axioms

MDPI

Article

Robust Fisher-Regularized Twin Extreme Learning Machine with Capped L_1-Norm for Classification

Zhenxia Xue [1,2,*] and Linchao Cai [1]

[1] School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China
[2] The Key Laboratory of Intelligent Information and Big Data Processing of NingXia Province, North Minzu University, Yinchuan 750021, China
* Correspondence: 2019015@num.edu.cn

Abstract: Twin extreme learning machine (TELM) is a classical and high-efficiency classifier. However, it neglects the statistical knowledge hidden inside the data. In this paper, in order to make full use of statistical information from sample data, we first come up with a Fisher-regularized twin extreme learning machine (FTELM) by applying Fisher regularization into TELM learning framework. This strategy not only inherits the advantages of TELM, but also minimizes the within-class divergence of samples. Further, in an effort to further boost the anti-noise ability of FTELM method, we propose a new capped L_1-norm FTELM (CL_1-FTELM) by introducing capped L_1-norm in FTELM to dwindle the influence of abnormal points, and CL_1-FTELM improves the robust performance of our FTELM. Then, for the proposed FTELM method, we utilize an efficient successive overrelaxation algorithm to solve the corresponding optimization problem. For the proposed CL_1-FTELM, an iterative method is designed to solve the corresponding optimization based on re-weighted technique. Meanwhile, the convergence and local optimality of CL_1-FTELM are proved theoretically. Finally, numerical experiments on manual and UCI datasets show that the proposed methods achieve better classification effects than the state-of-the-art methods in most cases, which demonstrates the effectiveness and stability of the proposed methods.

Keywords: twin extreme learning machine; within-class scatter; fisher regularization; capped L_1-norm; robustness

Citation: Xue, Z.; Cai, L. Robust Fisher-Regularized Twin Extreme Learning Machine with Capped L_1-Norm for Classification. *Axioms* **2023**, *12*, 717. https://doi.org/10.3390/axioms12070717

Academic Editor: Miljan Kovačević

Received: 27 June 2023
Revised: 15 July 2023
Accepted: 17 July 2023
Published: 24 July 2023

1. Introduction

Extreme learning machine [1,2], as a remarkable single hidden layer feed-forward neural networks (SLFNs) [3] training method, has been widely studied and applied in many fields such as efficient modeling [4], fashion retailing forecasting [5], fingerprint matching [6], metagenomic taxonomic classification [7], online sequential learning [8], and feature selection [9]. The weights of the input layer and hidden layer offsets are randomly generated. The output weight of the network is calculated effectively by minimizing the training error and the norm of the output weight. In addition, many researchers have tried to extend the extreme learning machine model to the support vector machine (SVM) learning framework to solve the classification problem [10]. Frenay et al. [11] found that the transformation performed by the first layer of ELM can be viewed as a kernel that can be plugged into SVM. Due to solving the support vector machine (SVM) type of optimization method that can be utilized to resolve the ELM model, an extreme learning machine based on the optimization method (OPTELM) was proposed in [12]. For binary classification problems, traditional ELM needs to compute all the sample points of training data at the same time in the training stage, which is time-consuming. The singe hyperplane was trained to perform the classification task in the traditional ELM, which enormously restricts its application prospect and the direction of evolution. Jayadeva et al. [13] proposed twin SVM (TWSVM), which is a famous non-parallel hyper-plane classification algorithm for

binary classification. Inspired by TWSVM, Wan et al. [14] proposed the twin extreme learning machine (TELM). Compared with ELM, TELM trains two non-parallel hyperplanes for classification tasks by solving two smaller quadratic programming problems (QPPs). Compared with TWSVM, TELM's optimization problem has fewer constraints, so the training speed is faster and the application prospect is broader. In recent years, researchers have made many improvements to TELM, such as sparse twin extreme learning machine [15], robust twin extreme learning machine [16], time efficient varient of twin extreme learning machine [17], and a generalized adaptive robust distance metric driven smooth regularization learning framework [18], etc.

Although the above ELM-based algorithm has a good classification effect, the statistical knowledge from the data itself is ignored. However, the knowledge of mathematical statistics from the data is very important to construct an efficient classifier. Fisher discriminant analysis (FDA) is an effective discriminant tool by minimizing the intra-class divergence while keeping the inter-class divergence of the data constant. From the above discussion, it can be known that it is necessary to reconstruct a new classification model by combining the characteristics of ELM model and FDA. In recent years, Ma et al. [19] have successfully combined them and proposed a Fisher-regularized extreme learning machine (Fisher-ELM), which not only has the advantages of efficient solution of ELM but also fully considers the statistical knowledge of the data.

Although the above models have good classification performance, most of them consider the L_2-norm. When the data contains noise or outliers, they can not deal with noise and outliers well, which degrades the classification performance of the model. In recent years, researchers have tried to introduce the L_1-norm into various models [20–23] to reduce the impact of outliers. This studies have shown that the L_1-norm was able to reduce the effect of outliers to some extent. However, it was still unsatisfactory when the data contains a large number of outliers. Recently, researchers have introduced the idea of truncation into the L_1-norm, constructed a new capped L_1-norm, and applied it to various models [24–26]. Many studies [27,28] show that the capped L_1-norm not only inherits the advantages of the L_1-norm, but also is bounded. So it is more robust and it approaches the L_0-norm to some degree. For instance, by applying the capped L_1-norm to the twin SVM, Wang et al. [29] proposed a new robust twin support vector machine (CL_1-TWSVM). Based on twin support vector machine with privileged information [30] (TWSVMPI), a new robust TWSVMPI [31] is proposed by replacing the L_2-norm with capped L_1-norm. The new model further improves the anti-noise ability of the pattern.

In order to utilize the advantanges of the twin extreme learning machine and FDA, we first put forward to a novel classifier named Fisher-regularized twin extreme learning machine (FTELM). Also considering the instability of the L_2-norm for the outliers, we introduce the capped L_1-norm into the FTELM model and propose a more robust capped L_1-norm FTELM (CL_1-FTELM) model.

The main contributions of this paper are as follows:

(1) Based on twin extreme learning machine and Fisher-regularization extreme learning machine (FELM), a new Fisher-regularized twin extreme learning machine (FTELM) is proposed. FTELM minimizes intra-class divergence while fixing the inter-class divergence of samples. FTELM takes full account of the statistical information of the sample data, and the training speed is faster than FELM.

(2) Considering the instability of L_2-norm and Hinge loss used by FTELM, we introduce capped L_1-norm instead of them and propose a new capped L_1-norm FTELM model. CL_1-FTELM uses the capped L_1-norm to reduce the influence of noise points, and at the same time utilizes Fisher regularization to consider the statistical knowledge of the data.

(3) Two algorithms are designed by utilizing the successive overrelaxation (SOR) [32] technique and the re-weighted technique [27] to solve the optimization problems of the proposed FTELM and CL_1-FTELM, respectively.

(4) Two theorems about convergence and local optimality of CL_1-FTELM are proved.

The organizational structure of this paper is as follows. In Section 2, we briefly review related work. In Section 3, we describe the FTELM model in detail. The robust capped L_1-norm FTELM learning framework along with related theoretical proofs are described in detail in Section 4. In Section 5, we describes numerical experiments on artificial and benchmark datasets. We summarize this paper in Section 6.

2. Related Work

In this section, we first define some concepts of symbols needed for this paper, and then we briefly review Fisher regularization, Fisher-ELM, TELM and successive overrelaxation algorithm.

2.1. The Concept of Symbols

e is a vector whose components are all ones, an identity matrix is represented by **I**, and a matrix(vector) of zeros is represented by **0**. Then, $\|\cdot\|_2$ is the L_2 norm, and $\|\cdot\|_F$ stands for the Frobenius norm.

A binary classification problem in Euclidean space (R^d) can be formulated in the following form:

$$T = \{x_i, y_i\} \in (\mathcal{X}, \mathcal{Y}), (i = 1, ..., m) \tag{1}$$

where $x_i \in \mathcal{X} \subset R^d$ is expressed as an input sample in a d-dimensional Euclidean space. Similarly, $y_i \in \mathcal{Y} = \{-1, +1\}$ is represented as an output label corresponding to an input instance x_i. In addition, m_1 and m_2 represent the number of sample data of the positive class and negative class, respectively, and $m = m_1 + m_2$.

2.2. Fisher Regularization

Fisher regularization has the following form:

$$\|f\|_F^2 = \mathbf{f}^T \mathbf{N} \mathbf{f} = \sum_{i \in \mathcal{I}_+} \left(f(x_i) - \bar{f}_+\right)^2 + \sum_{i \in \mathcal{I}_-} \left(f(x_i) - \bar{f}_-\right)^2 \tag{2}$$

where $\mathbf{f} = [f(x_1), f(x_2), \ldots, f(x_m)]^T$, $\mathbf{N} = \mathbf{I} - \mathbf{G}$, $\mathbf{I} \in R^{m \times m}$ is the identity matrix and $\mathbf{G}$ is the matrix with the elements:

$$\mathbf{G}_{ij} = \begin{cases} \frac{1}{m_1}, for\ i, j \in \mathcal{I}_+ \\ \frac{1}{m_2}, for\ i, j \in \mathcal{I}_- \\ 0, \quad otherwise \end{cases} \tag{3}$$

where $\mathcal{I}_\pm$ are the index sets of positive and negative training data, $m_1 = |\mathcal{I}_+|$, $m_2 = |\mathcal{I}_-|$. The average value of $\mathbf{f}(\mathbf{x})$ over the positive sample set is expressed as $\bar{f}_+$, the average value of $\mathbf{f}(\mathbf{x})$ over the negative sample set is expressed as $\bar{f}_-$. From Equation (2), we can see that the meaning of the Fisher regularization is the intra-class divergence of the data.

The proof of Formula (2) is as follows:

$$\begin{aligned} &\sum_{i \in \mathcal{I}_+} \left(f(x_i) - \bar{f}_+\right)^2 = \sum_{i \in \mathcal{I}_+} \left(f^2(x_i) - 2 \cdot f(x_i) \cdot \bar{f}_+ + \bar{f}_+^2\right) = \sum_{i \in \mathcal{I}+} f^2(x_i) - m_1 \cdot \bar{f}_+^2 \\ &= \mathbf{f}_+^\mathbf{T} \cdot \mathbf{f}_+ - \frac{1}{m_1} \cdot \left(\mathbf{f}_+^\mathbf{T} \cdot \mathbf{e} \cdot \mathbf{e}^\mathbf{T} \cdot \mathbf{f}_+^\mathbf{T}\right) = \mathbf{f}_+^\mathbf{T} \cdot \mathbf{I}_+ \mathbf{f}_+ - \mathbf{f}_+^\mathbf{T} \cdot \mathbf{M}_+ \cdot \mathbf{f}_+^\mathbf{T} \\ &= \mathbf{f}_+^\mathbf{T} \cdot (\mathbf{I}_+ - \mathbf{M}_+) \cdot \mathbf{f}_+ = \mathbf{f}_+^\mathbf{T} \cdot (\mathbf{N_1}) \cdot \mathbf{f}_+ \end{aligned} \tag{4}$$

where $\mathbf{e} = [1, \ldots, 1]^T$ is a vector of m_1 dimensions, $\mathbf{f}_+ = (f(x_1), f(x_2), \ldots, f(x_i), \ldots, f(x_{m_1}))$, $i \in \mathcal{I}_+$, $\mathbf{I}_+ \in R^{m_1 \times m_1}$ is the identity matrix. $\mathbf{M}_+ \in R^{m_1 \times m_1}$, and all the entries in the matrix $\mathbf{M}_+$ are $\frac{1}{m_1}$.

Similarly, it can be obtained:

$$\sum_{j\in\mathcal{I}_-}\left(f(x_i)-\bar{f}_-\right)^2=\mathbf{f}_-^{\mathbf{T}}\cdot(\mathbf{I}_--\mathbf{M}_-)\cdot\mathbf{f}_-=\mathbf{f}_-^{\mathbf{T}}\cdot(\mathbf{N_2})\cdot\mathbf{f}_- \tag{5}$$

where $\mathbf{f}_-=(f(x_1),f(x_2),\ldots,f(x_i),\ldots,f(x_{m_2})), i\in\mathcal{I}_-$, $\mathbf{I}_-\in R^{m_2\times m_2}$ is the identity matrix. $\mathbf{M}_-\in R^{m_2\times m_2}$, and all the entries in the matrix $\mathbf{M}_-$ are $\frac{1}{m_2}$.

Combining Equations (4) and (5), we can get another form of Equation (2):

$$\begin{aligned}&\mathbf{f}_+^{\mathbf{T}}\cdot(\mathbf{I}_+-\mathbf{M}_+)\cdot\mathbf{f}_++\mathbf{f}_-^{\mathbf{T}}\cdot(\mathbf{I}_--\mathbf{M}_-)\cdot\mathbf{f}_-\\&=(\mathbf{f}_+,\mathbf{f}_-)^T\cdot\left[\mathbf{I}-\begin{bmatrix}\mathbf{M}_+ & \mathbf{0_1}\\ \mathbf{0_2} & \mathbf{M}_-\end{bmatrix}\right]\cdot(\mathbf{f}_+,\mathbf{f}_-)\\&=\mathbf{f}^{\mathbf{T}}\cdot(\mathbf{I}-\mathbf{G})\cdot\mathbf{f}=\mathbf{f}^{\mathbf{T}}\cdot\mathbf{N}\cdot\mathbf{f}\end{aligned} \tag{6}$$

where $\mathbf{0_1}\in\mathbf{0}^{m_1\times m_2}$, $\mathbf{0_2}\in\mathbf{0}^{m_2\times m_1}$, $G=\begin{bmatrix}\mathbf{M}_+ & \mathbf{0_1}\\ \mathbf{0_2} & \mathbf{M}_-\end{bmatrix}$.

2.3. Fisher-Regularized Extreme Learning Machine

The primal problem of Fisher-regularized extreme learning machine (FELM) is as follows:

$$\begin{aligned}\min_{\alpha,\xi}\quad&\frac{1}{2}\beta^T\cdot\beta+C_1\cdot\mathbf{e}^{\mathbf{T}}\cdot\xi+\frac{1}{2}C_2\cdot\alpha^T\cdot\mathbf{K}_{ELM}\cdot\mathbf{N}\cdot\mathbf{K}_{ELM}.\alpha\\ s.t.\quad&\mathbf{Y}\cdot\mathbf{H}\cdot\beta\geq\mathbf{e}-\xi\\&\xi\geq\mathbf{0}\end{aligned} \tag{7}$$

According to the representer theorem $\beta=\sum_{i=1}^{m}\alpha_i\mathbf{h}(x_i)=\mathbf{H}^T\alpha$, problem (7) can be written as problem (8):

$$\begin{aligned}\min_{\alpha,\xi}\quad&\frac{1}{2}\alpha^T\cdot\mathbf{K}_{ELM}\cdot\alpha+C_1\cdot\mathbf{e}^{\mathbf{T}}\cdot\xi+\frac{1}{2}C_2\cdot\alpha^T\cdot\mathbf{K}_{ELM}\cdot\mathbf{N}\cdot\mathbf{K}_{ELM}.\alpha\\ s.t.\quad&\mathbf{Y}\cdot\mathbf{K_{ELM}}\cdot\alpha\geq\mathbf{e}-\xi\\&\xi\geq\mathbf{0}\end{aligned} \tag{8}$$

where $\mathbf{K}_{ELM}\in R^{m\times m}$ is a Gram matrix with elements $k_{ELM}(x_i,x_j)$, $k_{ELM}(x_i,x)=\mathbf{h}(x)^T\cdot\mathbf{h}(x_i)$, $\mathbf{h}(x)$ denotes the output of some hidden node, $\mathbf{Y}\in R^{m\times m}$ is a diagonal matrix with elements y_i, C_1, C_2 are the regularization parameters, and ξ is a nonnegative slack vector.

According to the optimization theory, the dual form of the problem (8) can be obtained as follows:

$$\begin{aligned}&\min_{\theta}\frac{1}{2}\theta^T\cdot\mathbf{Q}\cdot\theta-\mathbf{e}^T\cdot\theta\\&s.t.\quad\mathbf{0}\leq\theta\leq C_1\cdot\mathbf{e}\end{aligned} \tag{9}$$

where $\mathbf{Q}=\mathbf{Y}\cdot\left((\mathbf{I}+C_2\cdot\mathbf{N}\cdot\mathbf{K}_{ELM})^{-1}\right)^T\cdot\mathbf{K}_{ELM}\cdot\mathbf{Y}$.

The decision function of Fisher-regularized extreme learning machine is:

$$f(x)=sign\left(\sum_{i=1}^{m}\alpha_i\cdot k_{ELM}(x_i,x)\right) \tag{10}$$

2.4. Twin Extreme Learning Machine

Similar to the form of TWSVM [13], the primal problem of TELM [14] can be expressed in the following:

$$\begin{aligned} \textbf{Primal} \quad \textbf{TELM}_\mathbf{1}: \quad & \min_{\beta_1} \frac{1}{2}\|\mathbf{H}_1 \cdot \beta_1\|_2^2 + C_1 \cdot \mathbf{e_2^T} \cdot \xi \\ & s.t. \quad -\mathbf{H_2} \cdot \beta_1 \geq \mathbf{e_2} - \xi \\ & \xi \geq \mathbf{0} \end{aligned} \tag{11}$$

$$\begin{aligned} \textbf{Primal} \quad \textbf{TELM}_\mathbf{2}: \quad & \min_{\beta_2} \frac{1}{2}\|\mathbf{H}_2 \cdot \beta_2\|_2^2 + C_2 \cdot \mathbf{e_1^T} \cdot \eta \\ & s.t. \quad \mathbf{H_1} \cdot \beta_2 \geq \mathbf{e_1} - \eta \\ & \eta \geq \mathbf{0} \end{aligned} \tag{12}$$

where $\mathbf{H}_1$ and $\mathbf{H}_2$ respresent the outputs of the hidden layer for positive and negative samples, ξ and η represent the slack vectors, $\mathbf{0}$ is a zero vector, C_1, $C_2 \geq 0$ are penalty parameters, $\mathbf{e}_1 \in R^{m_1}$ and $\mathbf{e}_2 \in R^{m_2}$ are vectors of ones.

By introducing Lagrange multipliers α and ϑ, the dual problem of (11) and (12) can be written as follows:

$$\begin{aligned} \textbf{Dual} \quad \textbf{TELM}_\mathbf{1}: \quad & \min_{\alpha} \frac{1}{2}\alpha^T \cdot \mathbf{H}_2\left(\mathbf{H}_1^T \cdot \mathbf{H_1}\right)^{-1} \cdot \mathbf{H}_2^T \cdot \alpha - \mathbf{e_2^T} \cdot \alpha \\ & s.t. \quad \mathbf{0} \leq \alpha \leq C_1 \cdot \mathbf{e}_2 \end{aligned} \tag{13}$$

$$\begin{aligned} \textbf{Dual} \quad \textbf{TELM}_\mathbf{2}: \quad & \min_{\vartheta} \frac{1}{2}\vartheta^T \cdot \mathbf{H}_1\left(\mathbf{H}_2^T \cdot \mathbf{H_2}\right)^{-1} \cdot \mathbf{H}_1^T \cdot \vartheta - \mathbf{e_1^T} \cdot \vartheta \\ & s.t. \quad \mathbf{0} \leq \vartheta \leq C_2 \cdot \mathbf{e}_1 \end{aligned} \tag{14}$$

The solution of (13) and (14) are as follows:

$$\beta_1 = -\left(\mathbf{H_1^T} \cdot \mathbf{H_1} + \triangle_1 \mathbf{I}\right)^{-1} \cdot \mathbf{H}_2^T \cdot \alpha \tag{15}$$

$$\beta_2 = -\left(\mathbf{H_2^T} \cdot \mathbf{H_2} + \triangle_2 \mathbf{I}\right)^{-1} \cdot \mathbf{H}_1^T \cdot \vartheta \tag{16}$$

where $\triangle_1$ and $\triangle_2$ are two small positive constants and $\mathbf{I}$ is an identity matrix. The decision function of twin extreme learning machine is:

$$f(x) = arg \min_{k=1,2} d_k(x) = arg \min_{k=1,2} \left|\beta_k^T \cdot \mathbf{h}(\mathbf{x})\right| \tag{17}$$

2.5. Successive Overrelaxation Algorithm

The successive overrelaxation algorithm [32] mainly aims at the following optimization problems:

$$\begin{aligned} & \min_{\mu} \frac{1}{2}\left\|\mathbf{H}^T\mu\right\|_2^2 - \mathbf{e}^T\mu \\ & s.t. \quad \mu \in S = \{\mu | \mathbf{0} \leq \mu \leq C\mathbf{e}\} \end{aligned} \tag{18}$$

Let $\mathbf{H}\mathbf{H}^T = \mathbf{L} + \mathbf{E} + \mathbf{L}^T$, the strictly lower triangular matrix of the matrix $\mathbf{H}\mathbf{H}^T$ is $\mathbf{L}$, and the diagonal elements of the matrix $\mathbf{H}\mathbf{H}^T$ form the diagonal matrix $\mathbf{E}$.

The gradient projection optimality condition is the necessary and sufficient optimality condition for Equation (18):

$$\mu = \left(\mu - \pi\mathbf{E}^{-1}\left(\mathbf{H}\mathbf{H}^T\mu - \mathbf{e}\right)\right)_{\#}, \pi \geq 0$$

where the 2-norm projection onto the feasible region of Equation (18) is denoted by $(\cdot)_{\#}$, that is:

$$((\mu)_{\#})_i = \begin{cases} 0, if \quad \mu_i \leq 0, i = 1, 2, \ldots, m \\ \mu_i, if \quad 0 < \mu_i < C, i = 1, 2, \ldots, m \\ C, if \quad \mu_i \geq C, i = 1, 2, \ldots, m \end{cases} \tag{19}$$

The matrix $\mathbf{H}\mathbf{H}^T$ is expressed in the following form:

$$\begin{aligned} &\mathbf{H}\mathbf{H}^T = \pi^{-1}\mathbf{E}(\mathbf{B} + \mathbf{C}) \\ &s.t.\ \mathbf{B} - \mathbf{C}\ is\ positive\ definite \end{aligned} \tag{20}$$

Here:

$$\mathbf{B} = \left(\mathbf{I} + \pi\mathbf{E}^{-1}\mathbf{L}\right), \mathbf{C} = \left((\pi - 1)\mathbf{I} + \pi\mathbf{E}^{-1}\mathbf{L}^T\right), 0 < \pi\ < 2 \tag{21}$$

According to the [33], the matrix splitting algorithm is as follows:

$$\mu^{i+1} = \left(\mu^{i+1} - \mathbf{B}\mu^{i+1} - \mathbf{C}\mu^i + \pi\mathbf{E}^{-1}\mathbf{e}\right)_{\#} \tag{22}$$

Substituting Equation (21) into Equation (22), it can be obtained:

$$\mu^{i+1} = \left(\mu^i - \pi\mathbf{E}^{-1}\left(\mathbf{H}\mathbf{H}^T\mu^i - \mathbf{e} + \mathbf{L}\left(\mu^{i+1} - \mu^i\right)\right)\right)_{\#} \tag{23}$$

3. Fisher-Regularized Twin Extreme Learning Machine

3.1. Model Formulation

As mentioned above, TELM solves two smaller QPPs, which can get the solution quickly. However, it ignores the prior statistical knowledge from data. FELM minimizes the within-class scatter while controlling the between-class scatter of samples, but FELM needs to solve a large-scale quadratic programming problems which is time-consuming. In this paper, by combining the advantages of FELM and TELM, we first propose the Fisher-regularized twin extreme learning machine (FTELM) by introducing the Fisher regularization into the TELM feature space. FTELM only needs to solve two smaller quadratic programming problems and meanwhile utilizes the prior statistical knowledge from data. The pair of FTELM primal problems is as follows:

$$\begin{aligned} \textbf{Primal} \quad \textbf{FTELM}_\textbf{1}: \quad \min_{\beta_1, \xi} \quad & \frac{1}{2}\|\mathbf{H}_1 \cdot \beta_1\|^2 + C_1 \cdot \mathbf{e}_2^T \cdot \xi + \frac{C_2}{2} \cdot \mathbf{f}_1(x)^T \cdot \mathbf{N}_1 \cdot \mathbf{f}_1(x) \\ & s.t. \quad -\mathbf{H}_2 \cdot \beta_1 + \xi \geq \mathbf{e}_2 \\ & \xi \geq \mathbf{0} \end{aligned} \tag{24}$$

$$\begin{aligned} \textbf{Primal} \quad \textbf{FTELM}_\textbf{2}: \quad \min_{\beta_2, \eta} \quad & \frac{1}{2}\|\mathbf{H}_2 \cdot \beta_2\|^2 + C_3 \cdot \mathbf{e}_1^T \cdot \eta + \frac{C_4}{2} \cdot \mathbf{f}_2(x)^T \cdot \mathbf{N}_2 \cdot \mathbf{f}_2(x) \\ & s.t. \quad \mathbf{H}_1 \cdot \beta_2 + \eta \geq \mathbf{e}_1 \\ & \eta \geq \mathbf{0} \end{aligned} \tag{25}$$

From the Equations (4) and (5), we can know that $\mathbf{N}_1 = \mathbf{I}_+ - \mathbf{M}_+$ and $\mathbf{N}_2 = \mathbf{I}_- - \mathbf{M}_-$, $C_1, C_2, C_3, C_4 > 0$ are regularization parameters, ξ and η are the error vectors, and all the elements in vectors $\mathbf{e}_1 \in R^{m_1}$ and $\mathbf{e}_2 \in R^{m_2}$ are one. FTELM first inherits the advantage of the classical twin extreme learning machine, which computes two non-parallel hyperplanes to solve the classification problem. Secondly, FTELM takes full account of the statistical information of the samples and further improves the classification accuracy of the classifier. The optimization objective function in (24) of FTELM mainly has three terms: minimizing the distance from the positive class sample points to the positive class hyperplane, minimizing empirical loss, and minimizing the intra-class divergence from the samples.

The constraint condition in (24) of the optimization objective function is that the distance between the negative class sample points and the positive class hyperplane is greater than or equal to one. In a word, FTELM makes the positive class sample points closer to the positive class hyperplane, and the negative class sample points far away from the positive class hyperplane. At the same time, the positive class sample points are more concentrated in the center of the positive class sample points. There is a similar explanation for the model (25).

According to the representer theorem $\beta = \sum_{i=1}^{m} \alpha_i \mathbf{h}(x_i) = \mathbf{H}^T \alpha$, then $\beta_1 = \mathbf{H}_1^T \cdot \alpha_1$ and $\beta_2 = \mathbf{H}_2^T \cdot \alpha_2$. We know that $\mathbf{f} = \mathbf{H} \cdot \beta$. Therefore, the problem (24) and (25) can be written in the following forms:

$$\begin{aligned} \min_{\alpha_1, \xi} \quad & \frac{1}{2}\alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 + C_1 \cdot \mathbf{e}_2^T \cdot \xi + \frac{C_2}{2}\alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 \\ s.t. \quad & -\mathbf{H_2} \cdot \mathbf{H}_1^T \cdot \alpha_1 + \xi \geq \mathbf{e_2} \\ & \xi \geq \mathbf{0} \end{aligned} \tag{26}$$

$$\begin{aligned} \min_{\alpha_2, \eta} \quad & \frac{1}{2}\alpha_2^T \cdot \mathbf{K}_{ELM2} \cdot \mathbf{K}_{ELM2} \cdot \alpha_2 + C_3 \cdot \mathbf{e}_1^T \cdot \eta + \frac{C_4}{2}\alpha_2^T \cdot \mathbf{K}_{ELM2} \cdot \mathbf{N}_2 \cdot \mathbf{K}_{ELM2} \cdot \alpha_2 \\ s.t. \quad & \mathbf{H_1} \cdot \mathbf{H}_2^T \cdot \alpha_2 + \eta \geq \mathbf{e_1} \\ & \eta \geq \mathbf{0} \end{aligned} \tag{27}$$

where $\mathbf{K}_{ELM1} = \mathbf{H}_1 \cdot \mathbf{H_1^T}$ and $\mathbf{K}_{ELM2} = \mathbf{H}_2 \cdot \mathbf{H_2^T}$ are Gram matrices.

3.2. Model Solution

Introducing Lagrange multipliers $\theta = (\theta_1, \ldots, \theta_{m_2})^T$ and $\vartheta = (\vartheta_1, \ldots, \vartheta_{m_2})^T$, the Lagrange function of (26) can be written as follows:

$$\begin{aligned} \mathcal{L}(\alpha_1, \xi, \theta, \vartheta) = & \frac{1}{2}\alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot (\mathbf{I}_1 + C_2 \cdot \mathbf{N}_1) \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 + C_1 \cdot \mathbf{e}_2^T \cdot \xi \\ & - \theta^T \cdot \left(-\mathbf{H}_2 \cdot \mathbf{H}_1^T \cdot \alpha_1 + \xi - \mathbf{e}_2\right) - \vartheta^T \cdot \xi \end{aligned} \tag{28}$$

According to the KKT conditions, we get:

$$\frac{\partial \mathcal{L}}{\partial \alpha_1} = \mathbf{K}_{ELM1} \cdot (\mathbf{I}_1 + C_2 \cdot \mathbf{N}_1) \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 + \mathbf{H}_1 \cdot \mathbf{H_2^T} \cdot \theta = \mathbf{0} \tag{29}$$

$$\frac{\partial \mathcal{L}}{\partial \xi} = C_1 \cdot \mathbf{e}_2 - \theta - \vartheta = \mathbf{0} \tag{30}$$

$$\theta^T \cdot \left(-\mathbf{H}_2 \cdot \mathbf{H}_1^T \cdot \alpha_1 + \xi - \mathbf{e}_2\right) = \mathbf{0} \tag{31}$$

$$\vartheta^T \cdot \xi = \mathbf{0} \tag{32}$$

$$\theta \geq \mathbf{0} \tag{33}$$

$$\vartheta \geq \mathbf{0} \tag{34}$$

From (29) and (30), we can get:

$$\alpha_1^* = -(\mathbf{K}_{ELM1} \cdot (\mathbf{I}_1 + C_2 \cdot \mathbf{N}_1) \cdot \mathbf{K}_{ELM1})^{-1} \cdot \mathbf{H}_1 \cdot \mathbf{H}_2^T \cdot \theta \tag{35}$$

$$\mathbf{0} \leq \theta \leq C_1 \cdot \mathbf{e}_2 \tag{36}$$

By substituting (29)–(34) into (28), the dual optimization problem for (26) can be written in the following form:

$$\begin{aligned} \textbf{Dual} \quad \textbf{FTELM}_\textbf{1}: \min_{\theta} \quad & \frac{1}{2}\theta^T \cdot \mathbf{Q_1} \cdot \theta - \mathbf{e_2^T} \cdot \theta \\ s.t. \quad & \mathbf{0} \leq \theta \leq C_1 \cdot \mathbf{e}_2 \end{aligned} \tag{37}$$

Here $\mathbf{Q}_1 = \mathbf{H}_2 \cdot \mathbf{H}_1^T \cdot (\mathbf{K}_{ELM1} \cdot (\mathbf{I}_1 + C_2 \cdot \mathbf{N}_1) \cdot \mathbf{K}_{ELM1})^{-1} \cdot \mathbf{H}_1 \cdot \mathbf{H}_2^T$.
Similarly, we can obtain the dual of (27) as:

$$\begin{aligned} \textbf{Dual} \quad \textbf{FTELM}_\textbf{2}: \min_{\lambda} \quad & \frac{1}{2}\lambda^T \cdot \mathbf{Q_2} \cdot \lambda - \mathbf{e_1^T} \cdot \lambda \\ s.t. \quad & \mathbf{0} \leq \lambda \leq C_3 \cdot \mathbf{e}_1 \end{aligned} \tag{38}$$

Here $\lambda = (\lambda_1, \ldots, \lambda_{m_1})^T$ is the vector of Lagrange multipliers and we can get: $\mathbf{Q}_2 = \mathbf{H}_1 \cdot \mathbf{H}_2^T \cdot (\mathbf{K}_{ELM2} \cdot (\mathbf{I}_2 + C_4 \cdot \mathbf{N}_2) \cdot \mathbf{K}_{ELM2})^{-1} \cdot \mathbf{H}_2 \cdot \mathbf{H}_1^T$.

We use the successive overrelaxation (SOR) [32] technique to solve the convex quadratic optimization problems of (37) and (38) (The SOR-FTELM algorithm is summarized as Algorithm 1). We can get θ and λ. Therefore, we can obtain the solution for problems of (24) and (25) in the following :

$$\beta_1 = -\mathbf{H}_1^T \cdot (\mathbf{K}_{ELM1} \cdot (\mathbf{I}_1 + C_2 \cdot \mathbf{N}_1) \cdot \mathbf{K}_{ELM1} + \delta_1 \cdot \mathbf{I_1})^{-1} \cdot \mathbf{H}_1 \cdot \mathbf{H}_2^T \cdot \theta \tag{39}$$

$$\beta_2 = \mathbf{H}_2^T \cdot (\mathbf{K}_{ELM2} \cdot (\mathbf{I}_2 + C_4 \cdot \mathbf{N}_2) \cdot \mathbf{K}_{ELM2} + \delta_2 \cdot \mathbf{I_2})^{-1} \cdot \mathbf{H}_2 \cdot \mathbf{H}_1^T \cdot \lambda \tag{40}$$

The decision function of FTELM is:

$$f(x) = arg \min_{k=1,2} \left| \beta_k^T \cdot \mathbf{h}(\mathbf{x}) \right| \tag{41}$$

Algorithm 1 The procedure of SOR-FTELM.

Input:
Training set $T = \{x_i, y_i\}_{i=1}^m$, where $x_i \in R^d$, $y_i = \pm 1$, the number of hidden node number L, tolerance ε, regularization parameters C_1, C_2, C_3, C_4.

Output:
β_1, β_2, and the decision function of FTELM.

1: Compute the graph matrix $\mathbf{N}_1$, $\mathbf{N}_2$ by Equations (4) and (5).
2: Choose an activation function such as $G(x) = \frac{1}{1+e^{-x}}$ and compute the hidden layer output matrix $\mathbf{H}_1$, $\mathbf{H}_2$ by $\mathbf{h}(x_i) = G\left(\sum_{j=1}^d \omega_{ji} x_j + b_i\right)$ and compute $\mathbf{K}_{ELM1}$=$\mathbf{H}_1\mathbf{H}_1^T$ and $\mathbf{K}_{ELM2}$=$\mathbf{H}_2\mathbf{H}_2^T$.
3: Choose $t \in (0, 2)$, start with any $\theta^0 \in R^{m_2}$,Having θ^i, compute θ^{i+1} as follows:

$$\theta^{i+1} = \left(\theta^i - t\mathbf{E}_1\left(\mathbf{Q}_1\theta^i - \mathbf{e}_2 + \mathbf{L}_1\left(\theta^{i+1} - \theta^i\right)\right)\right)_\#$$

until $\left|\theta^{i+1} - \theta^i\right| \leq \varepsilon$, where $\mathbf{e}_2$ is a vector of ones of appropriate dimensions. $\mathbf{L}_1 \in R^{m_2 \times m_2}$ is the strictly lower triangular matrix, where $l_{ij} = q_{ij}, i > j$. $\mathbf{E}_1 \in R^{m_2 \times m_2}$ is the diagonal matrix, where $e_{ij} = q_{ij}, i > j$.
Then, given any $\lambda^0 \in R^{m_1}$, Having λ^i, compute λ^{i+1} as follows

$$\lambda^{i+1} = \left(\lambda^i - t\mathbf{E}_2\left(\mathbf{Q}_2\lambda^i - \mathbf{e}_1 + \mathbf{L}_2\left(\lambda^{i+1} - \lambda^i\right)\right)\right)_\#$$

4: Compute the output weights β_1, β_2 using Equations (39) and (40).
5: Construct the following decision functions:

$$f(x) = arg \min_{k=1,2} \left| \beta_k^T \cdot \mathbf{h}(\mathbf{x}) \right|$$

4. Capped L_1-Norm Fisher-Regularized Twin Extreme Learning Machine

4.1. Model Formulation

The Fisher-regularized twin extreme learning machine proposed in the previous section not only inherits the advantages of the twin extreme learning machine but also

makes full use of the statistical information of the samples. However, due to the use of the squared L_2-norm distance and hinge loss function, the Fisher-regularized twin extreme learning machine is not robust enough when noisy points are present, which often enlarges the impact of abnormal values. In order to reduce the influence of outliers and improve the robustness of the FTELM, we propose a capped L_1-norm Fisher twin extreme machine (CL_1-FTELM) by replacing the L_2-norm and hinge loss in the FTELM with capped L_1-norm. The primal CL_1-FTELM is in the following:

Primal CL_1-FTELM$_1$:

$$\begin{aligned}\min_{\alpha_1,\xi} \sum_{i=1}^{m_1} \min\left(\left\|\mathbf{h}^T(x_i)\cdot \mathbf{H_1}^T\cdot\alpha_1\right\|_1, \varepsilon_1\right) + C_1\cdot\sum_{j=1}^{m_2}\min\left(\|\xi_j\|_1,\varepsilon_2\right)\\ +\frac{C_2}{2}\cdot\alpha_1^T\cdot\mathbf{K}_{ELM1}\cdot\mathbf{N}_1\cdot\mathbf{K}_{ELM1}\cdot\alpha_1\\ s.t.\quad -\mathbf{H}_2\cdot\mathbf{H}_1^T\cdot\alpha_1+\xi\geq\mathbf{e}_2\end{aligned} \tag{42}$$

Primal CL_1-FTELM$_2$:

$$\begin{aligned}\min_{\alpha_2,\eta} \sum_{j=1}^{m_2} \min\left(\left\|\mathbf{h}^T(x_j)\cdot \mathbf{H_2}^T\cdot\alpha_2\right\|_1, \varepsilon_3\right) + C_3\cdot\sum_{i=1}^{m_1}\min(\|\eta_i\|_1,\varepsilon_4)\\ +\frac{C_4}{2}\cdot\alpha_2^T\cdot\mathbf{K}_{ELM2}\cdot\mathbf{N}_2\cdot\mathbf{K}_{ELM2}\cdot\alpha_2\\ s.t.\quad \mathbf{H}_1\cdot\mathbf{H}_2^T\cdot\alpha_2+\eta\geq\mathbf{e}_1\end{aligned} \tag{43}$$

where $C_1, C_2, C_3, C_4 > 0$ are regularization parameters, $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ are thresholding parameters.

CL_1-FTELM uses the capped L_1-norm to reduce the influence of noise points, and at the same time utilizes Fisher regularization to consider the statistical knowledge of the data. Based on FTELM, CL_1-FTELM changes the L_2-norm metric and Hinge loss function of the original model to the capped L_1-norm. The capped L_1-norm is bounded and can constrain the impact of noise within a certain range. Therefore, the anti-noise ability of the model can be improved. The optimization objective function in (42) of CL_1-FTELM also contains three terms: minimizing the distance between the positive class sample points and the positive class hyperplane by using capped L_1-norm metric, minimizing empirical loss by using capped L_1-norm loss function, and minimizing the within-class scatter of the samples. The constraints in (42) of CL_1-FTELM are as follows: the distance between the negative class sample points and the positive class hyperplane is greater than or equal to one. In summary, CL_1-FTELM inherits the advantages of FTELM, while further improving the noise immunity of the model by replacing the metric and loss function with the capped L_1-norm. However, the CL_1-FTELM is a non-convex and non-smooth problem. Here, we use the reweighting technique [27] to solve the problem corresponding to the CL_1-FTELM model, which is shown below:

CL_1-FTELM$_1$:

$$\begin{aligned}\min_{\alpha_1,\xi}\quad & \frac{1}{2}\alpha_1^T\cdot\mathbf{K}_{ELM1}\cdot\mathbf{F}\cdot\mathbf{K}_{ELM1}\cdot\alpha_1+\frac{C_1}{2}\cdot\xi^T\cdot\mathbf{D}\cdot\xi\\ &+\frac{C_2}{2}\cdot\alpha_1^T\cdot\mathbf{K}_{ELM1}\cdot\mathbf{N}_1\cdot\mathbf{K}_{ELM1}\cdot\alpha_1\\ &s.t.\quad -\mathbf{H}_2\cdot\mathbf{H}_1^T\cdot\alpha_1+\xi\geq\mathbf{e}_2\end{aligned} \tag{44}$$

where **F** and **D** are two diagonal matrices with i-th and j-th diagonal elements as:

$$f_i=\begin{cases}\frac{1}{\left|\mathbf{h}^T(x_i)\cdot\mathbf{H}_1^T\cdot\alpha_1\right|}, \left|\mathbf{h}^T(x_i)\cdot\mathbf{H}_1^T\cdot\alpha_1\right|\leq\varepsilon_1, i\in(1,\ldots,m_1)\\ \sigma_1,\quad otherwise\end{cases} \tag{45}$$

$$d_j = \begin{cases} \frac{1}{|\xi_j|}, |\xi_j| \leq \varepsilon_2, j \in (1, \ldots, m_2) \\ \\ \sigma_2, \quad otherwise \end{cases} \tag{46}$$

Here σ_1, σ_2 are two small constants.

CL_1-FTELM$_2$:

$$\begin{aligned} \min_{\alpha_2, \eta} \quad & \frac{1}{2}\alpha_2^T \cdot \mathbf{K}_{ELM2} \cdot \mathbf{R} \cdot \mathbf{K}_{ELM2} \cdot \alpha_2 + \frac{C_3}{2} \cdot \eta^T \cdot \mathbf{S} \cdot \eta \\ & + \frac{C_4}{2} \cdot \alpha_2^T \cdot \mathbf{K}_{ELM2} \cdot \mathbf{N}_2 \cdot \mathbf{K}_{ELM2} \cdot \alpha_2 \\ & s.t. \quad \mathbf{H}_1 \cdot \mathbf{H}_2^T \cdot \alpha_2 + \eta \geq \mathbf{e}_1 \end{aligned} \tag{47}$$

where **R** and **S** are two diagonal matrices with j-th and i-th diagonal elements as:

$$r_j = \begin{cases} \frac{1}{|\mathbf{h}^T(x_j) \cdot \mathbf{H}_2^T \cdot \alpha_2|}, |\mathbf{h}^T(x_j) \cdot \mathbf{H}_2^T \cdot \alpha_2| \leq \varepsilon_3, j \in (1, \ldots, m_2) \\ \\ \sigma_3, \quad otherwise \end{cases} \tag{48}$$

$$s_i = \begin{cases} \frac{1}{|\eta_i|}, |\eta_i| \leq \varepsilon_4, i \in (1, \ldots, m_1) \\ \\ \sigma_4, \quad otherwise \end{cases} \tag{49}$$

Here σ_3, σ_4 are two small constants.

4.2. Model Solution

Introducing Lagrange multipliers α, the Lagrange function of (44) can be written as follows:

$$\begin{aligned} \mathcal{L}(\alpha_1, \xi, \alpha) = & \frac{1}{2}\alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot (\mathbf{F} + C_2 \cdot \mathbf{N}_1) \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 + \frac{C_1}{2} \cdot \xi^T \cdot \mathbf{D} \cdot \xi \\ & - \alpha^T \cdot \left(-\mathbf{H}_2 \cdot \mathbf{H}_1^T \cdot \alpha_1 + \xi - \mathbf{e}_2 \right) \end{aligned} \tag{50}$$

According to the KKT conditions, we can get the following formula:

$$\frac{\partial \mathcal{L}}{\partial \alpha_1} = \mathbf{K}_{ELM1} \cdot (\mathbf{F} + C_2 \cdot \mathbf{N}_1) \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 + \mathbf{H}_1 \cdot \mathbf{H}_2^T \cdot \alpha = 0 \tag{51}$$

$$\frac{\partial \mathcal{L}}{\partial \xi} = C_1 \cdot \mathbf{D} \cdot \xi - \alpha = 0 \tag{52}$$

$$\alpha^T \cdot \left(-\mathbf{H}_2 \cdot \mathbf{H}_1^T \cdot \alpha_1 + \xi - \mathbf{e}_2 \right) = 0 \tag{53}$$

$$\alpha \geq 0 \tag{54}$$

From Equations (51) and (52), we can get:

$$\begin{aligned} \alpha_1 &= -(\mathbf{K}_{ELM1} \cdot (\mathbf{F} + C_2 \cdot \mathbf{N}_1) \cdot \mathbf{K}_{ELM1})^{-1} \cdot \mathbf{H}_1 \cdot \mathbf{H}_2^T \cdot \alpha \\ \xi &= \frac{1}{C_1} \cdot \mathbf{D}^{-1} \cdot \alpha \end{aligned}$$

Similarly, we can get:

$$\begin{aligned} \alpha_2 &= (\mathbf{K}_{ELM2} \cdot (\mathbf{R} + C_4 \cdot \mathbf{N}_2) \cdot \mathbf{K}_{ELM2})^{-1} \cdot \mathbf{H}_2 \cdot \mathbf{H}_1^T \cdot \lambda \\ \eta &= \frac{1}{C_3} \cdot \mathbf{S}^{-1} \cdot \lambda \end{aligned}$$

Thus, we can get the dual problem of (44) as follows:

Dual CL_1-FTELM$_1$

$$\min_{\alpha \geq \mathbf{0}} \frac{1}{2}\alpha^T \cdot \left(\left(\mathbf{H}_2\mathbf{H}_1^T\right)\mathbf{Q}_1^{-1}\left(\mathbf{H}_1\mathbf{H}_2^T\right) + \frac{1}{C_1}\mathbf{D}^{-1} \right) \cdot \alpha - \mathbf{e}_2^T \cdot \alpha \tag{55}$$

where $\mathbf{Q}_1 = \mathbf{K}_{ELM1} \cdot (\mathbf{F} + C_2 \cdot \mathbf{N}_1) \cdot \mathbf{K}_{ELM1}$.

In the same way, we can obtain the dual problem of the Equation (47) as follows:

Dual CL_1-FTELM$_2$

$$\min_{\lambda \geq \mathbf{0}} \frac{1}{2}\lambda^T \cdot \left(\left(\mathbf{H}_1\mathbf{H}_2^T\right)\mathbf{Q}_2^{-1}\left(\mathbf{H}_2\mathbf{H}_1^T\right) + \frac{1}{C_3}\mathbf{S}^{-1} \right) \cdot \lambda - \mathbf{e}_1^T \cdot \lambda \tag{56}$$

where $\mathbf{Q}_2 = \mathbf{K}_{ELM2} \cdot (\mathbf{R} + C_4 \cdot \mathbf{N}_2) \cdot \mathbf{K}_{ELM2}$.

After solving (55) and (56), α and λ are derived, and then α_1 and α_2 are obtained. So, the decision function of CL_1-FTELM is as follows:

$$y = arg \min_{k=1,2} \left| \alpha_k^T \cdot \mathbf{H}_k \cdot \mathbf{h}(\mathbf{x}) \right| = arg \min_{k=1,2} \sum_{i=1}^{m_k} \alpha_{k_i} \cdot k_{ELM_k}(x, x_i) \tag{57}$$

Based on the above discussion, our algorithm will be presented in Algorithm 2. Next, we give the convergence analysis of Algorithm 2.

Algorithm 2 The procedure of CL_1-FTELM.

Input:

Training set $T = \{x_i, y_i\}_{i=1}^m$, where $x_i \in R^d$, $y_i = \pm 1$, the number of hidden node number L, regularization parameters $C_1, C_2, C_3, C_4 > 0$, $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 > 0$, $\rho_1, \rho_2, \sigma_1, \sigma_2, \sigma_3, \sigma_4$.

Output:

α_1^*, α_2^* and the decision function of CL_1-FTELM.

1: Initialize $\mathbf{F}_0 \in R^{m_1 \times m_1}$, $\mathbf{D}_0 \in R^{m_2 \times m_2}$, $\mathbf{R}_0 \in R^{m_2 \times m_2}$, $\mathbf{S}_0 \in R^{m_1 \times m_1}$.

2: Compute the graph matrix $\mathbf{N}_1$, $\mathbf{N}_2$ by Equations (4) and (5).

3: Choose an activation function such as $G(x) = \frac{1}{1+e^{-x}}$ and compute the hidden layer output matrix $\mathbf{H}_1$, $\mathbf{H}_2$ by $\mathbf{h}(x_i) = G\left(\sum_{j=1}^d \omega_{ji} x_j + b_i\right)$ and compute $\mathbf{K}_{ELM1} = \mathbf{H}_1\mathbf{H}_1^T$ and $\mathbf{K}_{ELM2} = \mathbf{H}_2\mathbf{H}_2^T$.

4: Set t = 0.

5: **While**

Solving (55) and (56), the α^t and λ^t can be obtained.

Then get the solution α_1^t, α_2^t, ξ^t, and η^t by

$$\alpha_1^t = -(\mathbf{K}_{ELM1} \cdot (\mathbf{F}_t + C_2 \cdot \mathbf{N}_1) \cdot \mathbf{K}_{ELM1})^{-1} \cdot \mathbf{H}_1 \cdot \mathbf{H}_2^T \cdot \alpha^t, \xi^t = \frac{1}{C_1} \cdot \mathbf{D}_t^{-1} \cdot \alpha^t$$

$$\alpha_2^t = (\mathbf{K}_{ELM2} \cdot (\mathbf{R}_t + C_4 \cdot \mathbf{N}_2) \cdot \mathbf{K}_{ELM2})^{-1} \cdot \mathbf{H}_2 \cdot \mathbf{H}_1^T \cdot \lambda^t, \eta^t = \frac{1}{C_3} \cdot \mathbf{S}_t^{-1} \cdot \lambda^t$$

Update the matrices F_{t+1}, D_{t+1}, R_{t+1}, and S_{t+1} by (45), (46), (48) and (49), respectively.

Compute the objective function values J_1^{t+1} and J_2^{t+1}, by

$$J_1^{t+1} = \frac{1}{2}(\alpha_1^t)^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{F}_{t+1} \cdot \mathbf{K}_{ELM1} \cdot \alpha_1^t + \frac{C_1}{2} \cdot (\xi^t)^T \cdot \mathbf{D}_{t+1} \cdot \xi^t + \frac{C_2}{2} \cdot (\alpha_1^t)^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \alpha_1^t \tag{58}$$

$$J_2^{t+1} = \frac{1}{2}(\alpha_2^t)^T \cdot \mathbf{K}_{ELM2} \cdot \mathbf{R}_{t+1} \cdot \mathbf{K}_{ELM2} \cdot \alpha_2^t + \frac{C_3}{2} \cdot (\eta^t)^T \cdot \mathbf{S}_{t+1} \cdot \eta^t + \frac{C_4}{2} \cdot (\alpha_2^t)^T \cdot \mathbf{K}_{ELM2} \cdot \mathbf{N}_2 \cdot \mathbf{K}_{ELM2} \cdot \alpha_2^t \tag{59}$$

if $\left|J_1^{t+1} - J_1^t\right| \leq \rho_1$ and $\left|J_2^{t+1} - J_2^t\right| \leq \rho_2$.

break

else

t=t+1

6: **end while**

7: Stop the iteration process and get the solution of α_1^*, α_2^*.

4.3. Convergence Analysis

Before we prove the convergence of the iterative algorithm, we first review two lemmas [34].

Lemma 1. *For any non-zeros vectors* $x, y \in R^n$, *if* $f(x) = \|x\|_1 - \frac{\|x\|_1^2}{2\|y\|_1}$, *then the following inequalities* $f(x) \leq f(y)$ *hold.*

Lemma 2. *For any non-zeros vectors* $x, y, p, q \in R^n$, *if* $f(x,p) = \|x\|_1 - \frac{\|x\|_1^2}{2\|y\|_1} + C\left(\|p\|_1 - \frac{\|p\|_1^2}{2\|q\|_1}\right)$, $C \in R^+$, *then the following inequalities* $f(x,p) \leq f(y,q)$ *hold.*

The proof of two lemmas is detailed in [34].

Theorem 1. *Algorithm 2 monotonically decreases the objectives of problems (42) and (43) in each iteration until it converges.*

Proof. Here, we only use problem (42) as an example to prove Theorem 1.

$$\begin{aligned} J(\alpha_1, \xi) = \min_{\alpha_1, \xi} \sum_{i=1}^{m_1} \min\left(\left\|\mathbf{h}^T(x_i) \cdot \mathbf{H_1}^T \cdot \alpha_1\right\|_1, \varepsilon_1\right) + C_1 \cdot \sum_{j=1}^{m_2} \min\left(\|\xi_j\|_1, \varepsilon_2\right) \\ + \frac{C_2}{2} \cdot \alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 \end{aligned} \tag{60}$$

when $\left\|\mathbf{h}^T(x_i) \cdot \mathbf{H}_1^T \cdot \alpha_1\right\|_1 < \varepsilon_1$ and $\|\xi_j\|_1 < \varepsilon_2$,we have:

$$\begin{aligned} J(\alpha_1, \xi) = \min_{\alpha_1, \xi} \sum_{i=1}^{m_1} \left\|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right\|_1 + C_1 \sum_{j=1}^{m_2} \|\xi_j\|_1 \\ + \frac{C_2}{2} \cdot \alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 \end{aligned} \tag{61}$$

We take the derivative of Equation (61) with respect to α_1 and ξ separately and then obtain that:

$$\begin{cases} \sum_{i=1}^{m_1} \frac{\mathbf{H}_1\mathbf{h}(x_i)\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1}{\left|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right|} + C_2 \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N_1} \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 = 0 \\ C_1 \cdot \sum_{j=1}^{m_2} \frac{\xi_j}{|\xi_j|} = 0 \end{cases} \tag{62}$$

by the above Equation (62), we can get:

$$\begin{aligned} \sum_{i=1}^{m_1} \frac{\mathbf{H}_1\mathbf{h}(x_i)\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1}{\left|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right|} + C_1 \cdot \sum_{j=1}^{m_2} \frac{\xi_j}{|\xi_j|} \\ + C_2 \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N_1} \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 = 0 \end{aligned} \tag{63}$$

We define $f_i = \frac{1}{\left|\mathbf{h}^T(x_i) \cdot \mathbf{H}_1^T \cdot \alpha_1\right|}$ and $d_j = \frac{1}{|\xi_j|}$ as the diagonal entries of $\mathbf{F}$ and $\mathbf{D}$, respectively. Thus we can rewrite Equation (63) as follows:

$$\mathbf{H}_1 \cdot \mathbf{H}_1^T \cdot \mathbf{F} \cdot \mathbf{H}_1^T \cdot \mathbf{H}_1 \cdot \alpha_1 + C_1 \cdot \mathbf{D} \cdot \xi + C_2 \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N_1} \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 = 0 \tag{64}$$

Obviously, Equation (64) is the optimal solution to the following problem:

$$\begin{aligned} \min_{\alpha_1, \xi} \quad & \frac{1}{2}\alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{F} \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 + \frac{C_1}{2} \cdot \xi^T \cdot \mathbf{D} \cdot \xi \\ & + \frac{C_2}{2} \cdot \alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 \end{aligned} \tag{65}$$

Now, assume that $\bar{\alpha}_1$ and $\bar{\xi}$ denote the updated α_1 and ξ of Algorithm 2, respectively. Thus we can get:

$$\begin{aligned}
&\frac{1}{2}\bar{\alpha}_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{F} \cdot \mathbf{K}_{ELM1} \cdot \bar{\alpha}_1 + \frac{C_1}{2} \cdot \bar{\xi}^T \cdot \mathbf{D} \cdot \bar{\xi} \\
&+ \frac{C_2}{2} \cdot \bar{\alpha}_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \bar{\alpha}_1 \\
&\leq \frac{1}{2}\alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{F} \cdot \mathbf{K}_{ELM1} \cdot \alpha_1 + \frac{C_1}{2} \cdot \xi^T \cdot \mathbf{D} \cdot \xi \\
&+ \frac{C_2}{2} \cdot \alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \alpha_1
\end{aligned} \tag{66}$$

we have rewritten Equation (66) as follows

$$\begin{aligned}
&\sum_{i=1}^{m_1} \frac{(\mathbf{K}_{ELM1}\bar{\alpha}_1)^T(\mathbf{K}_{ELM1}\bar{\alpha}_1)}{2\left|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right|} + \sum_{j=1}^{m_2} \frac{C_1\left(\bar{\xi}_j\right)^2}{2|\xi_j|} \\
&+ \frac{C_2}{2}\bar{\alpha}_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \bar{\alpha}_1 \\
&\leq \sum_{i=1}^{m_1} \frac{(\mathbf{K}_{ELM1}\alpha_1)^T(\mathbf{K}_{ELM1}\alpha_1)}{2\left|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right|} + \sum_{j=1}^{m_2} \frac{C_1\left(\xi_j\right)^2}{2|\xi_j|} \\
&+ \frac{C_2}{2}\alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \alpha_1
\end{aligned} \tag{67}$$

Here, we let $x = \mathbf{K}_{ELM1}\bar{\alpha}_1$, $y = \mathbf{K}_{ELM1}\alpha_1$, $C = C_1$, $p = \bar{\xi}_j$, $q = \xi_j$. Based on Lemma 2, we have

$$\begin{aligned}
&|\mathbf{K}_{ELM1} \cdot \bar{\alpha}_1| - \frac{|\mathbf{K}_{ELM1} \cdot \bar{\alpha}_1|^2}{2|\mathbf{K}_{ELM1} \cdot \alpha_1|} + C_1 \cdot \left(|\bar{\xi}_j| - \frac{|\bar{\xi}_j|^2}{2|\xi_j|}\right) \\
&\leq |\mathbf{K}_{ELM1} \cdot \alpha_1| - \frac{|\mathbf{K}_{ELM1} \cdot \alpha_1|^2}{2|\mathbf{K}_{ELM1} \cdot \alpha_1|} + C_1 \cdot \left(|\xi_j| - \frac{|\xi_j|^2}{2|\xi_j|}\right)
\end{aligned} \tag{68}$$

then we can get

$$\begin{aligned}
&\sum_{i=1}^{m_1}\left(\left|\mathbf{h}^T(x_i) \cdot \mathbf{H}_1^T \cdot \bar{\alpha}_1\right| - \frac{|\mathbf{K}_{ELM1} \cdot \bar{\alpha}_1|^2}{2|\mathbf{K}_{ELM1} \cdot \alpha_1|}\right) + C_1 \sum_{j=1}^{m_2}\left(|\bar{\xi}_j| - \frac{|\bar{\xi}_j|^2}{2|\xi_j|}\right) \\
&\leq \sum_{i=1}^{m_1}\left(\left|\mathbf{h}^T(x_i) \cdot \mathbf{H}_1^T \cdot \alpha_1\right| - \frac{|\mathbf{K}_{ELM1} \cdot \alpha_1|^2}{2|\mathbf{K}_{ELM1} \cdot \alpha_1|}\right) + C_1 \sum_{j=1}^{m_2}\left(|\xi_j| - \frac{|\xi_j|^2}{2|\xi_j|}\right)
\end{aligned} \tag{69}$$

combining (67) and (69), we can get the following inequalities

$$\begin{aligned}
&\sum_{i=1}^{m_1}\left(\left|\mathbf{h}^T(x_i) \cdot \mathbf{H}_1^T \cdot \bar{\alpha}_1\right|\right) + C_1 \sum_{j=1}^{m_2}\left(|\bar{\xi}_j|\right) \\
&+ \frac{C_2}{2}\bar{\alpha}_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \bar{\alpha}_1 \\
&\leq \sum_{i=1}^{m_1}\left(\left|\mathbf{h}^T(x_i) \cdot \mathbf{H}_1^T \cdot \alpha_1\right|\right) + C_1 \sum_{j=1}^{m_2}\left(|\xi_j|\right) \\
&+ \frac{C_2}{2}\alpha_1^T \cdot \mathbf{K}_{ELM1} \cdot \mathbf{N}_1 \cdot \mathbf{K}_{ELM1} \cdot \alpha_1
\end{aligned} \tag{70}$$

further, we can get

$$\begin{aligned}
&\sum_{i=1}^{m_1} \min\left(\left|\mathbf{h}^T(x_i)\cdot\mathbf{H}_1^T\cdot\bar{\alpha}_1\right|\right) + C_1\sum_{j=1}^{m_2}\min\left(\left|\bar{\xi}_j\right|\right)\\
&+\frac{C_2}{2}\bar{\alpha}_1^T\cdot\mathbf{K}_{ELM1}\cdot\mathbf{N}_1\cdot\mathbf{K}_{ELM1}\cdot\bar{\alpha}_1\\
&\le\sum_{i=1}^{m_1} \min\left(\left|\mathbf{h}^T(x_i)\cdot\mathbf{H}_1^T\cdot\alpha_1\right|\right) + C_1\sum_{j=1}^{m_2}\min\left(\left|\xi_j\right|\right)\\
&+\frac{C_2}{2}\alpha_1^T\cdot\mathbf{K}_{ELM1}\cdot\mathbf{N}_1\cdot\mathbf{K}_{ELM1}\cdot\alpha_1
\end{aligned} \tag{71}$$

Therefore, we have $J(\bar{\alpha}_1,\bar{\xi}) \le J(\alpha_1,\xi)$. Similarly, when $\left\|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right\|_1 \le \varepsilon_1$ and $\|\xi_j\| \ge \varepsilon_2$, or $\left\|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right\|_1 \ge \varepsilon_1$ and $\|\xi_j\| \le \varepsilon_2$, or $\left\|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right\|_1 \ge \varepsilon_1$ and $\|\xi_j\| \ge \varepsilon_2$, we can obviously get $J(\bar{\alpha}_1,\bar{\xi}) \le J(\alpha_1,\xi)$. Thus, the inequality $J(\bar{\alpha}_1,\bar{\xi}) \le J(\alpha_1,\xi)$ holds. The three terms in Equation (60) are equal to or greater than 0. Meaning that Algorithm 2 decreases objective of problem (42) until convergence. □

Theorem 2. *Algorithm 2 will converge to a local optimum to the problem in (42).*

Proof. Here, we only use (42) as an example to prove Theorem 2.

When $\left\|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right\|_1 \le \varepsilon_1$ and $\left\|\xi_j\right\|_1 \le \varepsilon_2$, we write out the formula of (42) Lagrange function:

$$\begin{aligned}
\mathcal{L}_1(\alpha_1,\xi,\lambda) &= \sum_{i=1}^{m_1}\left(\left\|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right\|_1\right) + C_1\sum_{j=1}^{m_2}\left(\left\|\xi_j\right\|_1\right)\\
&+\frac{C_2}{2}\alpha_1^T\cdot\mathbf{K}_{ELM1}\cdot\mathbf{N}_1\cdot\mathbf{K}_{ELM1}\cdot\alpha_1 - \lambda^T\sum_{j=1}^{m_2}\left(\mathbf{h}^T(x_j)\mathbf{H}_1^T\alpha_1+\xi_j-1\right)
\end{aligned} \tag{72}$$

Then, we take the derivative of $\mathcal{L}(\alpha_1,\xi,\lambda)$ with respect to α_1

$$\begin{aligned}
\frac{\partial\mathcal{L}_1}{\partial\alpha_1} &= \sum_{i=1}^{m_1}\frac{\mathbf{H}_1\mathbf{h}(x_i)\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1}{\left|\mathbf{h}^T(x_i)\mathbf{H}_1^T\alpha_1\right|} + C_2\cdot\mathbf{K}_{ELM1}\cdot\mathbf{N}_1\cdot\mathbf{K}_{ELM1}\cdot\alpha_1\\
&+\mathbf{H}_1\cdot\mathbf{H}_2^T\cdot\lambda = \mathbf{K}_{ELM1}\cdot(\mathbf{F}+C_2\cdot\mathbf{N}_1)\cdot\mathbf{K}_{ELM1}\cdot\alpha_1+\mathbf{H}_1\cdot\mathbf{H}_2^T\cdot\lambda = \mathbf{0}
\end{aligned} \tag{73}$$

Similarly, we get the Lagrangian function of problem (44):

$$\begin{aligned}
\mathcal{L}_2(\alpha_1,\xi,\alpha) = \frac{1}{2}\alpha_1^T\cdot\mathbf{K}_{ELM1}\cdot(\mathbf{F}+C_2\cdot\mathbf{N}_1)\cdot\mathbf{K}_{ELM1}\cdot\alpha_1+\frac{C_1}{2}\cdot\xi^T\cdot\mathbf{D}\cdot\xi\\
-\lambda^T\cdot\left(-\mathbf{H}_2\cdot\mathbf{H}_1^T\cdot\alpha_1+\xi-\mathbf{e}_2\right)
\end{aligned} \tag{74}$$

Taking the derivative of $\mathcal{L}_2(\alpha_1,\xi,\alpha)$ with respect to α_1:

$$\frac{\partial\mathcal{L}_2}{\partial\alpha_1} = \mathbf{K}_{ELM1}\cdot(\mathbf{F}+C_2\cdot\mathbf{N}_1)\cdot\mathbf{K}_{ELM1}\cdot\alpha_1+\mathbf{H}_1\cdot\mathbf{H}_2^T\cdot\lambda = \mathbf{0} \tag{75}$$

The other three cases are similar. From the discussion above, we may safely draw that Equations (73) and (75) are equivalent, so we can use problem (44) instead of problem (42) to solve CL_1-FTELM, which further illustrates that Algorithm 2 can converge to a local optimal solution. □

5. Experiments

Description of the four comparison algorithms:

OPTELM: The optimization function of the model consists of minimizing the L_2-norm of the weight vector and minimizing empirical loss. It neither consider the establishment of two non-parallel hyperplanes to deal with classification tasks, nor consider the statistical information of samples. At the same time, since it uses L_2-norm metric and Hinge loss, it has weak anti-noise ability.

TELM: The optimization function of the model consists of minimizing the distance from the sample points to the hyperplane as well as minimizing empirical loss. TELM does not fully consider the statistical information of the sample. At the same time, its metric uses the L_2-norm metric and the loss function uses the Hinge loss. When there is noise in the data set, the influence of noise data will be amplified and the accuracy of classification will be reduced.

FELM: The optimization function of the model includes minimizing the L_2-norm of the weight vector, minimizing empirical loss, and minimizing the within-class scatter of the number sample data. Although FELM takes into account the statistics of the sample, it has to deal with a much larger optimization problem than the twin extreme learning machines, which is time-consuming. At the same time, FELM still continues the metric and loss used by OPTELM, so its anti-noise ability is weak.

CL_1-TWSVM: CL_1-TWSVM is formed on the basis of twin support vector machines by changing the model's metric and loss to capped L_1-norm. Although CL_1-TWSVM has the ability to resist noise, it does not fully take into account the statistics of the data. Meanwhile, CL_1-TWSVM not only needs to solve the weight vector of the hyperplane, but also needs to solve the bias of the hyperplane, so it is time-consuming.

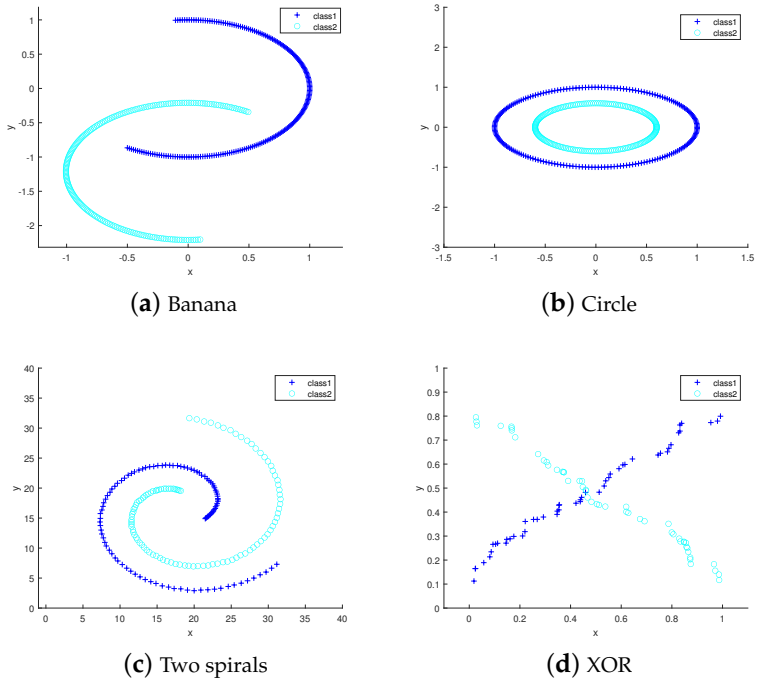
We systematically compare our algorithm above advanced algorithms (OPTELM [12], TELM [14], FELM [19], and CL_1-TWSVM [29]) on artificial synthetic datasets and UCI real datasets to verify the effectiveness of our FTELM and CL_1-FTELM. In Section 5.1, we describe the relevant experimental setting in detail. We describe their performance in different cases in Sections 5.2 and 5.3, respectively. In Section 5.4, we use the one-versus-rest multi-classification method to perform data classification tasks in four image datasets: Yale "http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html (accessed on 15 February 2023)", ORL "http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html (accessed on 15 February 2023)", USPS "http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html (accessed on 15 February 2023)" handwritten digit dataset and MNIST "http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html (accessed on 15 February 2023)" dataset.

5.1. Experimental Setting

All experiments were implemented in MATLAB R2020a installed in a personal computer (PC) with an AMD Radeon Graphics processor (3.2 GHz), and 16 GB random-access memory (RAM). For CL_1-TWSVM, and CL_1-FTELM, we take the maximum number of iterations to be 100 and the iteration stopping threshold to be 0.001. The activation functions used in a total of five models (OPTELM, TELM, FELM, FTELM, and CL_1-FTELM) are $G(x) = \frac{1}{1+e^{-x}}$. The Gaussian kernel function $K(x,z) = e^{-\frac{\|x-z\|^2}{2\sigma^2}}$ was used for CL_1-TWSVM. The parameters selected by all the above algorithms are as follows: ε_1, ε_2, ε_3, ε_4 were selected from $\{10^i | -6, -5, -4\}$, C_1, C_2, C_3, C_4 were selected from $\{10^i | -5, -4, \ldots, 4, 5\}$, σ was chosen from $\{2^i | -3, -2, \ldots, 2, 3\}$, and the hidden layer node number L was chosen from $\{50, 100, 200, 500, 1000, 2000, 5000, 10{,}000\}$. The optimal parameters used by the model are selected by 10-fold cross-validation and grid search. Normalization was performed for both artificial and UCI datasets. For image datasets, we randomly select 20% of the data as the test set to get the classification accuracy of the algorithm. All experimental processes are repeated 10 times and the average of the 10 test results is used as the performance measure, and the evaluation criterion selected in this paper is classification accuracy (ACC).

5.2. Experiments on Artificial Datasets

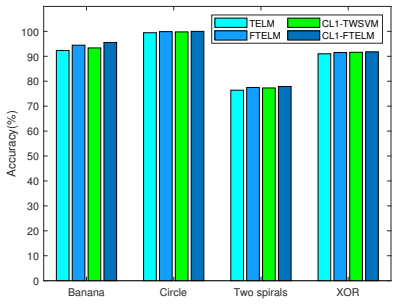
We first do experiments on the Banana, Circle, Two spirals, and XOR datasets which are generated by trigonometric function(sine, cosine), two circle lines, two spirals lines, and two intersecting lines, respectively. The two-dimensional distributions of the four synthetic datasets are shown in Figure 1. Dark blue '+' represents class 1, and cyan '○' represents class 2. Figure 2 illustrates the experimental results of four twin algorithms namely TELM, FTELM, CL_1-TWSVM, and CL_1-FTELM for four datasets with 0%, 20%, and 25% noise in terms of accuracy. From Figure 2a, we can observe that the classification accuracy of our FTELM and CL_1-FTELM in Banana and Two spirals datasets is higher than the other two methods. In the Circle and XOR datasets, the classification accuracy of the four methods is similar. The experimental results show that fully considering the statistical information of the data can effectively improve the classification accuracy of the classifier, which shows that our CL_1-FTELM method is effective. From Figure 2b,c, we can see that the overall effect of FTELM is better than TELM. This shows the importance of fully considering the statistical information of the sample. At the same time, we can see that CL_1-FTELM has the best effect, followed by CL_1-TWSVM. It shows that the capped L_1-norm can control the influence of noise on the model in a certain range, and further shows the effectiveness of using the capped L_1-norm. In summary, Figure 2 illustrates the effectiveness of considering sample statistics information and changing the distance metric and loss of the model into capped L_1-norm at the same time.



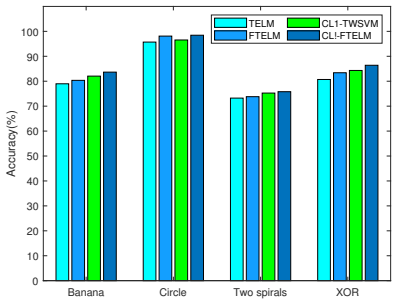
(**a**) Banana (**b**) Circle

(**c**) Two spirals (**d**) XOR

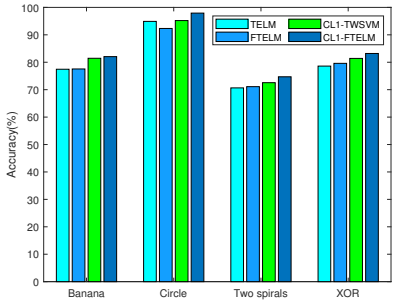
Figure 1. Four types of data without noise.



(**a**) Accuracy for four algorithms on four types of data without noise.



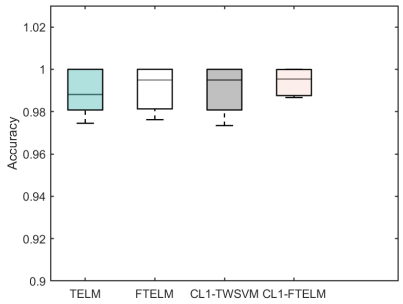
(**b**) Accuracy for four algorithms on four types of data with 20% noise.



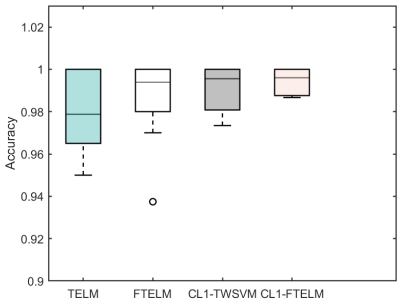
(**c**) Accuracy for four algorithms on four types of data with 25% noise.

Figure 2. Accuracy for TELM, FTELM, CL_1-TWSVM, and CL_1-FTELM on four types of data with 0%, 20%, and 25% noise.

To further show the robustness of CL_1-FTELM, we add noise with different ratios to the Circle dataset. Figure 3 shows the accuracy of TELM, FTELM, CL_1-TWSVM, and CL_1-FTELM algorithms on the Circle dataset in different noises ratios. The ratio is set in the range of $\{0.1, 0.15, 0.2, 0.25\}$. We plot the accuracy results of ten experiments with different noise ratios in a box-shaped plot. By observing the median of the four subgraphs, we can find that the median of CL_1-FTELM algorithm is much higher than the other three algorithms. And CL_1-FTELM method in four different noise ratios results is relatively concentrated. In other words, the variance of ten experimental results obtained by the CL_1-FTELM algorithm is smaller and the mean value is larger. The above results show that our CL_1-FTELM has better stability and better classification effect in environments containing noise. This shows the effectiveness and noise resistance of the distance metric and loss functions of the model using the capped L_1-norm.

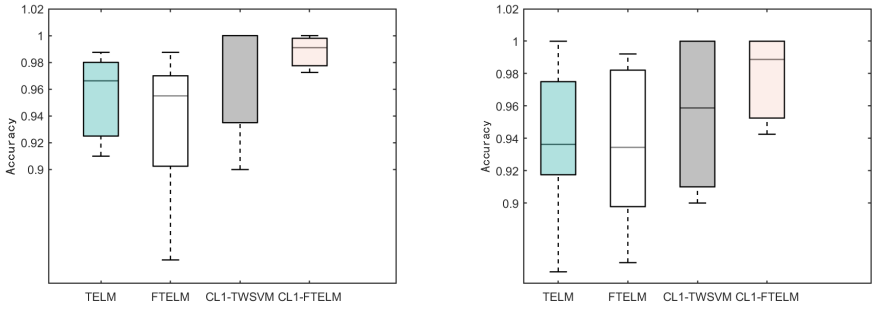


(**a**) Accuracy for four algorithms on Circle dataset with 10% noise.



(**b**) Accuracy for four algorithms on Circle dataset with 15% noise.

Figure 3. *Cont*.



(**c**) Accuracy for four algorithms on Circle dataset with 20% noise.

(**d**) Accuracy for four algorithms on Circle dataset data with 25% noise.

Figure 3. Accuracy for TELM, FTELM, CL_1-TWSVM, and CL_1-FTELM on Circle dataset with noise in different ratios.

5.3. Experiments on UCI Datasets

In this section, we conduct the numerical simulation on UCI datasets. Table 1 describes the features of the UCI datasets used in detail. We also added two algorithms (OPTELM, FELM) to verify the classification performance of FTELM and CL_1-FTELM in ten UCI data sets.

Table 1. Characteristics of UCI datasets.

Datasets	Instances	Attributes	Datasets	Instances	Attributes
Australian	690	14	Vote	432	16
German	1000	24	Ionosphere	351	35
Breast cancer	699	9	Pima	768	8
WDBC	569	30	QSAR	1055	41
Wholesale	440	7	Spam	4601	57

All experimental results obtained based on the optimal parameters are shown in Table 2. Here, the average running time according to the optimal parameters is denoted by Times(s), and the average classification plus or minus standard deviation is denoted by $ACC\pm$. From Table 2, we can see that FTELM performs better than OPTELM, TELM, and FELM on all ten datasets. This indicates that adding Fisher regularization term on the basis of TELM framework can significantly improve the accuracy of model classification. In addition, the average training time of FTELM algorithm on most data sets is smaller than that of FELM algorithm, which indicates that FTELM has inherited the advantages of TELM's short training time. In addition, we also can draw our CL_1-FTELM in most data sets has achieved the highest classification accuracy besides the WDBC data set. Through the analysis of the above results, we can conclude that the Fisher regularization and capped L_1-norm added to the TELM learning framework can effectively improve the performance of the classifier. It is shown that the proposed FTELM and CL_1-FTELM are efficient algorithms.

In order to more significantly verify the robustness of CL_1-FTELM to outliers, we added 20% and 25% Gaussian noise to 10 data sets, respectively. All experimental results are presented in Tables 3 and 4. From Tables 3 and 4, we find that the classification accuracy of all six algorithms decreases after adding noise. However, the classification accuracy of our algorithm CL_1-FTELM is the highest of the eight datasets, which further reveals the effectiveness of our method using capped L_1-norm instead of Hinge loss and L_2-norm distance metric. Compared with the other five algorithms, our CL_1-FTELM algorithm is more time-consuming. This is due to that CL_1-FTELM requires a lot of time in the process of training to iterative calculation, eliminating outliers, and computing graph matrices. In addition, we used different noise factor values (0.1, 0.15, 0.2, 0.25, 0.3) on the Cancer,

German, Ionosphere, and WDBC for the six algorithms. The experimental results are given in Figure 4. It can be seen from Figure 4a that when the Breast Cancer dataset contains 10% noise, the effects of our FTELM and CL_1-FTELM are comparable. This shows that it is important to consider the statistical information of the sample. As the ratio of noise increases, the classification accuracy of all methods decreases, but our CL_1-FTELM still has the highest accuracy. This illustrates the effectiveness of our using the capped L_1-norm. Figure 4b shows that with the increase of noise ratio, the decline trend of accuracy of CL_1-TWSVM and CL_1-FTELM is similar, but CL_1-FTELM is still the most stable among the six methods when facing the influence of noise. From both Figure 4c,d, we can clearly observe that the anti-noise effect of our CL_1-FTELM is the best. This illustrates the effectiveness of using the Fisher regularization term as well as the capped L_1-norm.

Table 2. Experimental results on UCI datasets, The best results are marked in bold.

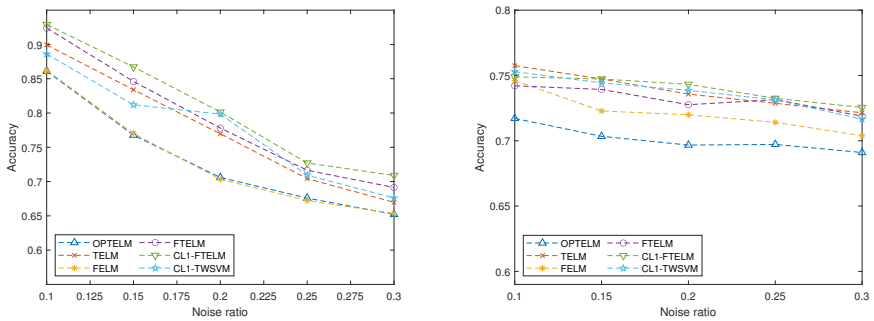
Datasets	**OPTELM** ACC ± S (%) Times (s)	**TELM** ACC ± S (%) Times (s)	**FELM** ACC ± S (%) Times (s)	**FTELM** ACC ± S (%) Times (s)	**CL_1-TWSVM** ACC ± S (%) Times (s)	**CL_1-FTELM** ACC ± S (%) Times (s)
Australian	85.31 ± 0.34	85.60 ± 0.44	85.46 ± 0.19	86.79 ± 0.33	85.82 ± 0.28	**87.13 ± 0.52**
	0.682	0.593	1.698	**0.456**	1.676	2.533
German	76.26 ± 0.52	76.40 ± 0.16	76.50 ± 0.42	76.56 ± 0.47	76.70 ± 0.25	**77.15 ± 1.18**
	1.182	0.979	4.555	**0.474**	5.318	7.006
Breast cancer	95.70 ± 0.24	96.35 ± 0.15	96.45 ± 0.09	97.07 ± 0.15	96.39 ± 0.13	**97.32 ± 0.53**
	0.601	0.668	1.646	**0.505**	4.011	3.902
WDBC	96.71 ± 0.27	97.13 ± 0.48	97.55 ± 0.17	**98.55 ± 0.26**	97.09 ± 0.25	97.86 ± 0.21
	0.416	0.605	1.144	0.578	3.618	4.551
Wholesale	87.35 ± 0.93	89.86 ± 0.84	90.26 ± 0.12	90.56 ± 0.33	89.89 ± 0.30	**90.70 ± 0.56**
	0.278	2.091	0.665	0.359	1.246	1.377
Vote	95.31 ± 0.16	95.56 ± 0.30	96.04 ± 0.24	96.12 ± 0.31	95.21 ± 0.54	**96.43 ± 0.35**
	0.256	0.502	0.651	0.445	1.077	0.992
Ionosphere	90.59 ± 0.84	91.38 ± 0.52	92.32 ± 0.32	92.74 ± 0.83	92.56 ± 0.54	**93.32 ± 1.21**
	0.184	0.476	0.421	0.268	1.128	2.237
Pima	76.83 ± 0.73	77.51 ± 0.08	77.79 ± 0.10	78.24 ± 0.49	77.49 ± 0.37	**78.82 ± 0.98**
	0.858	**0.795**	2.099	0.932	1.743	4.708
QSAR	83.91 ± 0.66	86.56 ± 0.19	87.12 ± 0.18	87.35 ± 0.23	85.72 ± 0.59	**87.50 ± 0.56**
	1.442	**0.979**	2.489	2.864	2.665	14.288
Spam	85.57 ± 0.65	91.38 ± 0.52	89.67 ± 0.21	91.94 ± 1.23	90.56 ± 1.23	**92.27 ± 0.54**
	125.498	**64.314**	488.251	108.232	158.145	170.261

Table 3. Experimental results on UCI datasets with 20% noise, The best results are marked in bold.

Datasets	**OPTELM** ACC ± S (%) Times (s)	**TELM** ACC ± S (%) Times (s)	**FELM** ACC ± S (%) Times (s)	**FTELM** ACC ± S (%) Times (s)	**CL_1-TWSVM** ACC ± S (%) Times (s)	**CL_1-FTELM** ACC ± S (%) Times (s)
Australian	79.68 ± 1.75	80.37 ± 0.56	79.06 ± 1.36	80.44 ± 1.34	81.98 ± 0.87	**82.78 ± 0.57**
	0.621	0.728	1.756	**0.224**	1.708	3.224
German	69.67 ± 0.97	73.57 ± 1.85	71.99 ± 1.35	72.76 ± 0.88	73.86 ± 1.35	**74.32 ± 1.12**
	1.318	0.981	4.102	**0.398**	5.673	6.764
Breast cancer	70.60 ± 0.45	76.97 ± 0.42	70.32 ± 0.37	77.81 ± 0.56	79.84 ± 0.37	**80.14 ± 0.91**
	0.803	0.706	1.552	**0.315**	4.572	5.034
WDBC	82.98 ± 0.15	84.38 ± 1.01	83.29 ± 0.68	89.43 ± 1.15	89.98 ± 0.30	**93.77 ± 0.32**
	0.419	**0.204**	0.992	0.376	3.899	4.861
Wholesale	73.40 ± 0.93	73.77 ± 0.69	73.74 ± 0.76	74.77 ± 0.56	78.74 ± 0.91	**79.47 ± 2.58**
	0.275	0.543	0.659	0.404	0.849	1.420
Vote	93.48 ± 0.62	**94.36 ± 0.60**	94.24 ± 0.82	94.10 ± 0.94	93.90 ± 0.44	94.29 ± 0.61
	0.277	0.619	0.549	**0.114**	1.048	1.398
Ionosphere	80.79 ± 2.88	82.71 ± 2.09	81.00 ± 3.11	86.06 ± 1.67	85.76 ± 1.58	**87.74 ± 1.08**
	0.159	**0.021**	0.456	0.737	0.391	2.081
Pima	65.79 ± 0.23	67.07 ± 0.56	66.12 ± 0.12	66.30 ± 1.34	70.25 ± 1.57	**71.42 ± 0.94**
	0.873	**0.649**	2.051	1.492	1.758	3.968
QSAR	68.32 ± 2.48	68.80 ± 0.95	68.54 ± 2.50	72.28 ± 2.18	71.09 ± 2.02	**72.31 ± 1.98**
	1.534	3.089	4.578	**0.892**	1.828	9.151
Spam	83.16 ± 0.57	87.38 ± 2.31	85.66 ± 0.65	**87.98 ± 0.87**	85.77 ± 2.21	86.75 ± 0.45
	128.798	**60.565**	432.257	106.267	147.365	160.231

Table 4. Experimental results on UCI datasets with 25% noise, The best results are marked in bold.

Datasets	OPTELM ACC ± S (%) Times (s)	TELM ACC ± S (%) Times (s)	FELM ACC ± S (%) Times (s)	FTELM ACC ± S (%) Times (s)	CL_1-TWSVM ACC ± S (%) Times (s)	CL_1-FTELM ACC ± S (%) Times (s)
Australian	73.68 ± 2.20	75.41 ± 1.52	74.25 ± 2.01	76.40 ± 1.19	80.56 ± 1.07	**81.63 ± 0.71**
	0.585	0.673	1.627	**0.206**	2.205	2.261
German	69.72 ± 0.13	72.87 ± 0.82	71.41 ± 0.88	73.15 ± 0.87	73.13 ± 1.16	**73.25 ± 0.76**
	1.565	0.871	3.855	**0.342**	5.233	6.798
Breast cancer	67.59 ± 0.18	70.43 ± 0.79	67.23 ± 0.24	71.65 ± 0.58	70.93 ± 0.52	**72.71 ± 0.49**
	0.654	0.513	1.438	**0.309**	4.476	5.124
WDBC	79.61 ± 0.78	81.66 ± 0.84	79.83 ± 0.72	87.96 ± 1.13	88.50 ± 0.74	**92.43 ± 0.76**
	0.417	**0.197**	0.887	0.334	3.675	4.861
Wholesale	71.79 ± 1.03	71.63 ± 0.89	69.63 ± 0.38	71.60 ± 1.02	75.53 ± 1.02	**75.74 ± 3.48**
	0.570	2.021	0.623	**0.338**	1.147	1.387
Vote	92.62 ± 0.88	92.95 ± 0.50	93.12 ± 0.80	93.21 ± 0.80	93.21 ± 0.68	**93.50 ± 1.00**
	0.252	0.503	0.514	**0.121**	1.213	1.390
Ionosphere	78.15 ± 2.94	78.79 ± 3.01	76.62 ± 3.67	83.59 ± 1.49	82.94 ± 2.90	**85.03 ± 2.28**
	0.229	**0.058**	0.313	0.737	0.576	1.987
Pima	65.67 ± 0.12	65.45 ± 1.55	65.89 ± 0.12	65.79 ± 0.14	**69.01 ± 1.55**	68.51 ± 2.75
	0.803	0.761	2.182	**0.471**	5.532	4.012
QSAR	67.49 ± 3.08	67.81 ± 1.63	70.30 ± 2.33	**71.53 ± 3.00**	70.49 ± 2.13	69.72 ± 2.14
	2.067	2.730	4.251	**0.849**	1.783	12.564
Spam	71.77 ± 1.05	75.35 ± 0.72	70.89 ± 1.23	76.43 ± 1.16	83.56 ± 0.26	**84.75 ± 0.78**
	99.541	**61.254**	462.221	116.267	142.365	165.214

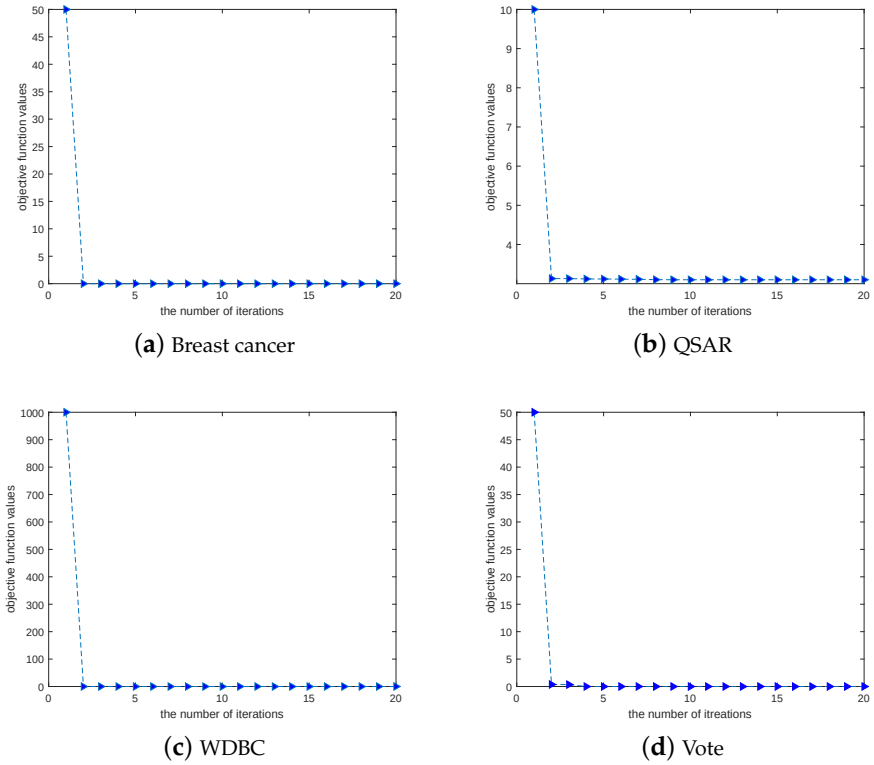


(**a**) Breast cancer (**b**) German

(**c**) Ionosphere (**d**) WDBC

Figure 4. Accuracies of six algorithms via different noises factors.

We also conduct experiments on four data sets (Breast cancer, QSAR, WDBC, and Vote) to verify the convergence of the proposed Algorithm 2. As shown in Figure 5, we plot the objective function value of each iteration. It can be seen that the objective function value converges to a fixed value rapidly with the increase in the number of iterations. This shows that our algorithm can make the objective function value can converge to a local optimal value within a limited number of iterations. The effectiveness and convergence of the Algorithm 2 are demonstrated.



(**a**) Breast cancer (**b**) QSAR

(**c**) WDBC (**d**) Vote

Figure 5. Objective values of CL_1-FTELM on four datasets.

5.4. Experiments on Image Datasets

The image datasets include Yale, ORL, USPS, and MNIST. Figure 6 illustrates examples of four high-dimensional image datasets. The number of samples and characteristics of the four image datasets are shown in Table 5. These four image datasets are used to investigate the performance of our FTELM and CL_1-FTELM for multi-classification. Specifically, for the MNIST dataset, we only select the first 2000 samples to participate in the experiment.

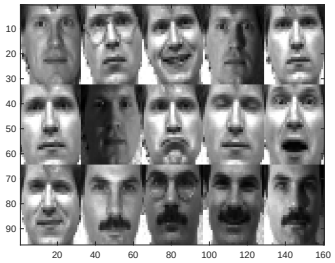
Table 5. Characteristics of image datasets.

Datasets	Instances	Attributes	Datasets	Instances	Attributes
Yale	165	1024	ORL	400	1024
USPS	9298	256	MNIST	70,000	784

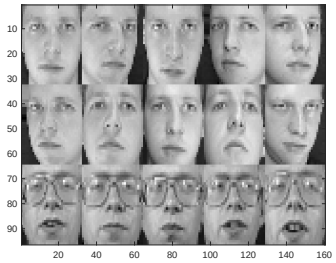
Table 6 shows the specific experimental results. As can be seen from the results of the experiment, our CL_1-FTELM and CL_1-TWSVM have similar training times. This is because this paper uses an iterative algorithm to solve non-convex optimization problem of CL_1-FTELM , which is time-consuming. Simultaneously, the CL_1- FTELM at Yale, ORL, USPS, and MNIST four datasets classification accuracy is highest among the six algorithms. In addition, the classification accuracy of our FTELM algorithm on the four image datasets is the second highest after our CL_1-FTELM. The above results fully show the effectiveness of our two algorithms in dealing with multi-classification tasks.

Table 6. Experimental results on images and handwritten digits datasets. The best results are marked in bold.

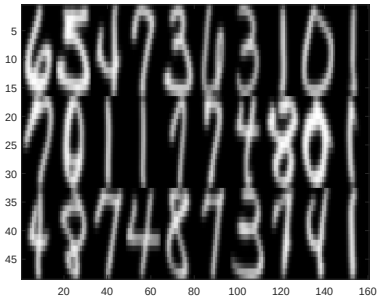
Datasets	OPTELM ACC ± S (%) Times (s)	TELM ACC ± S (%) Times (s)	FELM ACC ± S (%) Times (s)	FTELM ACC ± S (%) Times (s)	CL_1-TWSVM ACC ± S (%) Times (s)	CL_1-FTELM ACC ± S (%) Times (s)
Yale	89.39 ± 2.85 0.126	91.44 ± 1.58 **0.101**	90.54 ± 2.01 0.262	92.23 ± 1.29 0.136	91.54 ± 1.07 0.135	**93.12 ± 1.71** 0.492
ORL	87.72 ± 1.53 1.169	90.87 ± 0.52 **0.483**	90.41 ± 0.78 3.064	92.45 ± 0.67 0.529	92.32 ± 1.16 1.338	**93.25 ± 0.46** 2.695
USPS	98.76 ± 0.18 118.729	98.83 ± 0.69 17.536	98.23 ± 0.24 134.438	99.65 ± 0.68 **6.795**	99.23 ± 0.42 358.368	**99.89 ± 0.89** 355.762
MNIST	89.61 ± 0.58 8.723	90.66 ± 0.74 1.237	89.83 ± 0.75 41.656	91.26 ± 1.13 **0.868**	90.88 ± 0.14 14.258	**91.53 ± 0.56** 14.973



(**a**) Yale



(**b**) ORL



(**c**) USPS

(**d**) MINST

Figure 6. Examples of four high-dimensional image datasets.

6. Conclusions

In this paper, we have proposed FTELM and CL_1-FTELM. FTELM not only inherits the advantages of TELM but also takes full account of the statistical information of samples, so as to further improve the classification performance of the classifier. Specifically, when there is no noise in the data or the ratio of noise is very small, our FTELM algorithm can deal with the classification problem very well, not only time-saving but also with high classification accuracy. CL_1-FTELM further improves the anti-noise ability of the model by replacing the L_2-norm and hinge loss in FTELM with capped L_1-norm. It not only utilizes the distribution information of the data but also improves the anti-noise ability of the model. Furthermore, we have designed two algorithms to solve the problems of FTELM and CL_1-FTELM. In addition, we present two theorems to prove the convergence of our CL_1-FTELM. However, in terms of computational cost, FTELM is better than CL_1-FTELM to some extent. Therefore, in future work, we will propose some new tricks to accelerate the

computation of the CL_1-FTELM. In addition, trying to extend FTELM and CL_1-FTELM from supervised learning to semi-supervised learning framework is also a future research focus.

Author Contributions: Z.X., conceptualization, methodology, validation, investigation, project administration, writing—original draft. L.C., methodology, software, validation, formal analysis, investigation, data curation, writing—original draft. All authors have read and agreed to the published version of the manuscript.

Funding: The authors wish to acknowledge the financial support of the National Nature Science Youth Foundation of China (No. 61907012), the Start-up Funds of Scientific Research for Personnel Introduced by North Minzu University (No. 2019KYQD41), the Special project of North Minzu University (No. FWNX01), the Basic Research Plan of Key Scientific Research Projects of Colleges and Universities in Henan Province (No. 19A120005), the Construction Project of First-Class Disciplines in Ningxia Higher Education (NXYLXK2017B09), the Young Talent Cultivation Project of North Minzu University (No. 2021KYQD23), the Natural Science Foundation of Ningxia Provincial of China (No. 2022A0950), the Fundamental Research Funds for the Central Universities (No. 2022XYZSX03).

Informed Consent Statement: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement: The UCI machine learning repository is available at "http://archive.ics.uci.edu/ml/datasets.php (accessed on 15 February 2023)". The image data are available at "http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html (accessed on 15 February 2023)".

Acknowledgments: The authors thank the anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: A new learning scheme of feedforward neural networks. In Proceedings of the 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541), Budapest, Hungary, 25–29 July 2004; Volume 2, pp. 985–990.
2. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]
3. Huang, G.B.; Chen, Y.Q.; Babri, H.A. Classification ability of single hidden layer feedforward neural networks. *IEEE Trans. Neural Networks* **2000**, *11*, 799–801. [CrossRef] [PubMed]
4. Chen, X.; Cui, B. Efficient modeling of fiber optic gyroscope drift using improved EEMD and extreme learning machine. *Signal Process.* **2016**, *128*, 1–7.
5. Xia, M.; Zhang, Y.; Weng, L.; Ye, X. Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs. *Knowl.-Based Syst.* **2012**, *36*, 253–259. [CrossRef]
6. Yang, J.; Xie, S.; Yoon, S.; Park, D.; Fang, Z.; Yang, S. Fingerprint matching based on extreme learning machine. *Neural Comput. Appl.* **2013**, *22*, 435–445. [CrossRef]
7. Rasheed, Z.; Rangwala, H. Metagenomic Taxonomic Classification Using Extreme Learning Machines. *J. Bioinform. Comput. Biol.* **2012**, *10 5*, 1250015. [CrossRef]
8. Zou, Q.Y.; Wang, X.J.; Zhou, C.J.; Zhang, Q. The memory degradation based online sequential extreme learning machine. *Neurocomputing* **2018**, *275*, 2864–2879.
9. Fu, Y.; Wu, Q.; Liu, K.; Gao, H. Feature Selection Methods for Extreme Learning Machines. *Axioms* **2022**, *11*, 444. [CrossRef]
10. Liu, Q.; He, Q.; Shi, Z. Extreme support vector machine classifier. In Proceedings of the Advances in Knowledge Discovery and Data Mining: 12th Pacific-Asia Conference, PAKDD 2008, Osaka, Japan, 20–23 May 2008; pp. 222–233.
11. Frénay, B.; Verleysen, M. Using SVMs with randomised feature spaces: An extreme learning approach. In Proceedings of the 18th European Symposium on Artificial Neural Networks, ESANN 2010, Bruges, Belgium, 28–30 April 2010.
12. Huang, G.B.; Ding, X.; Zhou, H. Optimization method based extreme learning machine for classification. *Neurocomputing* **2010**, *74*, 155–163. [CrossRef]
13. Khemchandani, R.; Chandra, S. Twin support vector machines for pattern classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 905–910.
14. Wan, Y.; Song, S.; Huang, G.; Li, S. Twin extreme learning machines for pattern classification. *Neurocomputing* **2017**, *260*, 235–244. [CrossRef]
15. Shen, J.; Ma, J. Sparse Twin Extreme Learning Machine with ε -Insensitive Zone Pinball Loss. *IEEE Access* **2019**, *7*, 112067–112078. [CrossRef]

16. Yuan, C.; Yang, L. Robust twin extreme learning machines with correntropy-based metric. *Knowl.-Based Syst.* **2021**, *214*, 106707. [CrossRef]
17. Anand, P.; Bharti, A.; Rastogi, R. Time efficient variants of Twin Extreme Learning Machine. *Intell. Syst. Appl.* **2023**, *17*, 200169. [CrossRef]
18. Ma, J.; Yu, G. A generalized adaptive robust distance metric driven smooth regularization learning framework for pattern recognition. *Signal Process.* **2023**, *211*, 109102. [CrossRef]
19. Ma, J.; Wen, Y.; Yang, L. Fisher-regularized supervised and semi-supervised extreme learning machine. *Knowl. Inf. Syst.* **2020**, *62*, 3995–4027. [CrossRef]
20. Gao, S.; Ye, Q.; Ye, N. 1-Norm least squares twin support vector machines. *Neurocomputing* **2011**, *74*, 3590–3597. [CrossRef]
21. Yan, H.; Ye, Q.L.; Zhang, T.A.; Yu, D.J. Efficient and robust TWSVM classifier based on L1-norm distance metric for pattern classification. In Proceedings of the 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), Nanjing, China, 26–29 November 2017; pp. 436–441.
22. Ye, Q.; Yang, J.; Liu, F.; Zhao, C.; Ye, N.; Yin, T. L1-norm distance linear discriminant analysis based on an effective iterative algorithm. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *28*, 114–129. [CrossRef]
23. Wu, Q.; Wang, F.; An, Y.; Li, K. L-1-Norm Robust Regularized Extreme Learning Machine with Asymmetric C-Loss for Regression. *Axioms* **2023**, *12*, 204. [CrossRef]
24. Wu, M.J.; Liu, J.X.; Gao, Y.L.; Kong, X.Z.; Feng, C.M. Feature selection and clustering via robust graph-laplacian PCA based on capped L 1-norm. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017; pp. 1741–1745.
25. Nie, F.; Huang, H.; Cai, X.; Ding, C. Efficient and robust feature selection via joint L2, 1-norms minimization. *Adv. Neural Inf. Process. Syst.* **2010**, *23*, 1813–1821.
26. Ma, J.; Yang, L.; Sun, Q. Capped L1-norm distance metric-based fast robust twin bounded support vector machine. *Neurocomputing* **2020**, *412*, 295–311. [CrossRef]
27. Jiang, W.; Nie, F.; Huang, H. Robust Dictionary Learning with Capped L1-Norm. In Proceedings of the 24th International Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015; pp. 3590–3596.
28. Nie, F.; Huo, Z.; Huang, H. Joint Capped Norms Minimization for Robust Matrix Recovery. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 2557–2563.
29. Wang, C.; Ye, Q.; Luo, P.; Ye, N.; Fu, L. Robust capped L1-norm twin support vector machine. *Neural Netw.* **2019**, *114*, 47–59. [CrossRef]
30. Pal, A.; Khemchandani, R.R.n. Learning TWSVM using Privilege Information. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 1548–1554.
31. Li, Y.; Sun, H.; Yan, W.; Cui, Q. R-CTSVM+: Robust capped L1-norm twin support vector machine with privileged information. *Inf. Sci.* **2021**, *574*, 12–32. [CrossRef]
32. Mangasarian, O.; Musicant, D. Successive overrelaxation for support vector machines. *IEEE Trans. Neural Netw.* **1999**, *10*, 1032–1037. [CrossRef] [PubMed]
33. Luo, Z.Q.; Tseng, P. Error bounds and convergence analysis of feasible descent methods: A general approach. *Ann. Oper. Res.* **1993**, *46*, 157–178. [CrossRef]
34. Yang, Y.; Xue, Z.; Ma, J.; Chang, X. Robust projection twin extreme learning machines with capped L1-norm distance metric. *Neurocomputing* **2023**, *517*, 229–242. [CrossRef]