型进行迭代求解,避开复杂的优化求解过程,提升了模型的训练速率。虽然通过 SMO 算法使 SVM 的训练速率得以提升,但是模型依旧求解若干个子模型,因而需要一定的训练时间。坐标下降法^[11]也常用于求解 SVM 模型,通过在梯度方向的迭代以接近模型最优值,然而训练时间仍较长,并且容易陷入局部最小值。因此, SVM 模型的求解仍存在一定的局限性。Suykens 等^[12]提出最小二乘支持向量机(least squares support vector machine, LSSVM)模型,将 SVM 中的不等式约束转化为等式约束,将凸二次规划问题求解转化为线性方程组求解,使模型求解变得简单,大大减少了模型的训练时间,求解过程也更容易被理解。Zhang 等^[13]对 SVM 间隔最大化的想法进行创新,认为间隔最大化思想并没有考虑到支持向量的概率分布,所以在支持向量模型中引入了概率分布均值和方差;加入概率分布的优化分布机为最优间隔分布学习机(optimal margin distribution machine, ODM),可以更好地表述数据样本的分布方向。

最小二乘支持向量机通过求解线性方程组获得

极高的训练速率,节省了大量的训练时间。优化分布机表征了概率分布,将间隔带边缘的数据分布考虑其中,使模型更贴合数据的分布方向。结合两者的优点,本研究拟在 LSSVM 中引入概率分布,旨在提高 LSSVM 模型的分类效果,并在相同数据集下与 SVM、LSSVM、优化分布机进行对比实验。

1 相关理论知识

本文的变量及描述属性见表 1。传统 SVM 是通过寻找一个超平面进行数据划分。对于线性可分数据,原始模型就可达到数据分类的效果;对于非线性可分数据,在原始空间中不存在可以完全划分两类的超平面。SVM 的一个优点是可以在模型中引入核函数,将原始空间数据映射到高维空间,从而实现数据的线性可分,因此 SVM 的功能足以实现大量的分类任务。

表 1 变量及描述
Tab. 1 Variables and descriptions

变量	描述
$\{x_1, x_2, \dots, x_n\}$	数据集
$y = [y_1, y_2, \dots, y_n]$	类别标签, $y_i = \{1, -1\}, i = 1, \dots, n$
$\varphi(x)$	核函数, $K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle$
n	训练集样本数量
ξ_i	松弛变量
$X = [\varphi(x_1), \dots, \varphi(x_n)]$	核映射空间
$e = \underbrace{[1, \dots, 1]}_{n \times 1}$	全 1 列向量
$Y = \text{diag}(y)$	对角矩阵
$\xi = [\xi_1, \dots, \xi_n]$	松弛变量列向量

1.1 最小二乘支持向量机

传统 SVM 解决的是一个凸二次规划问题,是一个不等式约束问题,而将 SVM 转化为对偶问题后,其参数的个数与训练样本的数量成正比。当训练样本数量较大时,参数的调整工作就变得繁琐,并且与不等式约束问题的求解相比,等式约束问题的求解会简单许多,训练速率也会大幅提升。因此, Suykens 等^[12]提出的 LSSVM 模型将 SVM 模型中的不等式约束变为等式约束,考虑每一个训练样本到平面间的距离,大大提升了模型的训练速率。同时,将凸二次规划问题求解转变为求解一组线性方程组,具体模型为

$$\begin{aligned} \min_{w, b, \xi} & \frac{1}{2} \|w\|^2 + c \sum_{i=1}^n \xi_i^2 \\ \text{s.t. } & y_i (w^T \varphi(x_i) + b) = 1 - \xi_i \end{aligned}$$

$$\xi_i \geq 0, i=1, \dots, n \quad (1)$$

式中: $\|\mathbf{w}\|^2$ 为正则项; c 为平衡系数; $\mathbf{b}$ 为偏置项; ξ_i 为松弛变量; $\mathbf{w}$ 为法向量, 其参数个数与数据集维度一致. 该模型旨在寻找一个超平面 $\mathbf{w}^T \varphi(x) + \mathbf{b} = 0$, 使两类样本点尽可能靠近对应的平行超平面 $\mathbf{w}^T \varphi(x) + \mathbf{b} = \pm 1$. 直观理解为: 寻找两个平行的超平面, 使每一类样本尽可能在对应的超平面附近, 而用于决策的超平面位于两个平行超平面的中间位置. 引入拉格朗日方程, 令其对各参数的偏置为 0, 可将原问题转化为求解以下线性方程组

$$\begin{bmatrix} 0 & -\mathbf{y}^T \\ \mathbf{y} & \mathbf{YX}^T \mathbf{XY} + \frac{\mathbf{I}}{2c} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{e} \end{bmatrix} \quad (2)$$

式中: $\boldsymbol{\alpha} = [\alpha_1; \alpha_2; \dots; \alpha_n]$ 为拉格朗日乘子, $\mathbf{I}$ 为维度为 n 的单位矩阵. 通过求解上式可获得偏置项 $\mathbf{b}$ 以及参数 $\boldsymbol{\alpha}$ 的值, 得到参数后即可根据决策函数 $f(x) = \text{sgn}\left(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + \mathbf{b}\right)$ 判断新样本的类别, 其中: $\text{sgn}(x)$ 为符号函数, 当 $x > 0$ 时取 1, 当 $x < 0$ 时取 -1, $x = 0$ 时取 0; $K(x_i, x)$ 为 x_i 与 x 的核映射函数.

1.2 大间隔分布学习机

传统 SVM 与 LSSVM 都通过最大化样本距离建立模型, 而基于 SVM 的大间隔分布学习机 (large margin distribution machine, LDM) 引入了数据的距离均值以及方差, 用其值描述概率分布^[14], 其中均值 (γ_m) 与方差 (γ_v) 为

$$\begin{aligned} \gamma_m &= \frac{1}{n} \sum_{i=1}^n y_i \mathbf{w}^T \varphi(x_i) = \frac{1}{n} (\mathbf{XY})^T \mathbf{w} \\ \gamma_v &= \frac{1}{n} \sum_{i=1}^n (y_i \mathbf{w}^T \varphi(x_i) - \gamma_m)^2 = \mathbf{w}^T \mathbf{X} \frac{n\mathbf{I} - \mathbf{yy}^T}{n^2} \mathbf{X}^T \mathbf{w} \end{aligned} \quad (3)$$

LDM 模型^[15]为

$$\begin{aligned} \min_{\mathbf{w}, \xi_i} & \frac{1}{2} \|\mathbf{w}\|^2 + \mu_1 \gamma_v - \mu_2 \gamma_m + \frac{\lambda}{n} \sum_{i=1}^n \xi_i \\ \text{s.t. } & y_i \mathbf{w}^T \varphi(x_i) \geq 1 - \xi_i, \xi_i \geq 0, \forall i \end{aligned} \quad (4)$$

其中: μ_1 、 μ_2 为平衡系数, 需要提前设定, 当其值都为 0 时, 模型类似于传统 SVM 模型; λ 为松弛变量的平衡系数, 需要提前设定. 在传统 SVM 模型中引入最大化均值以及最小化方差的概念, 可将模型进一步简化为

$$\begin{aligned} \min_{\mathbf{w}, \xi_i} & \frac{1}{2} \|\mathbf{w}\|^2 + \mathbf{w}^T \mathbf{X} \frac{\mu_1 (n\mathbf{I} - \mathbf{yy}^T)}{n^2} \mathbf{X}^T \mathbf{w} - \\ & \frac{\mu_2}{n} (\mathbf{XY})^T \mathbf{w} + \frac{\lambda}{n} \mathbf{e}^T \boldsymbol{\xi} \\ \text{s.t. } & \mathbf{YX}^T \mathbf{w} \geq \mathbf{e} - \boldsymbol{\xi}, \boldsymbol{\xi} \geq 0 \end{aligned} \quad (5)$$

通过引入拉格朗日方程, 可以得到模型的对偶问题为

$$\begin{aligned} \min_{\boldsymbol{\alpha}} & \frac{1}{2} \left(\frac{\mu_2}{n} \mathbf{e} + \boldsymbol{\alpha} \right)^T \mathbf{YX}^T \mathbf{Q}^{-1} \mathbf{XY} \left(\frac{\mu_2}{n} \mathbf{e} + \boldsymbol{\alpha} \right) - \mathbf{e}^T \boldsymbol{\alpha} \\ \text{s.t. } & 0 \leq \boldsymbol{\alpha} \leq \frac{\lambda}{n} \mathbf{e} \end{aligned} \quad (6)$$

其中 $\mathbf{Q} = \mathbf{I} + 2\mu_1 \mathbf{X} (n\mathbf{I} - \mathbf{yy}^T) \mathbf{X}^T / n^2$.

通过求解上述模型, 获得最优参数 $\boldsymbol{\alpha}^*$, 为

$$\mathbf{w} = \mathbf{X} (\mathbf{I} + \mathbf{AG})^{-1} \mathbf{Y} \left(\frac{\mu_2}{n} \mathbf{e} + \boldsymbol{\alpha}^* \right) = \mathbf{X} \mathbf{v} \quad (7)$$

其中 $\mathbf{A} = 2\mu_1 (n\mathbf{I} - \mathbf{yy}^T) / n^2$, $\mathbf{G} = \mathbf{X}^T \mathbf{X}$ 为核函数, $\mathbf{v} = (\mathbf{I} + \mathbf{AG})^{-1} \mathbf{Y} (\mu_2 \mathbf{e} / n + \boldsymbol{\alpha}^*)$, 其模型决策函数为

$$\begin{aligned} f(x) &= \text{sgn}(\mathbf{w}^T \varphi(x)) = \\ & \text{sgn}\left(\sum_{i=1}^n v_i K(x_i, x)\right) \end{aligned} \quad (8)$$

其中 v_i 为 $\mathbf{v}$ 的第 i 个元素.

1.3 最优间隔分布学习机

Zhang 等^[13]对 LDM 算法进行改进, 受传统 SVM 最大间隔默认为 1 的启发, 在 LDM 的基础上也将均值固定为 1, 得到模型

$$\begin{aligned} \min_{\mathbf{w}, \xi_i, \varepsilon_i} & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{\lambda}{n} \sum_{i=1}^n (\xi_i^2 + \varepsilon_i^2) \\ \text{s.t. } & y_i \mathbf{w}^T \varphi(x_i) \geq 1 - \xi_i \\ & y_i \mathbf{w}^T \varphi(x_i) \leq 1 + \varepsilon_i, \forall i \end{aligned} \quad (9)$$

此时, λ 为模型的平衡系数, 由于模型中的 ξ_i 与 ε_i 中至少有一个大于等于 0, 其值为样本点到间隔带的距离, 因此第 2 项相当于间隔带的概率分布方差. SVM 中为保证间隔带具有一定的容错率, 在模型中引入松弛变量, 因此 ODM 模型为距离均值也引入松弛变量 θ , 同时引入参数 μ 用于平衡两种不同偏差, 模型变为

$$\begin{aligned} \min_{\mathbf{w}, \xi_i, \varepsilon_i} & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{\lambda}{n} \sum_{i=1}^n \frac{\xi_i^2 + \mu \varepsilon_i^2}{(1 - \theta)^2} \\ \text{s.t. } & y_i \mathbf{w}^T \varphi(x_i) \geq 1 - \theta - \xi_i \\ & y_i \mathbf{w}^T \varphi(x_i) \leq 1 + \theta + \varepsilon_i, \forall i \end{aligned} \quad (10)$$

与 LDM 类似, 对原模型求拉格朗日方程, 令偏导为 0 后, 得到其对偶问题为

$$\min_{\alpha} \frac{1}{2} \alpha^T \begin{bmatrix} Q + \frac{n(1-\theta)^2}{2\lambda} I & -Q \\ -Q & Q + \frac{n(1-\theta)^2}{2\lambda\mu} I \end{bmatrix} \alpha + \begin{bmatrix} (\theta-1)e \\ (\theta+1)e \end{bmatrix}^T \alpha$$

s.t. $\alpha \geq 0$ (11)

其中 $Q = YX^T XY$, 通过求解模型, 获得最优参数 α^* , 参数 w 求解为

$$w = XY[I, -I]\alpha^* = Xv \quad (12)$$

其中 $v = Y[I, -I]\alpha^*$, 其模型决策函数为

$$f(x) = \text{sgn}(w^T \phi(x)) = \text{sgn}\left(\sum_{i=1}^n v_i K(x_i, x)\right) \quad (13)$$

上述两个模型通过变换都类似于求解以下二次规划问题

$$\min_{\alpha} f(\alpha) = \frac{1}{2} \alpha^T H \alpha + q^T \alpha$$

s.t. $0 \leq \alpha \leq u$ (14)

其中 H 为 n 维度方阵, q 为 n 维度列向量, 为求解上述框架模型, 采用坐标下降法求解模型, 如算法 1 所示.

算法 1: ODM_{dcd}

1. 输入: 数据集 X 、标签矩阵 Y .
2. 初始化 $\alpha = 0$, 计算 H 与 q .
3. 计算 $[\nabla f(\alpha)]_i = [H\alpha + q]_i$, $i = 1, \dots, n$.
4. 计算 $\alpha_i = \min\left(\max\left(\alpha_i - \frac{[\nabla f(\alpha)]_i}{h_i}, 0\right), u_i\right)$.
5. 若 α 不收敛, 重复此步骤 3 与步骤 4.
6. 输出 α .

2 模型构建

SVM 与 LSSVM 具有最大化最小边缘的想法, 而实验证明, 只考虑最大间隔宽度的模型无法有效拟合数据的分布方向, 使模型无法达到更好的泛化性能. 概率分布会对模型起到关键作用, 因此, 本文受 ODM 模型的启发, 将概率分布考虑到 LSSVM 模型中, 用训练样本到超平面距离的方差描述概率分布. 在模型中, 用最小化方差考虑数据的分布方向, 模型示意图如图 1 所示. 其中虚线为两类数据的基本分布情况, h_{LSSVM} 为传统 LSSVM 获得的超平面, h_{dist}

为本文模型获得的超平面. 通过观察可知, h_{dist} 可以更好表述数据的分布特征, 从而进行数据划分.

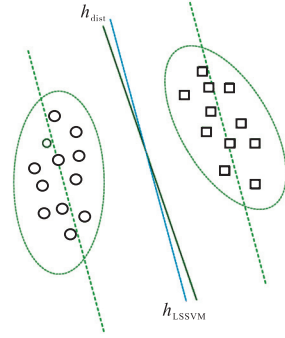


图 1 模型示意图

Fig. 1 Model diagram

本文利用概率分布的思想, 将概率分布方差引入 LSSVM 模型中, 利用方差影响决策面的方向, 从而改善模型获得新的 LSSVM_{rv} 模型, 其中概率分布的均值与方差为

$$\begin{aligned} \gamma_m &= \frac{1}{n} \sum_{i=1}^n y_i [\mathbf{w}^T \phi(x_i) + b] = \frac{1}{n} (Xy)^T \mathbf{w} + \frac{b}{n} e^T y \\ \gamma_v &= \frac{1}{n} \sum_{i=1}^n \{y_i [\mathbf{w}^T \phi(x_i) + b] - \gamma_m\}^2 = \\ &= \mathbf{w}^T X \frac{nI - yy^T}{n^2} X^T \mathbf{w} - \frac{2b}{n^2} e^T y (Xy)^T \mathbf{w} + \\ &+ \frac{2b}{n} e^T X^T \mathbf{w} - \frac{b^2}{n^2} (e^T y)^2 + b^2 \end{aligned} \quad (15)$$

其中 b 为模型的偏置项, 加入方差后的 LSSVM_{rv} 模型为

$$\begin{aligned} \min_{\mathbf{w}, b, \xi_i} & \frac{1}{2} \|\mathbf{w}\|^2 + c_1 \sum_{i=1}^n \xi_i^2 + c_2 \gamma_v \\ \text{s.t. } & y_i (\mathbf{w}^T \phi(x_i) + b) = 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, \dots, n \end{aligned} \quad (16)$$

其中 c_1 、 c_2 为该模型的平衡参数, 引入上述模型的拉格朗日方程为

$$\begin{aligned} L &= \frac{1}{2} \mathbf{w}^T \mathbf{w} + c \sum_{i=1}^n \xi_i^2 + \\ &+ c_2 \left(\mathbf{w}^T X \frac{nI - yy^T}{n^2} X^T \mathbf{w} - \frac{2b}{n^2} e^T y (Xy)^T \mathbf{w} \right) + \\ &+ c_2 \left(\frac{2b}{n} e^T X^T \mathbf{w} - \frac{b^2}{n^2} (e^T y)^2 + b^2 \right) - \\ &- \sum_{i=1}^n \alpha_i \{y_i [\mathbf{w}^T \phi(x_i) + b] - 1 + \xi_i\} \end{aligned} \quad (17)$$

将上述拉格朗日方程对 $\mathbf{w}$ 、 b 、 ξ_i 、 α_i 进行求导, 并令其为 0, 结果如下:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} + 2c_2 \mathbf{X} \frac{n\mathbf{I} - \mathbf{y}\mathbf{y}^T}{n^2} \mathbf{X}^T \mathbf{w} - \\
&\quad \frac{2bc_2}{n^2} \mathbf{e}^T \mathbf{y} \mathbf{X} \mathbf{y} + \frac{2bc_2}{n} \mathbf{X} \mathbf{e} - \mathbf{X} \mathbf{Y} \boldsymbol{\alpha} = 0 \\
\frac{\partial L}{\partial \mathbf{b}} &= \frac{2c_2}{n^2} \mathbf{e}^T \mathbf{y} (\mathbf{X} \mathbf{y})^T \mathbf{w} + \frac{2c_2}{n} \mathbf{e}^T \mathbf{X}^T \mathbf{w} - \\
&\quad \frac{2c_2}{n^2} (\mathbf{e}^T \mathbf{y})^2 \mathbf{b} + 2c_2 \mathbf{b} - \mathbf{y}^T \boldsymbol{\alpha} = 0 \\
\frac{\partial L}{\partial \xi_i} &= 2c_1 \xi_i - \alpha_i = 0 \\
\frac{\partial L}{\partial \alpha_i} &= y_i [\mathbf{w}^T \boldsymbol{\varphi}(x_i) + \mathbf{b}] - 1 + \xi_i = 0
\end{aligned} \quad (18)$$

其中 $\boldsymbol{\alpha} = [\alpha_1; \alpha_2; \dots; \alpha_n]$, 通过求解上式可得

$$\begin{aligned}
\mathbf{w} &= \left(\mathbf{I} + 2c_2 \mathbf{X} \frac{n\mathbf{I} - \mathbf{y}\mathbf{y}^T}{n^2} \mathbf{X}^T \right)^{-1} \left(\mathbf{X} \mathbf{Y} \boldsymbol{\alpha} + \right. \\
&\quad \left. \frac{2bc_2}{n^2} \mathbf{e}^T \mathbf{y} \mathbf{X} \mathbf{y} - \frac{2bc_2}{n} \mathbf{X} \mathbf{e} \right) \\
\xi &= \frac{1}{2c_1} \boldsymbol{\alpha}
\end{aligned} \quad (19)$$

将上述两个所得变量代入线性方程组, 可获得新的线性方程组

$$\begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_3 & \mathbf{H}_4 \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{e} \end{bmatrix} \quad (20)$$

$$\begin{aligned}
\text{其中: } \mathbf{H}_1 &= \left(\frac{2c_2}{n^2} \mathbf{e}^T \mathbf{y} \mathbf{y}^T + \frac{2c_2}{n} \mathbf{e}^T \right) \frac{2c_2}{n^2} \mathbf{e}^T \mathbf{y} \mathbf{S} \mathbf{y} - \\
&\quad \left(\frac{2c_2}{n^2} \mathbf{e}^T \mathbf{y} \mathbf{y}^T + \frac{2c_2}{n} \mathbf{e}^T \right) \frac{2c_2}{n} \mathbf{S} \mathbf{e} - \frac{2c_2}{n^2} (\mathbf{e}^T \mathbf{y})^2 + 2c_2 \\
\mathbf{H}_2 &= \left(\frac{2c_2}{n^2} \mathbf{e}^T \mathbf{y} \mathbf{y}^T + \frac{2c_2}{n} \mathbf{e}^T \right) \mathbf{S} \mathbf{y} - \mathbf{y}^T \\
\mathbf{H}_3 &= \frac{2c_2}{n^2} \mathbf{e}^T \mathbf{y} \mathbf{Y} \mathbf{S} \mathbf{y} - \frac{2c_2}{n} \mathbf{Y} \mathbf{S} \mathbf{e} + \mathbf{y} \\
\mathbf{H}_4 &= \mathbf{Y} \mathbf{S} \mathbf{Y} + \frac{1}{2c_1} \mathbf{I}
\end{aligned}$$

$$\text{式中: } \mathbf{S} = \mathbf{X}^T \left(\mathbf{I} + 2c_1 \frac{\mathbf{X} (n\mathbf{I} - \mathbf{y}\mathbf{y}^T) \mathbf{X}^T}{n^2} \right) \mathbf{X}$$

根据 Sherman-Morrison-Woodbury^[16-17] 公式: $(\mathbf{I} + \mathbf{X} \mathbf{A} \mathbf{X}^T)^{-1} = \mathbf{I} + \mathbf{X} (\mathbf{A}^{-1} + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, 可以进一步得到: $\mathbf{S} = \mathbf{G} (\mathbf{I} + \mathbf{A} \mathbf{G})^{-1}$, 其中 $\mathbf{A} = 2c_2 (n\mathbf{I} - \mathbf{y}\mathbf{y}^T) / n^2$, $\mathbf{G} = \mathbf{X}^T \mathbf{X}$ 为核函数. 通过计算以上线性方程组, 可以获得最优解 $\boldsymbol{\alpha}^*$, 利用最优解进行新样本的决策判断, 决策函数为 $f(x) = \text{sgn} \left(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + \mathbf{b} \right)$.

模型的算法求解过程见算法 2.

算法 2: LSSVM_rv

1. 输入: 数据集 $\mathbf{X}$ 、类别标签 $\mathbf{y}$.
2. 分别计算矩阵 $\mathbf{H}_1$ 、 $\mathbf{H}_2$ 、 $\mathbf{H}_3$ 、 $\mathbf{H}_4$.
3. 通过矩阵堆叠的方式获取 $\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_3 & \mathbf{H}_4 \end{bmatrix}$.
4. 创建列向量 $\mathbf{q} = \begin{bmatrix} 0 \\ \mathbf{e} \end{bmatrix}$.
5. 计算 $\mathbf{z} = \mathbf{H}^{-1} * \mathbf{q}$.
6. 计算 $\mathbf{b} = \mathbf{z}[1]$, $\boldsymbol{\alpha} = [\mathbf{z}[2]; \mathbf{z}[3]; \dots; \mathbf{z}[n+1]]$.
7. 输出 $\mathbf{b}$ 和 $\boldsymbol{\alpha}$.

在上述模型的基础上, 本文进一步加入了均值的概念, 形成新模型 LSSVM_rm_rv, 与原模型最小化到平行平面的距离不同的是, 尽可能最大化样本点到超平面的距离, 因此重新建立模型

$$\begin{aligned}
\min_{\mathbf{w}, \mathbf{b}, \xi_i} & \frac{1}{2} \|\mathbf{w}\|^2 + c_1 \sum_{i=1}^n \xi_i^2 + c_2 \gamma_v - c_3 \gamma_m \\
\text{s.t. } & y_i (\mathbf{w}^T \boldsymbol{\varphi}(x_i) + \mathbf{b}) = 1 - \xi_i \\
& \xi_i \geq 0, \quad i = 1, \dots, n
\end{aligned} \quad (21)$$

其中 c_1 、 c_2 、 c_3 为该模型的平衡参数, 与上述求解过程类似, 将模型变换为对偶问题

$$\begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_3 & \mathbf{H}_4 \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_5 \\ \mathbf{h}_6 \end{bmatrix} \quad (22)$$

其中 $\mathbf{H}_1$ 、 $\mathbf{H}_2$ 、 $\mathbf{H}_3$ 、 $\mathbf{H}_4$ 与最初模型一样, 不同的是 $\mathbf{h}_5$ 、 $\mathbf{h}_6$ 为

$$\begin{aligned}
\mathbf{h}_5 &= \frac{c_3}{n} \mathbf{e}^T \mathbf{y} - \frac{c_3}{n} \left(\frac{2c_2}{n^2} \mathbf{e}^T \mathbf{y} \mathbf{y}^T + \frac{2c_2}{n} \mathbf{e}^T \right) \mathbf{S} \mathbf{y}, \quad \mathbf{h}_5 \in \mathbf{R}^{1 \times 1} \\
\mathbf{h}_6 &= \mathbf{e} - \frac{c_3}{n} \mathbf{Y} \mathbf{S} \mathbf{y}, \quad \mathbf{h}_6 \in \mathbf{R}^{m \times 1}
\end{aligned} \quad (23)$$

因此, 新模型 LSSVM_rm_rv 的求解算法见算法 3.

算法 3: LSSVM_rm_rv

1. 输入: 数据集 $\mathbf{X}$ 、类别标签 $\mathbf{y}$.
2. 分别计算矩阵 $\mathbf{H}_1$ 、 $\mathbf{H}_2$ 、 $\mathbf{H}_3$ 、 $\mathbf{H}_4$ 、 $\mathbf{h}_5$ 、 $\mathbf{h}_6$.
3. 通过矩阵堆叠方式获取 $\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 \\ \mathbf{H}_3 & \mathbf{H}_4 \end{bmatrix}$ 、 $\mathbf{q} = \begin{bmatrix} \mathbf{h}_5 \\ \mathbf{h}_6 \end{bmatrix}$.
4. 计算 $\mathbf{z} = \mathbf{H}^{-1} * \mathbf{q}$.
5. 计算 $\mathbf{b} = \mathbf{z}[1]$, $\boldsymbol{\alpha} = [\mathbf{z}[2]; \mathbf{z}[3]; \dots; \mathbf{z}[n+1]]$.
6. 输出 $\mathbf{b}$ 和 $\boldsymbol{\alpha}$.

通过 LSSVM_rv 与 LSSVM_rm_rv 模型的计算, 可以考虑数据样本的分布方向, 获得更加合理的决策面, 并且继承 LSSVM 便于求解的优点, 其训练速率大幅提升. 通过在不同数据集上的实验, 在分类准确

率以及训练时间方面进行模型对比,从而验证本文模型的有效性。

3 验证实验

3.1 实验设置

上述模型都是在二分类的基础上进行搭建,因此本文将在以下 6 组常见的二分类 UCI 数据集上进行实验验证。与 ODM 实验类似,在实验之前对数据集特征进行最大最小归一化,将特征值映射到 $[0, 1]$ 区间内,同时随机取样本的 80% 作为训练数据集,利用训练数据集进行模型训练,每个模型的最优参数通过网格搜索法进行最优超参数的选取,剩余的样本作为测试数据集,验证训练好的模型,具体数据集信息见表 2。

由于上述数据集均为非线性可分数据集,为使数据可以实现线性可分,因此在模型中引入核函数,在实验中采用的是高斯核函数,具体形式为

$$\varphi(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle = \exp\left(-\frac{\|x_1 - x_2\|^2}{\sigma^2}\right) \quad (24)$$

核函数确定后进行参数设置,本文模型中的 c_1 、 c_2 与 c_3 参数范围为 $\{2^{-10}, 2^{-9}, \dots, 2^9, 2^{10}\}$, 其中每个模型都引入了核函数,因此核参数 σ 的范围均为

$\{2^{-5}, 2^{-4}, 2^{-2}, 2^0, 2^2, 2^5\}$ 。通过网格搜索的方法对各个模型进行实验,在确定了最优参数后,对上述每个数据集进行 10 次实验,记录每次实验的准确率以及训练时间,最后结果以“平均值 $\pm$ 标准差”的形式呈现。

表 2 数据集信息

Tab. 2 Datasets information

数据集	样本数量	维度
Sonar	208	60
Heart-statlog	270	13
House-votes	435	16
Breast	683	9
Australian	690	15
Diabetes	768	8

3.2 实验结果

通过上述的实验设置,经过不同的参数组训练,实现了模型搭建,对所有模型以及数据集分别进行 10 次实验,结果见表 3。

通过上述比较可以发现,本文提出的两个模型在以上数据集中表现效果较好。与传统的 SVM 和 LSSVM 相比,其最优间隔分布的引入有效地提高了数据的分类效果;与 ODM 以及 LDM 相比,也有不错的表现。实验充分证明了基于最优间隔分布的最小二乘支持向量机在分类预测上的优越性。

在实验时也记录了每次运行的训练时间,结果见表 4。

表 3 不同模型的平均准确率结果

Tab. 3 Average accuracy results of different models

%

数据集	SVM	LSSVM	LDM	ODM	LSSVM_rv	LSSVM_rm_rv
Sonar	85.24 $\pm$ 5.19	89.76 $\pm$ 3.85	86.19 $\pm$ 3.81	86.67 $\pm$ 4.01	89.52 $\pm$ 3.72	90.48 $\pm$ 3.69
Heart-statlog	84.44 $\pm$ 4.70	84.26 $\pm$ 4.16	84.63 $\pm$ 2.04	83.52 $\pm$ 5.00	84.44 $\pm$ 3.81	84.26 $\pm$ 3.82
House-votes	95.98 $\pm$ 1.65	96.21 $\pm$ 1.99	96.32 $\pm$ 2.23	93.79 $\pm$ 2.07	97.01 $\pm$ 1.72	97.01 $\pm$ 1.72
Breast	94.09 $\pm$ 1.55	96.42 $\pm$ 1.48	94.53 $\pm$ 2.12	96.57 $\pm$ 1.60	96.79 $\pm$ 1.14	97.23 $\pm$ 1.34
Australian	85.80 $\pm$ 2.08	86.74 $\pm$ 2.00	86.38 $\pm$ 2.42	86.67 $\pm$ 2.36	86.38 $\pm$ 2.77	86.59 $\pm$ 2.23
Diabetes	77.53 $\pm$ 2.50	77.66 $\pm$ 2.76	76.88 $\pm$ 2.47	74.81 $\pm$ 2.53	77.73 $\pm$ 2.29	77.60 $\pm$ 2.57

表 4 不同模型的训练时间结果

Tab. 4 Training time results of different models

s

数据集	SVM	LSSVM	LDM	ODM	LSSVM_rv	LSSVM_rm_rv
Sonar	0.001 6 $\pm$ 0.000 5	0.008 9 $\pm$ 0.000 8	0.157 6 $\pm$ 0.128 0	0.010 8 $\pm$ 0.000 9	0.019 0 $\pm$ 0.001 2	0.016 7 $\pm$ 0.001 3
Heart-statlog	0.001 7 $\pm$ 0.000 5	0.017 6 $\pm$ 0.005 4	0.370 3 $\pm$ 0.028 4	0.012 8 $\pm$ 0.001 1	0.098 6 $\pm$ 0.175 7	0.038 3 $\pm$ 0.024 7
House-votes	0.001 8 $\pm$ 0.000 6	0.024 6 $\pm$ 0.005 2	0.558 5 $\pm$ 0.049 0	0.815 0 $\pm$ 0.090 2	0.055 7 $\pm$ 0.002 6	0.056 3 $\pm$ 0.006 8
Breast	0.003 5 $\pm$ 0.001 0	0.130 6 $\pm$ 0.220 3	0.959 0 $\pm$ 0.099 3	1.546 1 $\pm$ 0.034 3	0.108 0 $\pm$ 0.006 4	0.111 6 $\pm$ 0.011 0
Australian	1.932 9 $\pm$ 0.817 5	0.074 2 $\pm$ 0.078 1	0.696 0 $\pm$ 0.188 4	0.154 5 $\pm$ 0.019 8	0.102 7 $\pm$ 0.004 7	0.114 8 $\pm$ 0.010 0
Diabetes	0.008 3 $\pm$ 0.001 4	0.053 5 $\pm$ 0.006 4	1.017 4 $\pm$ 0.070 7	0.254 6 $\pm$ 0.031 5	0.120 3 $\pm$ 0.008 0	0.127 5 $\pm$ 0.012 9

通过对训练时间的比较可以发现,在 Sonar 和 Heart-statlog 数据集上,ODM 的训练时间较少,这是因为 ODM 与 LDM 的运算速率与样本量有关,而这两个数据集的样本量较小,导致在此数据集上训练较

快;而对比其他数据集,本文两个模型的训练速率均优于 ODM 与 LDM,这是因为与 LDM 以及 ODM 相比,本文模型用 LSSVM 代替 SVM,使模型求解变为求解一组线性方程组,因此模型的训练速率大幅提

高;对比 SVM 以及 LSSVM,由于 SVM 采用的是 Python 中的模块进行训练,因此训练速率较快,而 LSSVM 本身就具有快速求解的优点,本模型的线性方程组相较于 LSSVM 而言比较复杂,因此在训练时间上也有较多的消耗。

以上实验结果证明:本文模型将最优间隔分布与最小二乘支持向量机结合,在准确率方面,比传统 SVM 以及 LSSVM 有所提升;在训练时间方面,用 LSSVM 代替 LDM 和 ODM 中的支持向量机模型,使训练速率得到大幅提升。

4 结 语

结合 LSSVM 易于求解的特点以及概率分布的优点,本文模型在 LSSVM 模型中引入最小化方差的概念,后续也引入了最大化均值的概念加以改进,形成了基于最优间隔分布的最小二乘支持向量机模型,使 LSSVM 模型分类更加合理,分类准确率也得到提升。通过引入对偶问题,将原模型变换为求解一组线性方程组,使模型的训练速率与基于 SVM 的 ODM 相比得到大幅提升,并且也可以达到相同的分类效果。在不同数据集上进行实验,也进一步验证了上述观点,准确率得到提升;训练速率方面,与 ODM 相比,也获得大幅提升,有效节省了大量的时间成本,使其在实际工作中可以更加高效快速。

参考文献:

- [1] 王万良. 人工智能导论[M]. 5版. 北京:高等教育出版社,2020.
- [2] 高蕊. 随机森林算法改进及其在医疗诊断系统中的应用[D]. 南京:南京邮电大学,2020.
- [3] 熊传文. 基于机器学习的网络支付欺诈交易动态识别模型的构建[J]. 惠州学院学报,2021,41(6):80-84.
- [4] 李家樑. 基于随机森林算法的移动电话骚扰号码识别策略研究[J]. 通讯世界,2019,26(8):93-94.
- [5] VALUEVA M V, NAGORNOV N N, LYAKHOV P A, et al. Application of the residue number system to reduce hard-ware costs of the convolutional neural network implementation[J]. Mathematics and computers in simulation, 2020, 177:232-243.
- [6] YUAN Z W, ZHANG J. Feature extraction and image retrieval based on Alex-Net[C]//IEEE. 2016 50th Asilomar Conference on Signals, Systems and Computers. New York: IEEE, 2016: 7869105.
- [7] TEKCHANDANI H, VERMA S, LONDHE N D, et al. Severity assessment of cervical lymph nodes using modified VGG-Net, and squeeze and excitation concept[C]//IEEE. 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC). New York: IEEE, 2021: 709-714.
- [8] LU Y, MA C, LU Y, et al. A mean field analysis of deep ResNet and beyond: towards provably optimization via overparameterization from depth[C]//PMLR. International Conference on Machine Learning. New York: PMLR, 2020: 6426-6436.
- [9] CORTES C, VAPNIK V. Support vector networks[J]. Machine learning, 1995, 20(3): 273-297.
- [10] PLATT J. Fast training of support vector machines using sequential minimal optimization[M]//Schölkopf B, Burges C J C. Advances in Kernel Methods-Support Vector Learning. Cambridge: MIT Press, 1998: 185-208.
- [11] HSIEH C, CHANG K W, LIN C J, et al. A dual coordinate descent method for large-scale linear SVM[C]//ACM. Proceedings of the 25th international conference on Machine learning. New York: ACM, 2016.
- [12] SUYKENS J, VANDEWALIE J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9(3): 293-300.
- [13] ZHANG T, ZHOU Z H. Optimal margin distribution machine[J]. IEEE Transactions on knowledge and data engineering, 2016, 32(6): 1143-1156.
- [14] GAO W, ZHOU Z H. On the doubt about margin explanation of boosting[J]. Artificial Intelligence, 2013, 203: 1-18.
- [15] ZHANG T, ZHOU Z H. Large margin distribution machine[C]//ACM. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014.
- [16] SHERMAN J, MORRISON W J. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix[J]. The annals of mathematical statistics, 1950, 21(1): 124-127.
- [17] BARTLETT M S. An inverse matrix adjustment arising in discriminant analysis[J]. The annals of mathematical statistics, 1951, 22(1): 107-111.

责任编辑: 郎婧