Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 9(11): 31444–31469.
DOI: 10.3934/math.20241514
Received: 12 September 2024
Revised: 18 October 2024
Accepted: 28 October 2024
Published: 05 November 2024

Research article

A novel adaptive safe semi-supervised learning framework for pattern extraction and classification

Jun Ma*, **Junjie Li and Jiachen Sun**

School of Mathematics and Information Sciences, North Minzu University, Yinchuan Ningxia 750021, China

* **Correspondence:** Email: jun_ma1990@num.edu.cn.
Junjie Li & Jiachen Sun are co-second authors.

Abstract: Manifold regularization semi-supervised learning is a powerful graph-based semi-supervised learning method. However, the performance of semi-supervised learning methods based on manifold regularization depends to some extent on the quality of the manifold graph and unlabeled samples. Intuitively speaking, the quality of the graph directly affects the final classification performance of the model. In response to the above problems, this paper first proposed an adaptive safety semi-supervised learning framework. The framework implements the weight assignment of the self-similarity graph during the model learning process. In order to adapt to the learning needs, accelerate the learning speed, and avoid the impact of the curse of dimensionality, the framework also optimizes the features of each sample point through an automatic weighting mechanism to extract effective features and eliminate redundant information in the learning task. In addition, the framework defines an adaptive risk measurement mechanism for the uncertainty and potential risks of unlabeled samples to determine the degree of risk of unlabeled samples. Finally, a new adaptive safe semi-supervised extreme learning machine was proposed. Comprehensive experimental results across various class imbalance scenarios demonstrated that our proposed method outperforms other methods in terms of classification accuracy, and other critical performance metrics.

Keywords: semi-supervised learning; adaptive learning; risk degree; feature extraction; manifold regularization
Mathematics Subject Classification: 68T10, 91C20

1. Introduction

Semi-supervised learning (SSL) has emerged as a remarkable framework, achieving notable success in both theoretical and applied domains over the past decade, as evidenced in [1, 2]. A pivotal factor contributing to its prevalence lies in the often arduous and expensive process of acquiring labeled samples, contrasted with the relative ease and cost-effectiveness of collecting unlabeled data in numerous practical scenarios. SSL ingeniously leverages a diverse array of assumptions, including smoothness, clustering, and the manifold assumption, to forge connections between labeled and unlabeled data instances [1, 2]. Among these, the manifold assumption stands out as one of the most prevalent and influential, as highlighted in [1]. For instance, Belkin et al. [1] introduced the Laplacian regularized least squares (Lap-RLS) and support vector machines (Lap-SVM) algorithms, empirically demonstrating that the manifold regularization approach adeptly harnesses the rich information latent in unlabeled samples, thereby enhancing the overall learning performance.

It is universally acknowledged that the construction of a manifold graph is paramount to the efficacy of manifold regularization (MR). A well-designed graph, capable of facilitating subsequent classification tasks, can significantly enhance classification performance [3]. Conversely, an inadequately constructed graph may fail to contribute or even detract from classification accuracy. Notably, the graph is typically predefined and remains static throughout the learning process, posing a challenge as assessing its performance beforehand is virtually impossible. Consequently, parameter tuning within the manifold graph becomes crucial, yet in the context of semi-supervised learning with scarce label information, parameter selection remains an unresolved challenge. Building an optimal graph prior to classification is exceedingly difficult, further complicating the graph construction process for MR. Recently, some excellent graph learning methods have been proposed [4–6]. For example, Kang et al. [4] proposed an innovative and robust graph learning scheme, which effectively addresses the challenges of real-world noisy data by dynamically eliminating noise and errors present in the raw data.

To date, most advancements in MR have focused on optimizing regularization parameters or enhancing MR efficiency, with limited attention given to graph construction. Recent studies have cautioned that unlabeled samples may harbor risks and potentially compromise SSL performance [7–9]. This limitation underscores the need for a safe semi-supervised learning (SaSSL) approach that guarantees performance no worse than its supervised learning (SL) counterpart using solely labeled samples [10, 11]. In recent years, numerous innovative SaSSL methods have been introduced [12–20]. Among them, the semi-supervised extreme learning machine (SS-ELM), a novel single hidden layer feed-forward network (SLFN) algorithm proposed by Huang et al. [21], stands out. However, SS-ELM lacks robust safety mechanisms when leveraging unlabeled samples, as their inclusion can occasionally diminish its performance. To address this issue, She et al. [22] introduced the safe semi-supervised extreme learning machine (Safe-SSELM). Experimental findings underscore the robustness of Safe-SSELM, with its performance rarely falling significantly below that of ELM utilizing only labeled samples. This development represents a significant step forward in ensuring the safe and effective utilization of unlabeled data in semi-supervised learning frameworks.

Semi-supervised learning has garnered significant attention and widespread application in the realm of machine learning, with numerous remarkable algorithms emerging, particularly graph-based approaches that have yielded impressive results [23, 24]. For example, Xu et al. proposed a new tensor-based semi-supervised classifier, namely the graph-embedded low-rank tensor learning machine (GELRTLM). By implementing the multi-sensor signal fusion strategy, GELRTLM is effectively applied to mechanical diagnosis tasks and deeply explored. Among these, manifold regularization semi-supervised learning stands out as a highly effective graph-based method. Nevertheless, the performance of such methods is inherently tied to the quality of the manifold graph and the unlabeled samples. Specifically, the graph's quality exerts a direct influence on the model's ultimate classification performance. To address these challenges, this paper introduces an innovative adaptive safety semi-supervised learning framework. This framework innovatively incorporates self-adaptive manifold graph weights into the model learning process, enabling it to dynamically adjust to the learning dynamics. Furthermore, to expedite the learning process and mitigate the impact of dimensionality curses, the framework employs automatic weighting to learn the salient features of each sample point, extracting essential information while discarding redundant features that are detrimental to the learning task. Moreover, the framework introduces an adaptive risk assessment mechanism that quantifies the uncertainty and potential risks associated with unlabeled samples, thereby determining their risk level. This mechanism ensures that the learning process is guided by a nuanced understanding of the data's inherent risks. Based on this framework, we present a specific model tailored for semi-supervised classification tasks: The adaptive safe semi-supervised extreme learning machine (ASSELM). Experimental evaluations conducted on diverse datasets demonstrate the efficacy and robustness of the proposed ASSELM algorithm.

The structure of the remainder of this paper is as follows. In Section 2 provides a concise overview of related work, encompassing ELM and SS-ELM, to establish the context and position of our work within the broader research landscape. In Section 3, we delve into the specifics of our proposed algorithm, outlining its key components and novel contributions in detail. In Section 5 presents the experimental outcomes, accompanied by a thorough analysis of the results, showcasing the effectiveness and advantages of our approach. Finally, Section 6 concludes the paper by summarizing our key findings and outlining directions for future research.

2. Background

2.1. ELM

The extreme learning machine (ELM) method stands out as a formidable learning tool in the realms of machine learning and pattern recognition [25, 26]. By randomly initializing the input weights and biases of its hidden layer, ELM boasts a streamlined structure, minimal computational overhead, and remarkable versatility when compared to conventional neural network algorithms. Furthermore, ELM transcends the limitations inherent in traditional neural networks, such as the proclivity towards local minima, imprecise learning rates, and sluggish convergence rates, thereby offering a more robust and efficient learning paradigm.

Let $\mathcal{T}_l = \{\boldsymbol{x}_i, y_i\}_{i=1}^{l}$ represent the training set, where l denotes the total number of training samples, with each $x_i \in \mathbb{R}^n$ representing an input vector, and $y_i \in \{-1, +1\}(i = 1, \ldots, l)$ indicating the corresponding binary label. Assuming there is a hidden layer comprising L neurons, the output function of the ELM can be expressed as follows:

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{\beta} \tag{2.1}$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_L]^T$ denotes the vector of output weights connecting the hidden layer, consisting of L nodes, to the output node. The vector $\boldsymbol{Y} = [y_1, y_2, \ldots, y_l]^T$ represents the target outputs for the training samples. Additionally, $\boldsymbol{H}$ signifies the hidden layer output matrix, which encapsulates the activations of the hidden layer nodes for the given input samples:

$$\boldsymbol{H} = \begin{bmatrix} h_1(x_1) & \cdots & h_L(x_1) \\ \vdots & \vdots & \vdots \\ h_1(x_l) & \cdots & h_L(x_l) \end{bmatrix}$$

where $h_i(x) = G(\mathbf{a_i}, b_i, \mathbf{x}) = \mathbf{a_i} \cdot \mathbf{x} + b_i$ for $i = 1, \ldots, L$, represents the activation function of the i-th hidden node. Here, $\boldsymbol{a_i}$ and b_i are the input weights and bias, respectively, of the i-th hidden node. Both $\boldsymbol{a_i}$ and b_i can be randomly generated from a continuous probability distribution, embodying the essence of ELM's simplicity and efficiency in terms of model initialization.

The regularization ELM framework can be formulated as:

$$\min_{\beta} \Gamma(\beta) = C||\boldsymbol{H}\beta - \boldsymbol{Y}||^2 + ||\beta||^2 \tag{2.2}$$

where C serves as a penalty coefficient for the training errors, balancing the trade-off between minimizing the training error and the complexity of the model. Subsequently, the output weight vector β is derived based on the Moore-Penrose pseudoinverse principle, ensuring an optimal solution that satisfies the given constraints.

Thus, we can obtain the output weight vector $\boldsymbol{\beta}$ via

$$\boldsymbol{\beta}^* = \begin{cases} (\boldsymbol{H}^T\boldsymbol{H} + \frac{\boldsymbol{I}_L}{C})^{-1}\boldsymbol{H}^T\boldsymbol{Y}, & l \geq L, \\ \boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{H}^T + \frac{\boldsymbol{I}_l}{C})^{-1}\boldsymbol{Y}, & l \leq L, \end{cases} \tag{2.3}$$

where $\boldsymbol{I}_L$ is an identity matrix of dimension L and $\boldsymbol{I}_l$ is an identity matrix of dimension l.

2.2. SS-ELM

While ELMs have garnered significant popularity across various domains, their primary application has been confined to supervised learning tasks, particularly classification and regression, thereby limiting their versatility [25, 26]. In practical scenarios, acquiring labeled data samples is often a challenging and costly endeavor, whereas gathering abundant unlabeled samples is comparatively straightforward and economical [1]. To address the limitation of supervised ELMs in leveraging unlabeled data, Huang et al. introduced the semi-supervised ELM (SS-ELM), which incorporates manifold regularization, enabling the utilization of both labeled and unlabeled samples for enhanced learning performance.

Let $\mathcal{T} = \mathcal{T}_l \cup \mathcal{T}_u = \{x_i, y_i\}_{i=1}^{l} \cup \{x_i\}_{i=l+1}^{l+u}$ represent the semi-supervised learning training dataset, where $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, +1\}$ for labeled samples, $\mathcal{T}_l$ denotes labeled sample set with l, and $\mathcal{T}_u$ denotes an unlabeled sample set with u; hence, the total number of samples is $n = l + u$. Assuming a general decision function f, the overarching SS-ELM learning framework can be formulated as the following optimization problem:

$$\min_{\boldsymbol{\beta}} \Psi(\boldsymbol{\beta}) = C||\boldsymbol{H}_l\boldsymbol{\beta} - \boldsymbol{Y}||^2 + ||\boldsymbol{\beta}||^2 + \lambda Tr(\boldsymbol{\beta}^T\boldsymbol{H}_n^T\boldsymbol{L}\boldsymbol{H}_n\boldsymbol{\beta}) \tag{2.4}$$

where $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ represents the graph Laplacian, constructed using both labeled and unlabeled samples. Here, $\boldsymbol{D}$ is a diagonal matrix, and $\boldsymbol{W}$ denotes the similarity matrix that captures the relationships between sample. The parameters C and λ serve as regularization coefficients, balancing the trade-off between minimizing the empirical risk and the complexity of the model, respectively. The notation $Tr(\cdot)$ denotes the trace of a matrix, which is a measure of the sum of the diagonal elements of the matrix.

Thus, we have

$$\boldsymbol{\beta}^* = \begin{cases} (\boldsymbol{I}_l + \boldsymbol{H}_l^T \boldsymbol{C}\boldsymbol{H}_l + \lambda \boldsymbol{H}_n^T \boldsymbol{L}\boldsymbol{H}_n)^{-1} \boldsymbol{H}_l^T \boldsymbol{C}\boldsymbol{Y}, & n \geq L \\ \boldsymbol{H}_l^T (\boldsymbol{I}_n + \boldsymbol{C}\boldsymbol{H}_l \boldsymbol{H}_l^T + \lambda \boldsymbol{L}\boldsymbol{H}_n \boldsymbol{H}_n^T)^{-1} \boldsymbol{C}\boldsymbol{Y}, & n \leq L \end{cases} \tag{2.5}$$

where $\boldsymbol{I}_l$ represents an identity matrix of dimension l, and $\boldsymbol{C}$ is a diagonal matrix with entries defined as $\boldsymbol{C}_{jj} = \frac{C}{l_{t_j}}$, where l_{t_j} denotes the count of training samples belonging to class j, with j ranging from 1 to l. Additionally, $\boldsymbol{I}_n$ is an identity matrix of dimension n, where $n = l+u$ is the total number of samples including both labeled and unlabeled instances.

3. Adaptive safe semi-supervised learning framework

In this section, we introduce the adaptive safe semi-supervised learning framework. The schematic diagram illustrating the proposed methodology is shown in Figure 1.

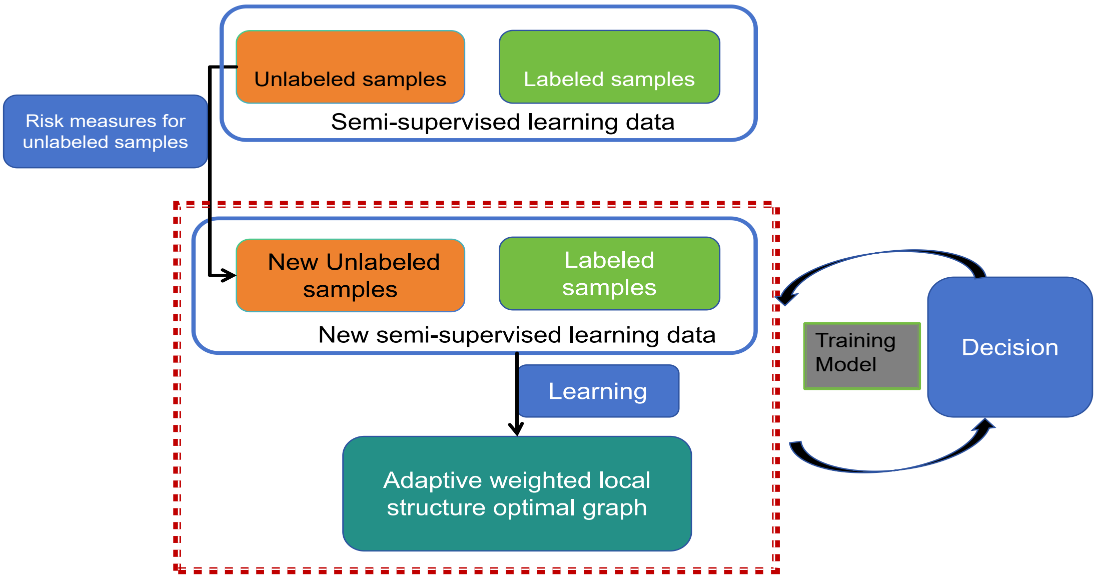

Unlabeled samples
Labeled samples
Semi-supervised learning data
Risk measures for unlabeled samples
New Unlabeled samples
Labeled samples
New semi-supervised learning data
Learning
Adaptive weighted local structure optimal graph
Training Model
Decision

Figure 1. The schematic diagram illustrating the proposed methodology.

3.1. Adaptive feature weighted local structure optimal graph

In manifold learning, the exploitation of local structural information, which possesses advantages over global structural considerations, has driven the development of numerous semi-supervised learning approaches rooted in the manifold hypothesis [27–29]. However, a pivotal limitation lies in the reliance of these methods on a fixed similarity matrix, which determines the weight map and, consequently, impacts the subsequent learning and decision-making processes to a significant extent [30–32]. Specifically, the manifold is preconstructed prior to classification and remains static throughout the learning phase, with the similarity matrix derived solely from the original data. This rigidity results in suboptimal graph learning as the original data often contains noise and redundant information, thereby hindering the effectiveness of the learned graph [33, 34].

In recent years, extensive research has underscored the paramount importance of graph quality in influencing the generalization performance of algorithms. Notably, the selection of the number of nearest neighbors plays a pivotal role in determining the ultimate performance, yet traditional methods often rely on suboptimal approaches such as k-nearest neighbors or ε-nearest neighbors to construct graphs. A more intuitive assumption is that the proximity between samples should be inversely proportional to their distance, implying that closer samples should have greater weights. Additionally, in graph-based semi-supervised learning (SSL), the manifold graph is formulated based on both relevant and irrelevant features of the samples. To this end, constructing the similarity matrix from a carefully selected set of primary features has the potential to yield a superior similarity matrix. This, in turn, can enhance the quality of the graph and subsequently boost the classification performance of the model. By focusing on the most informative features, we can mitigate the impact of noise and redundant information, leading to more effective graph representation and improved learning outcomes.

Drawing upon the aforementioned analysis, this paper introduces an innovative approach by proposing an adaptive weighted local structure optimal graph that builds upon previous research. This model aims to dynamically learn the optimal mapping tailored to the given sample set. The specific framework of this model is outlined as follows:

$$\begin{array}{ll}\min\limits_{\Theta,W} & \Upsilon(\Theta, W) = \sum_{i,j=1}^{n}\left(\|\Theta x_i - \Theta x_j\|_2^2 w_{ij} + \lambda w_{ij}^2\right) \\ \text{s.t.} & w_i^T\mathbf{1} = 1, 0 \le w_{ij} \le 1 \\ & \mathbf{1}^T\theta = 1, \theta \ge 0 \\ & \Theta = diag(\theta) \\ & rank(L_W) = n - c \end{array} \tag{3.1}$$

where λ denotes the regularization parameter, which necessitates careful adjustment. x_i represents the i-th sample, while w_{ij} signifies the similarity between the i-th and j-th samples within the similarity matrix W, which is subject to learning. Θ represents the sparse automatic weighting matrix that is also a target of the learning process. The inclusion of the regularization term serves to prevent the emergence of trivial solutions; in the absence of this term, the optimal solution would trivially assign a probability of 1 to the adjacency of the two closest samples. Furthermore, the rank constraint imposes a condition such that if $rank(L_W) = n - c$, then the similarity matrix W inherently comprises c connected components. Consequently, the similarity matrix W derived from solving the optimization problem (3.1) precisely contains c connected components, enhancing its capability to capture more precise local structural information.

Pursuant to Ky Fan's seminal theorem,

$$\sum_{i=1}^{c}\sigma_i(L_W) = \min_{F^TF=\boldsymbol{I}} Tr(F^T L_W F) \tag{3.2}$$

where $\boldsymbol{I}$ is the identity matrix.

As a consequence, the aforementioned optimization problem (3.1) can be formulated as

$$\begin{array}{ll} \min\limits_{\Theta,W,F} & \Upsilon(\Theta, W, F) = \left(\sum_{i,j=1}^{n}(\|\Theta x_i - \Theta x_j\|_2^2 w_{ij} + \lambda w_{ij}^2)\right) + \alpha Tr(F^T L_W F) \\ \text{s.t.} & w_i^T \mathbf{1} = 1, 0 \leq w_{ij} \leq 1 \\ & \mathbf{1}^T \theta = 1, \theta \geq 0 \\ & \Theta = diag(\theta) \\ & F^T F = \boldsymbol{I} \end{array} \tag{3.3}$$

where $L_W = D - \frac{(W^T+W)}{2}$ represents the Laplacian matrix, $D \in \mathbb{R}^{n\times n}$ denotes a diagonal matrix, with its diagonal elements being the sum of $\sum_j \frac{(w_{ij}+w_{ji})}{2}$ over all j.

3.2. Adaptive risk measurement mechanism for unlabeled samples

In the context of semi-supervised learning, unlabeled samples are typically regarded as benign and not detrimental to the overall performance. However, recent studies have revealed that these unlabelled samples possess dual properties, namely harmfulness and usefulness [7, 8]. In this section, we propose a safety mechanism aimed at evaluating the risk associated with unlabeled samples, which enhances the efficacy of semi-supervised learning algorithms to some extent. Notably, the risk attributed to unlabeled samples is often fixed, remaining constant throughout the learning phase. Nonetheless, the impact of this risk varies, influencing the performance of semi-supervised classifiers differently.

To address this, we introduce an adaptive safety mechanism designed to harness unlabeled samples more effectively. This mechanism operates on two main principles: (1) it enables the semi-supervised classifier to utilize all unlabeled samples while constraining their predictions to align closely with those of the supervised classifier, thereby mitigating the risk posed by unlabeled data; (2) it adaptively assigns distinct risk levels to each unlabeled sample, ensuring that samples deemed safe exhibit lower risk levels compared to those identified as risky. Consequently, the safety mechanism establishes a safety-driven trade-off between supervised and semi-supervised learning. In this manner, the adaptive risk assessment for unlabeled samples proposed herein integrates aspects of both supervised and semi-supervised learning, with the trade-off factor incorporated into the objective function of the semi-supervised learning framework.

The primary objective of this section is to develop an adaptive security mechanism designed to evaluate the security of each unlabeled sample and assign varying degrees of security accordingly. This evaluation is achieved through the application of an entropy maximization criterion. The details are presented as follows:

Definition 1. *Let $f(x)$ denote a semi-supervised classifier and $g(x)$ denote a supervised classifier. Thus, the adaptive risk measure mechanism for unlabelled samples can be defined as:*

$$\begin{array}{ll} \min\limits_{r_j} & \Xi(r_j) = \sum_{j=l+1}^{n} r_j\|f(x_j) - g(x_j)\|^2 + \sum_{j=l+1}^{n} r_j \ln(r_j) \\ s.t. & \sum_{j=l+1}^{n} r_j = 1, \\ & 0 < r_i \leq 1, \forall i = l+1, \ldots, n, \end{array} \tag{3.4}$$

where r_j describes the degree of safety of the unlabelled samples. The first term in the (3.4) *objective function represents the risk of the semi-supervised and supervised classifier trade-offs for the different unlabeled samples. The second term represents the regularisation term whose role is to prevent the emergence of a tame solution.*

Remark 1. *For unlabeled samples, the degree of safety remains uncertain. This section aims to quantify the risk associated with these unlabeled samples in a specific manner. It is widely recognized that the values of random variables are inherently uncertain. Prior to conducting the randomized test, only the probability distribution of the values is available; however, following the test, the values become known with certainty, thereby eliminating uncertainty. Through this randomized trial, information is acquired, and the amount of this information is precisely equivalent to the entropy of the random variable. Consequently, we can utilize entropy as an informative measure. Information entropy, also referred to as Shannon entropy, serves to reflect the level of disorder (or orderliness) within a system: The more ordered a system is, the lower its information entropy, and vice versa. In the context of unlabeled samples, a safer unlabeled sample corresponds to a lower level of risk, while a less secure unlabeled sample is indicative of higher risk.*

3.3. Adaptive safety semi-supervised learning framework

Given the comprehensive dataset $\mathcal{T} = \mathcal{T}_l \cup \mathcal{T}_u = \{\boldsymbol{x}_i, y_i\}_{i=1}^{l} \cup \{\boldsymbol{x}_i\}_{i=l+1}^{l+u}$, where $x_i \in \mathbb{R}^n$ represents the feature vectors and $y_i \in \{-1, +1\}$ denotes the corresponding binary labels for the labeled subset $\mathcal{T}_l$ comprising l instances. Conversely, $\mathcal{T}_u$ signifies the unlabeled subset containing u instances, for which the labels are unknown. Notably, in the realm of fully supervised learning, the unlabeled subset vanishes, i.e., $u = 0$.

Drawing inspiration from the prevalent regularized term in semi-supervised learning frameworks and adhering to the principle of structural risk minimization, we propose an innovative adaptive safety semi-supervised learning framework. This framework is designed to leverage both labeled and unlabeled data in an efficient and robust manner, as outlined in the subsequent sections.

$$
\begin{array}{ll}
\min\limits_{\Theta,W,F,R} & \sum_{i=1}^{l}(f(x_i) - y_j)^2 + \lambda_1\|f\|_{\mathcal{H}}^2 + \Upsilon(\Theta, W, F) + \lambda_2\Xi(r_i) \\
\text{s.t.} & w_i^T\mathbf{1} = 1, 0 \le w_{ij} \le 1 \\
 & \mathbf{1}^T\theta = 1, \theta \ge 0 \\
 & \theta = diag(\theta), F^TF = \boldsymbol{I} \\
 & \sum_{j=l+1}^{n} r_j = 1, \\
 & 0 < r_i \le 1, \forall i = l+1, \ldots, n,
\end{array}
\tag{3.5}
$$

where λ_1 and λ_2 are the regularization parameters. r_j describes the security of the unlabelled sample x_j. Within this framework, the flow graphs and their corresponding parameters undergo optimization during the learning process, rather than being pre-determined or statically defined. This dynamic approach ensures that the model adapts to the intrinsic characteristics of the data. Furthermore, the safety level of individual unlabeled samples is meticulously calculated and assigned in an adaptive manner, thereby enhancing the robustness and accuracy of the learning process.

(1) The first component of (3.5) encapsulates the empirical risk, quantifying the extent to which the model aligns with the training samples. It serves as a metric for assessing the goodness-of-fit between the model predictions and the observed data.

(2) The second term in (3.5) represents the structural risk, which aims to ensure the model's generalization capability and mitigate the risk of overfitting. By incorporating this term, the model is encouraged to learn patterns that generalize well to unseen data.

(3) The third term of (3.5) constitutes the joint regularization term, which leverages both discriminative information and the local geometric properties of new samples to enhance classification performance. In joint regularization (JR), samples residing in close proximity within the data manifold are encouraged to share the same class label if they belong to the same class, or to possess distinct labels otherwise. This approach harnesses the inherent structure of the data to further refine the classification boundaries.

(4) Lastly, the fourth term of (3.5) introduces risk-based regularization, which governs the balance between supervised and semi-supervised learning. The risk degrees play a pivotal role in determining the manner in which unlabeled samples are utilized during the learning process. By adjusting this term, the model can dynamically allocate weights to the labeled and unlabeled data, optimizing the overall learning strategy.

4. Adaptive safe semi-supervised extreme learning machine

4.1. AS3ELM

In this section, we introduce a novel semi-supervised extreme learning algorithm tailored for pattern classification, grounded in the adaptive safe semi-supervised learning framework. We term this innovative approach the adaptive safe semi-supervised extreme learning machine (AS3ELM). The specific formulation of this model is outlined as follows:

$$\begin{array}{ll}\min\limits_{\beta, w_{ij}, r_j, \theta_i} & \sum_{i=1}^{l}((h(x_i)\beta - y_j)^2 + \|\beta\|^2 + \Upsilon(\Theta, W, F) + \lambda_2 \Xi(r_i) \\ \text{s.t.} & w_i^T \mathbf{1} = 1, 0 \leq w_{ij} \leq 1 \\ & \mathbf{1}^T \theta = 1, \theta \geq 0 \\ & \Theta = diag(\theta), F^T F = \boldsymbol{I} \\ & \sum_{j=l+1}^{n} r_j = 1, \\ & 0 < r_i \leq 1, \forall i = l+1, \ldots, n. \end{array} \tag{4.1}$$

In order to solve the above problem, the model is solved by alternating iterative methods as follows:
Step 1. Fix variables w_{ij}, r_j, and θ_i, find output weights $\boldsymbol{\beta}$, and obtain the S3ELM:

$$\begin{aligned}\min_{\boldsymbol{\beta}} \Gamma(\boldsymbol{\beta}) = {} & C\|\boldsymbol{H}_l\boldsymbol{\beta} - \boldsymbol{Y}\|^2 + \|\boldsymbol{\beta}\|^2 + \lambda_1 Tr(\boldsymbol{\beta}^T \boldsymbol{H}_n^T \boldsymbol{L} \\ & \boldsymbol{H}_n\boldsymbol{\beta}) + \lambda_2 \sum_{j=l+1}^{l+u} r_j \|f(x_j) - g(x_j)\|^2. \end{aligned} \tag{4.2}$$

Within the AS3ELM framework, C, λ_1 and λ_2 serve as regularization parameters, governing the complexity and learning behavior of the model. The matrix $\boldsymbol{H}_l$ represents the hidden layer output for the labeled samples, while $\boldsymbol{H}_n$ encompasses the hidden layer output for the entire dataset, comprising both labeled and unlabeled samples. The Laplacian matrix $\boldsymbol{L}$, computed across all samples, captures the geometric relationships and intrinsic manifold structure within the data. The first three terms of the AS3ELM objective function jointly define the semi-supervised classifier, where λ_1 and λ_2 facilitate the integration of structural and joint regularization, respectively. This collaboration enables the model to leverage both labeled and unlabeled data effectively, enhancing its classification performance. The last term, controlled by C, meticulously balances the influence of ELM and SS-ELM, allowing for a smooth transition between fully supervised and semi-supervised learning modalities. Herein, $f(x)$ symbolizes the semi-supervised classifier instantiated through ELM, leveraging both labeled and unlabeled data. Conversely, $g(x)$ denotes the supervised classifier realized by SS-ELM, primarily relying on labeled data. The interplay between $f(x)$ and $g(x)$ within the AS3ELM framework underscores the adaptability and versatility of the proposed algorithm in handling complex pattern classification tasks.

A further simple mathematical collation gives

$$\begin{aligned} \min_{\boldsymbol{\beta}} \Gamma(\boldsymbol{\beta}) &= C\|\boldsymbol{H}_l\boldsymbol{\beta} - \boldsymbol{Y}\|^2 + \|\boldsymbol{\beta}\|^2 + \lambda_1 Tr(\boldsymbol{\beta}^T \boldsymbol{H}_n^T \boldsymbol{L}\boldsymbol{H}_n\boldsymbol{\beta}) \\ &+ \lambda_2(\boldsymbol{H}_u\boldsymbol{\beta} - \boldsymbol{H}_u\boldsymbol{\beta}_{ELM})^T \boldsymbol{R}(\boldsymbol{H}_u\boldsymbol{\beta} - \boldsymbol{H}_u\boldsymbol{\beta}_{ELM}). \end{aligned} \tag{4.3}$$

In the context of the AS3ELM framework, $\boldsymbol{\beta}_{ELM}$ represents the optimal solution derived from the ELM optimization process. This optimal solution encapsulates the weights that best map the hidden layer representations of the labeled data to their corresponding output labels. Additionally, $\boldsymbol{H}_u$ denotes the output matrix of the hidden layer specifically for the unlabeled samples, providing a rich feature space that can be leveraged in a semi-supervised manner. Furthermore, $\boldsymbol{R}$ is introduced as a diagonal matrix, playing a pivotal role in the regularization strategy employed within the AS3ELM algorithm. By incorporating $\boldsymbol{R}$ into the optimization objective, the algorithm is able to impose specific constraints or weights on the unlabeled samples, thereby enhancing the robustness and generalization capability of the resulting semi-supervised classifier. The careful design and utilization of R contribute significantly to the adaptive and safe nature of the AS3ELM approach.

The above problem (4.3) with respect to $\boldsymbol{\beta}$ is obtained by taking the partial derivative:

$$\begin{aligned} \frac{\partial \Gamma}{\partial \boldsymbol{\beta}} &= \boldsymbol{\beta} + \boldsymbol{H}_l^T \boldsymbol{C}(\boldsymbol{H}_l\boldsymbol{\beta} - \boldsymbol{Y}) + \lambda_1(\boldsymbol{H}_n^T \boldsymbol{L}\boldsymbol{H}_n\boldsymbol{\beta}) \\ &+ \lambda_2 \boldsymbol{H}_u^T \boldsymbol{R}(\boldsymbol{H}_u\boldsymbol{\beta} - \boldsymbol{H}_u\boldsymbol{\beta}_{ELM}). \end{aligned} \tag{4.4}$$

By setting (4.4) to 0, the optimal solution for $\boldsymbol{\beta}$ can be obtained as

$$\boldsymbol{\beta}^* = (\boldsymbol{H}_l^T \boldsymbol{C}\boldsymbol{H}_l + \boldsymbol{I} + \lambda_1 \boldsymbol{H}_n^T \boldsymbol{L}_l \boldsymbol{H}_n + \lambda_2 \boldsymbol{H}_u^T \boldsymbol{R}\boldsymbol{H}_u)^{-1} \cdot (\boldsymbol{H}_l^T \boldsymbol{C}\boldsymbol{Y} + \lambda_2 \boldsymbol{H}_u^T \boldsymbol{R}\boldsymbol{H}_u\boldsymbol{\beta}_{ELM}) \tag{4.5}$$

where $\boldsymbol{I}_l$ represents an identity matrix of dimension l, and $\boldsymbol{C}$ is a diagonal matrix with entries defined as $\boldsymbol{C}_{jj} = \frac{C}{l_{t_j}}$, where l_{t_j} denotes the count of training samples belonging to class j, with j ranging from 1 to l. Additionally, $\boldsymbol{I}_n$ is an identity matrix of dimension n, where $n = l + u$ is the total number of samples including both labeled and unlabeled instances.

Given a test set $\boldsymbol{X}_{new}$, we first meticulously compute its corresponding hidden layer output matrix, denoted as $\boldsymbol{H}_{new}$. This step transforms the input features of the test samples into a high-dimensional space, where they can be effectively classified using the learned decision boundary. Subsequently, we utilize the optimal solution obtained from the AS3ELM framework, denoted as $\boldsymbol{\beta}^*$, to derive the prediction results. Specifically, the prediction for each test sample in $\boldsymbol{X}_{new}$ is given by:

$$\boldsymbol{Y} = \boldsymbol{H}_{new}\boldsymbol{\beta}^*. \tag{4.6}$$

This formulation encapsulates the essence of the AS3ELM approach, leveraging both labeled and unlabeled data during the training phase to construct a robust classifier that can accurately predict the labels of unseen test samples.

Step 2. Fix β, w_{ij}, and θ_i to optimise r_j. Thus, the following optimization problem can be obtained:

$$\begin{array}{ll} \min\limits_{r_j} & \Xi(r_i) = \lambda_2 \sum_{j=l+1}^{n} r_j \|f(x_j) - g(x_j)\|^2 \\ & + \sum_{j=l+1}^{n} r_j \ln(r_j) \\ \text{s.t.} & \sum_{j=l+1}^{n} r_j = 1, \\ & 0 < r_i \leq 1, \forall i = l+1, \ldots, n. \end{array} \tag{4.7}$$

The Lagrangian function corresponding to the optimization problem (4.7) is

$$L = \lambda_2 \sum_{j=l+1}^{n} r_j \|f(x_j) - g(x_j)\|^2 + \sum_{j=l+1}^{n} r_j \ln(r_j) - \alpha_i (\sum_{j=l+1}^{n} r_j - 1). \tag{4.8}$$

The partial derivative of the above Lagrangian function L with respect to r_j gives

$$\frac{\partial L}{\partial r_j} = \lambda_2 \|f(x_j) - g(x_j)\|^2 + (1 + \ln r_j) - \alpha_i. \tag{4.9}$$

Therefore

$$r_j = \exp^{\alpha_i - \lambda_2 \|f(x_j) - g(x_j)\|^2}. \tag{4.10}$$

Also, $\sum_{j=l+1}^{n} r_j = 1$, and therefore

$$\exp^{\alpha_i} = \sum_{j=l+1}^{n} \exp^{\lambda_2 \|f(x_j) - g(x_j)\|^2}. \tag{4.11}$$

From (4.10) and (4.11), we have

$$r_j = \frac{\exp^{-\lambda_2 \|f(x_j) - g(x_j)\|^2}}{\sum_{j=l+1}^{n} \exp^{-\lambda_2 \|f(x_j) - g(x_j)\|^2}}. \tag{4.12}$$

As derived from (4.12), the discrepancy between $f(x_j)$ and $g(x_j)$ serves as an indicator of the safety level of the unlabeled sample x_j. Specifically, when the difference $f(x_j) - g(x_j)$ is minimal, it implies that the predictions made by the semi-supervised classifier $f(x)$ and the supervised classifier $g(x)$ are in close agreement for sample x_j. Consequently, x_j is likely to be a safe sample, and its corresponding safety score r_j should be high. In such cases, unlabeled samples with high safety scores exert a more significant influence on enhancing the performance of the semi-supervised learning process compared to those deemed risky. Conversely, if the difference $f(x_j) - g(x_j)$ is substantial, it indicates a potential disagreement between the predictions of the two classifiers for sample x_j. This discrepancy suggests that x_j may not be a safe sample, and hence, its safety score r_j should be low. By assigning lower weights to such risky unlabeled samples, the prediction for unlabeled data is effectively biased toward the supervised learning algorithm's prediction, thereby mitigating the potential risks introduced by uncertain or outlier unlabeled samples.

Step 3. Fixed β, w_{ij}m and r_j to optimize θ_i.

$$\begin{array}{ll} \min\limits_{\Theta} & \sum_{i,j=1}^{n} \|\Theta x_i - \Theta x_j\|^2 w_{ij} \\ \text{s.t.} & \mathbf{1}^T \theta = 1, \theta \geq 0 \\ & \Theta = diag(\theta). \end{array} \tag{4.13}$$

Further, the optimization problem (4.13) can be rewritten as

$$\begin{array}{ll} \min\limits_{\Theta} & Tr(\Theta^T X^T L_W X\Theta) \\ \text{s.t.} & \mathbf{1}^T\theta = 1, \theta \geq 0 \\ & \Theta = diag(\theta). \end{array} \tag{4.14}$$

Let $M = X^T L_W X$, and then the question (4.14) can be further written as follows:

$$\begin{array}{ll} \min\limits_{\Theta} & Tr(\Theta^T M\Theta) \\ \text{s.t.} & \mathbf{1}^T\theta = 1, \theta \geq 0 \\ & \Theta = diag(\theta). \end{array} \tag{4.15}$$

Let the i diagonal element of the matrix M be m_{ii}, and set $q_i = m_{ii}$. The above problem (4.15) can be simplified to

$$\begin{array}{ll} \min\limits_{\theta_i} & \sum_{i=1}^{d} \theta_i^2 q_i \\ \text{s.t.} & \mathbf{1}^T\theta = 1, \theta \geq 0. \end{array} \tag{4.16}$$

Assume that the matrix Q is a diagonal matrix and the diagonal elements are q_i. Therefore, the problem (4.16) can be written in vector form as:

$$\begin{array}{ll} \min\limits_{\theta} & \theta^T Q\theta \\ \text{s.t.} & \mathbf{1}^T\theta = 1, \theta \geq 0. \end{array} \tag{4.17}$$

The Lagrangian function of the above problem (4.17) can be written as

$$L(\theta, \rho) = \theta^T Q\theta - \rho(\mathbf{1}^T\theta - 1) \tag{4.18}$$

where ρ is the Lagrangian multiplier.

The above problem (4.18) is equal to 0 with respect to θ, which gives

$$2Q\theta - \rho\mathbf{1} = 0. \tag{4.19}$$

Since $\mathbf{1}^T\theta = 1$, it follows that

$$\theta_i = \frac{1}{q_i \sum_{j=1}^{d} \frac{1}{q_j}}. \tag{4.20}$$

Step 4. Fix β, θ_i, and r_j optimize w_{ij}.

$$\begin{array}{ll} \min\limits_{W} & \sum_{i,j=1}^{n} \left(\|\Theta x_i - \Theta x_j\|^2 w_{ij} + \lambda w_{ij}^2 \right) + \alpha \sum_{i,j=1}^{n} \|f_i - f_j\|_2^2 w_{ij} \\ \text{s.t.} & w_i^T\mathbf{1} = 1, 0 \leq w_{ij} \leq 0. \end{array} \tag{4.21}$$

Let $d_{ij}^x = \|x_i - x_j\|^2$, $d_{ij}^{\theta x} = \|\theta x_i - \theta x_j\|^2$, and $d_{ij}^f = \|f_i - f_j\|^2$. Let $d_{ij} = d_{ij}^{\theta x} + \alpha d_{ij}^f$, and we denote $d_i \in \mathbb{R}^{n\times 1}$ as a vector by the jth element. Then, the above problem (4.21) is written as

$$\underbrace{\min_{w_i^T\mathbf{1}=1,0\leq w_{ij}\leq 0}\sum_{j=1}^{n}(\lambda w_{ij}^2+d_{ij}w_{ij})} \tag{4.22}$$

$$\Updownarrow$$

$$\underbrace{\min_{w_i^T\mathbf{1}=1,0\leq w_{ij}\leq 0}\sum_{j=1}^{n}\left(w_{ij}^2+\frac{1}{\lambda}d_{ij}w_{ij}+\frac{1}{4\lambda^2}d_{ij}^2\right)} \tag{4.23}$$

$$\Updownarrow$$

$$\underbrace{\min_{w_i^T\mathbf{1}=1,0\leq w_{ij}\leq 0}\|w_i+\frac{1}{2\lambda}d_i\|^2} \qquad . \tag{4.24}$$

For each i, the Lagrangian function for the problem (4.22) can be written as

$$L(w_i,\eta,\delta_i)=\frac{1}{2}\|w_i+\frac{1}{2\lambda}d_i\|^2-\eta(w_i^T\mathbf{1}-1)-\delta_i w_i \tag{4.25}$$

where η and δ_i are Lagrangian multipliers.

According to the KKT condition, the optimal solution w_i of the problem (4.21) can be expressed as

$$w_{ij}=\left(-\frac{d_{ij}}{2\lambda}+\eta\right)_+ \tag{4.26}$$

where $\eta=\left(\frac{1}{k}+\frac{1}{2k\lambda}\sum_{j=1}^{k}d_{ij}\right)$, k is the number of nearest neighbors, and $\lambda=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{2}d_{i,k+1}^x-\frac{1}{2}\sum_{j=1}^{k}d_{ij}^x\right)$.

Based on the above discussion, the following algorithm is give.

Algorithm 1 AS3ELM algorithm

```
Input: Input: Semi-supervised learning datasets $\mathcal{T} = \mathcal{T}_l \cup \mathcal{T}_u = \{x_i, y_i\}_{i=1}^{l} \cup \{x_i\}_{i=l+1}^{l+u}$ represent the semi-supervised learning training dataset, where $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, +1\}$ for labeled samples, $\mathcal{T}_l$ denotes the labeled samples set with $l$, and $\mathcal{T}_u$ denotes the unlabeled samples set with $u$;
    Parameters Set: Hidden nodes $L$, regularization parameters $\lambda_1$, $\lambda_2$ and $\alpha$; maximum number of iterations $T$. .
Output: Output: Weight vector $\beta$;
    Initialization: Set t=0, $\beta^0 = 0$, and random initialization parameters $w_j$ and $b_j$ $(j = 1, 2, \ldots, L)$;
    Compute $\boldsymbol{\beta}_{ELM}$ by having labeled samples;
    Fix the variables $w_{ij}$, $r_j$, and $\theta_i$, and find the output weight $\boldsymbol{\beta}$ by the Eq (4.5);
    Fix $\beta$, $w_{ij}$, and $\theta_i$, optimizing $r_j$ by Eq (4.12);
    Fix $\beta$, $w_{ij}$, and $r_j$, optimizing $\theta_i$ by Eq (4.20);
    Fix $\beta$, $\theta_i$, and $r_j$, optimizing $w_{ij}$ by the Eq (4.26);
    Termination criterion: When $t > T$, the program terminates;
    Output: Weight vector $\boldsymbol{\beta}^*$.
```

4.2. Computational complexity analysis

In our proposed algorithm, the overarching computational complexity is predominantly attributed to four key operations: Updating the matrix $\boldsymbol{R}$, refining the weight vector β, constructing the graph Laplacian L, and identifying the k-nearest neighbors (k-NN). Specifically, the primary computational bottleneck within a single iteration of the algorithm stems from the $O(L^3)$ computational load associated with updating the current weight vector β, where L denotes a relevant dimensionality. Regarding the updated matrix $\boldsymbol{R}$, its computational complexity scales as $O((l + u)^2)$, with l and u representing specific dimensions pertinent to the matrix's dimensions or indices. The construction of the graph Laplacian matrix L, which encapsulates the topological structure of the data, involves a complexity of $O((l + u)^2 \log(l + u))$, primarily due to the efficient implementation of algorithms for constructing sparse matrices. Furthermore, the computational requirement for determining the adjacency matrix leveraging the k-nearest neighbor approach is $O(l + u)^2$, highlighting the quadratic dependence on the combined dimensions l and u. This complexity underscores the significance of efficient k-NN search algorithms in minimizing the overall computational overhead, particularly for large-scale datasets.

Hence, the overall computational complexity of the AS3ELM algorithm can be approximately formulated as $O(T \cdot (L^3 + (l + u)^2 \log(l + u) + 2(l + u)^2)))$, where T denotes the number of iterations required for the algorithm to converge. Based on our experimental findings, a value of $T = 10$ has been empirically demonstrated to yield satisfactory performance, indicating that the algorithm can effectively converge within a reasonable number of iterations. This observation underscores the practical feasibility and efficiency of the AS3ELM algorithm, especially when dealing with moderate to large-scale datasets. Collectively, these computational complexities underscore the need for optimization strategies and efficient algorithmic designs to facilitate the practical implementation of our algorithm, particularly in the context of high-dimensional data and large-scale graph constructions.

5. Experiments

In order to evaluate the performance of the proposed AS3ELM, this section systematically compares AS3ELM with other good methods, including: ELM [25], SS-ELM [21], and SASSELM [22], where the MATLAB source code for the algorithms ELM and SS-ELM The source code is available in [21] [*]. The activation function of $1/(1 + \exp(-(\boldsymbol{w} \cdot \boldsymbol{x} + b)))$ ($\boldsymbol{w}$, b is randomly generated) was used in the ELM, SS-ELM, SASSELM, and AS3ELM algorithms. Also, to measure the classification performance of all algorithms, the conventional accuracy classification precision (ACC) was used:

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{5.1}$$

where TP and TN denote true positives and true negatives, respectively, and FN and FP denote false negatives and false positives, respectively. In addition, the learning time is also used to indicate the computational efficiency of each algorithm.

Parameters have a significant impact on the performance of a model. Therefore, it is necessary to choose the parameters of the algorithm in a reasonable way to improve the performance of the model. The parameters of the algorithm involved in this experiment were chosen in the following range.

[*] http://www.ntu.edu.sg/home/egbhuang/elm.html

- $C \in \{10^{-5}, 10^{-5}, \ldots, 10^{4}, 10^{5}\}$
- $\lambda_1 \in \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$
- $\lambda_2 \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$
- $\alpha \in \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$
- $L \in \{100, 200, 300, 400, 500, 600, 700, 800, 1000, 1500, 2000\}$

To ensure a rigorous and valid comparison among the various methods evaluated, this experiment employed a rigorous 10-fold cross-validation framework coupled with a grid search technique. This comprehensive approach aimed to obtain a robust estimate of the average classification accuracy, serving as the primary metric for assessing the performance of the algorithms under investigation. To further strengthen the reliability of our findings, this process was reiterated ten times, and the arithmetic mean of the results from these ten independent experiments was adopted as the definitive performance measure. Moreover, to mitigate potential biases and ensure the objectivity of our experimental outcomes, all datasets underwent a normalization process, transforming their values into the unified interval of $[0, 1]$. This standardization step facilitated a fair comparison among algorithms, as it eliminated any scaling-related disparities that might otherwise skew the results. All experiments were meticulously conducted on a dedicated personal computer, equipped with MATLAB 2014a software running on a Windows 10 operating system. The computational hardware consisted of an Intel(R) Core(TM) i7-8700 processor, clocked at 3.40 GHz, complemented by 16 GB of RAM, ensuring sufficient resources for the smooth execution of the experiments and the timely processing of the data.

5.1. Experimental results on UCI dataset

To rigorously assess the efficacy of the proposed methodology, we sourced nine benchmark datasets from the prestigious UCI Machine Learning Repository[†]. To ensure a fair and unbiased comparison, these datasets underwent a normalization process, scaling their features to a uniform range of $[0, 1]$. This standardization step facilitated the analysis and interpretation of the results, eliminating any potential biases introduced by varying scales of the features. The comprehensive details of these datasets, including their characteristics and dimensions, are systematically presented in Table 1.

Table 1. Information description of UCI datasets.

ID	Datasets	Samples	Dimension
1	Austra	690	14
2	Balance	576	4
3	Banknote	1372	4
4	Ionosp	350	34
5	Pima	768	8
6	QSAR	1055	41
7	Vote	432	16
8	WDBC	569	30
9	Wholesale	440	7

[†] http://archive.ics.uci.edu/ml/datasets.html

First, a meticulous experimental design was formulated to rigorously validate the classification performance of the method introduced in this chapter, utilizing nine diverse datasets sourced from the UCI Machine Learning Repository. This validation process involved varying the proportion of labeled samples across the datasets to assess the robustness and adaptability of the proposed method. Specifically, for each dataset, 70% of the samples from each class were randomly allocated to form a comprehensive training set, while the remaining 30% constituted the test set for evaluation purposes. Within the training sets, distinct labeling scenarios were created by labeling different proportions of samples, namely 10% and 30%, respectively. This approach enabled us to evaluate the sensitivity of the method to varying degrees of supervision during the learning phase. To ensure reproducibility and comprehensiveness, the outcomes of all experiments were reported in terms of the average classification accuracy (ACC), accompanied by the standard deviation (S), denoted as $ACC \pm S$. This notation provides insights into both the central tendency and the variability of the results, thereby facilitating a nuanced understanding of the method's performance. The comprehensive results from these exhaustive experiments are systematically tabulated in Tables 2 and 3, offering a clear and concise overview of the method's classification capabilities across various benchmark datasets and labeling scenarios.

As can be seen from Tables 2 and 3, the performance of all the algorithms improves as the number of labeled samples increases. In addition, it can be seen that the proposed method AS3ELM shows comparable performance in most of the datasets compared to the other algorithms. Specifically, SS-ELM outperforms ELM on the QSAR, Vote, and Wholesale datasets in 10% of cases. AS3ELM outperforms ELM in 10% of cases, except for QSAR and Wholesale. Furthermore, SS-ELM outperforms ELM in 10% of cases. In addition, the proposed algorithm AS3ELM achieves comparable results to SS-ELM in the case where SS-ELM outperforms ELM. This indicates that the security mechanism used in our algorithm is effective and reduces the risk of untagged samples. In contrast to the SASSELM method, the method obtains optimal graphs and security measures through adaptive graph construction and adaptive security measures, respectively. This brings the performance of our method in line with that of SASSELM on most datasets. In contrast to SS-ELM, the approach in this paper incorporates risk level to control the trade-off between supervised and semi-supervised learning. Supervised learning is performed on high-risk unlabeled data, and semi-supervised learning is performed on low-risk data.

Table 2. Learning results of the four algorithms on UCI datasets with 10% labeled samples.

	ELM	**SS-ELM**	**SASSELM**	**AS3ELM**
Datasets	ACC±S(%)	ACC±S(%)	ACC±S(%)	ACC± S(%)
Australian	83.97±2.34	86.36±2.49	88.57±2.77	89.74±2.12
Balance	94.55±2.09	96.42±2.01	96.37±2.09	96.75±1.78
Banknote	81.75±1.63	91.70±1.74	90.67±1.49	88.89±1.67
Ionosp	78.24±1.39	76.99±1.41	76.38±1.57	78.79±1.46
Pima	70.79±1.35	75.37±1.58	76.44±1.78	75.85±1.94
QSAR	85.87±1.42	81.25±1.51	83.94±1.82	77.72±1.37
Vote	95.00±2.26	92.11±2.71	94.44±2.11	96.15±2.78
WDBC	91.96±1.33	96.55±1.39	96.67±1.51	94.14±1.05
Wholesale	90.23±2.38	86.37±2.43	86.66±2.27	86.89±2.57

Table 3. Learning results of the four algorithms on UCI datasets with 30% labeled samples.

	ELM	**SS-ELM**	**SASSELM**	**AS3ELM**
Australian	84.12±2.08	88.16±2.51	89.42±2.68	89.74±2.09
Balance	94.48±2.17	97.00±2.13	97.33±2.51	98.95±1.55
Banknote	87.96±1.63	92.68±1.57	92.84±1.66	90.67±1.67
Ionosp	77.06±1.39	82.47±1.48	82.68±1.55	82.94±1.26
Pima	75.13±1.42	82.76±1.33	82.83±1.58	81.46±1.39
QSAR	85.87±1.35	87.51±1.33	87.79±1.41	86.36±1.34
Vote	95.71±2.32	95.38±2.72	97.26±2.59	98.04±2.56
WDBC	94.11±1.18	98.85±1.79	98.86±1.84	98.29±1.78
Wholesale	87.67±2.22	90.91±2.29	91.05±2.25	94.70±2.16

5.2. Image dataset experimental results

In order to better validate the performance of the proposed method, a series of experiments are conducted on four image datasets in this subsection. The descriptions and information of the four image datasets are as follows.

- ORL[‡]: 10 different images of each of the 40 different subjects. For some subjects, images were taken at different times, varying lighting, facial expression (eyes open/closed, smiling/not smiling), and facial details (glasses on/not wearing glasses). All photographs were taken against a dark uniform background with the subject in a frontal upright position (with some side shifts tolerated). A preview image of the face database is available. Each pixel has 256 grey levels. The database was used for a face recognition project in collaboration with the Speech, Vision and Robotics Group at the University of Cambridge Engineering Department.

[‡] http://www.cl.cam.ac.uk/Research/DTG/attarchive:pub/data/att_faces.tar.Z

- YaleB[§]: Yale contains 165 grayscale images of 15 people in GIF format. Each subject has 11 images, each with a different facial expression or configuration: central light, with glasses, happy, left light, without glasses, normal, right light, sad, sleepy, surprised, and blinking. For YaleB, we simply used the cropped images and resized them to 32 × 32 pixels. This dataset now has 38 individuals and about 64 near-frontal images of individuals under different lighting conditions.
- COIL20 [¶]: It contains 20 objects. The images of each object are taken at a distance of 5 degrees as the object is rotated on a turntable, giving 72 images per object. The size of each image is 32 × 32 pixels, with 256 gray levels per pixel. Thus, each image is represented by a 1024-dimensional vector. In this experiment, the first 10 objects in COIL20 are classified as 1 and the remaining objects are classified as 2.
- USPST [‖]: The USPST dataset is a collection of handwritten digits from the USPS postal system. Each digit image is represented by a grayscale value of 16 × 16 pixels. It was constructed in this experiment by grouping the first 5 digits into class 1 and the remaining digits into class 2.

This experiment tested the performance of the four algorithms involved with different proportions of labeled samples, with the proportion of labeled samples in the training set taken to be 10%, 30%, 50%, 70%, and 90%, respectively. All the experimental results are presented in Figure 2. As shown in Figure 2, the performance of all four algorithms tends to increase when the proportion of labelled samples increases. Furthermore, it can be seen that the overall performance of the proposed method is comparable to the other three algorithms on the four datasets, and some satisfactory results are obtained. Details of the above dataset are given in Table 4.

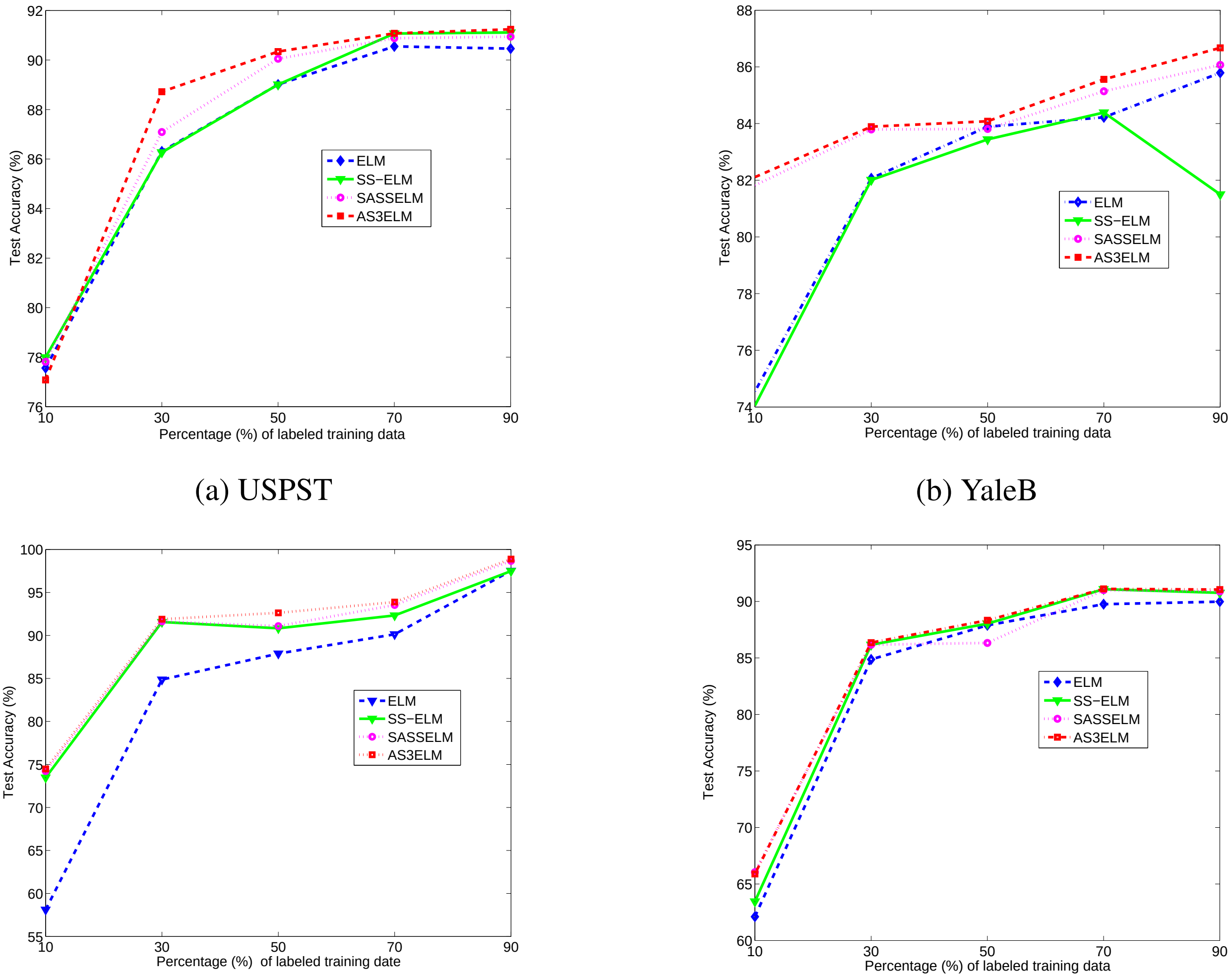
Table 4. Description of the dataset.

Datasets	Dimension	Samples	Class
USPST	256	2007	9
ORL	1024	400	40
YaleB	1024	2414	38
COIL20	1024	1440	20

[§] http://www.cad.zju.edu.cn/home/dengcai/Data/data.html

[¶] https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

[‖] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/multiclass.html



(a) USPST

(b) YaleB

(c) ORL

(d) COIL20

Figure 2. Comparison of the performance of the four algorithms on image datasets with different scales of labeled samples.

5.3. *Experimental results on large datasets*

To demonstrate and evaluate the effectiveness of our method, we conducted a systematic comparison with other mainstream semi-supervised classification techniques. We used the standard accuracy rate (ACC) as the metric to measure the classification efficiency of all algorithms. In our experiments, we randomly selected 30% of the samples from each NDC dataset (refer to the "Description of NDC Datasets" in Table 5) to establish the training set, while the remaining 70% constituted the test set. Furthermore, we divided the training set into labeled and unlabeled samples in a 2:8 proportion. We repeated the experiment ten times and computed the average classification accuracy. All the experimental results are presented in Table 6. Particularly, for certain algorithms, due to the shortage of computer memory, we halted the computation and substituted the results with NaN. From Table 6, it can be seen that our algorithm consistently achieved higher classification accuracy than the other algorithms in all instances. Obviously, due to memory limitations, the ELM and SS-ELM algorithms performed poorly on the NDC-31 and NDC-51 datasets.

Table 5. Description of NDC datasets.

Dataset	Samples	Features	Dataset	Samples	Features
NDC-5k	5000	32	NDC-10k	10000	32
NDC-11	100000	32	NDC-31	300000	32
NDC-51	500000	32			

Table 6. Experimental results on large datasets.

	ELM	**SS-ELM**	**SASSELM**	**AS3ELM**
Datasets	ACC %	ACC %	ACC %	ACC%
NDC-5k	82.01	84.45	85.19	86.67
NDC-10k	83.26	84.17	84.21	85.36
NDC-11	72.04	73.08	74.11	76.73
NDC-3l	NaN	NaN	71.48	73.21
NDC-51	NaN	NaN	65.03	67.18

5.4. Ablation study experiments and results

To verify the performance of the proposed method, we conducted an ablation study. Specifically, AS3ELM1: AS3ELM without adaptive feature weighted local structure optimal graph, and AS3ELM2: AS3ELM without adaptive risk measurement mechanism for unlabeled samples. The experimental results are presented in Table 7.

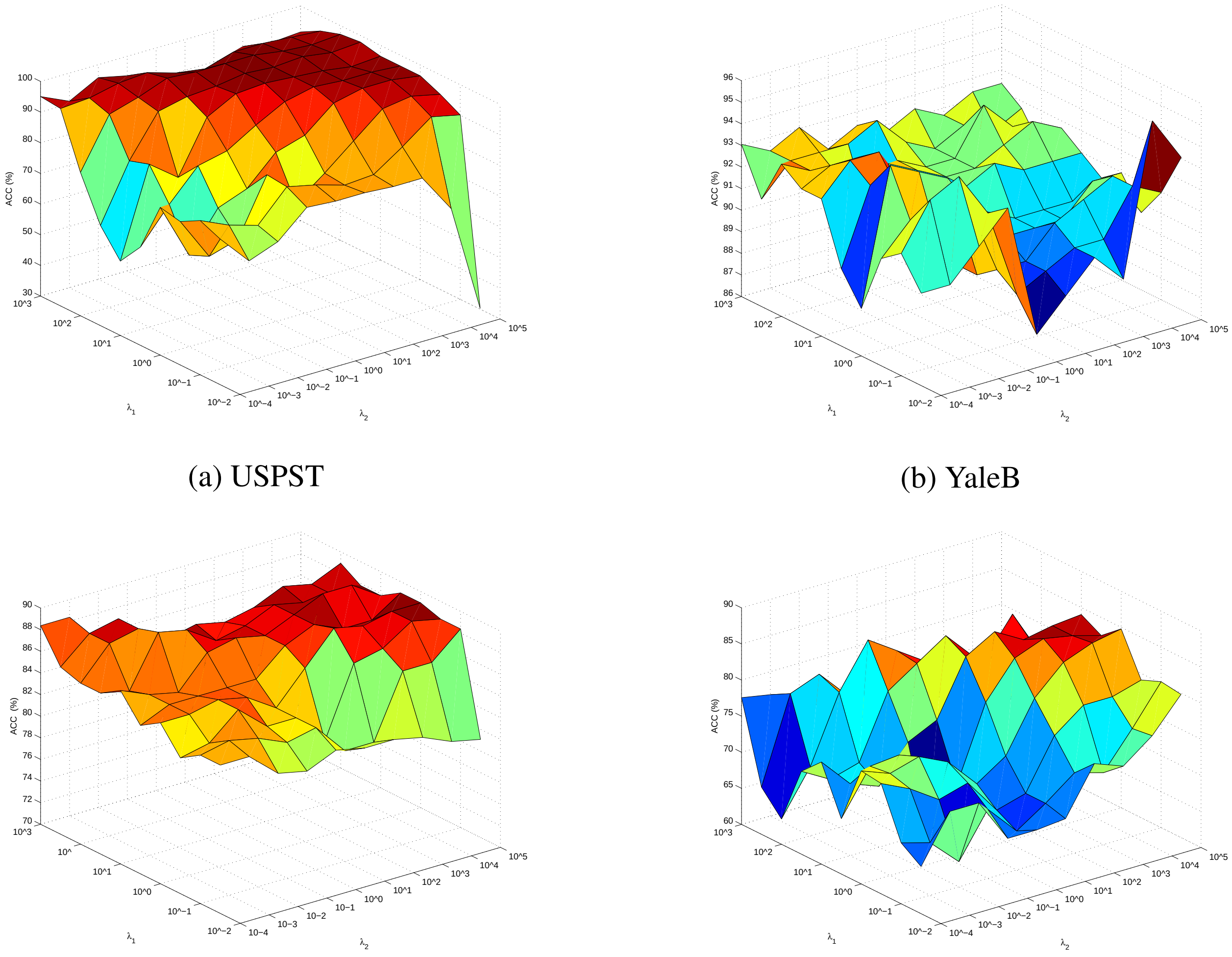
Table 7. Ablation study experiments and results.

	AS3ELM	**AS3ELM1**	**AS3ELM2**
Datasets	ACC %	ACC %	ACC %
NDC-5k	86.67	81.04	83.29
NDC-10k	85.36	79.58	81.36
NDC-11	76.73	69.33	71.05

Through Table 7, we can see that on all three datasets, AS3ELM achieves better classification accuracy than both AS3ELM1 and AS3ELM2. This observation is particularly noteworthy as it indicates a consistent performance advantage across diverse data environments. The results suggest that the proposed method not only enhances the model's ability to classify instances correctly but also implies an improvement in its generalization performance. Generalization performance refers to a model's capability to perform well on unseen data, which is crucial for practical applications where models are deployed in real-world scenarios. By outperforming its counterparts, AS3ELM demonstrates a robust learning mechanism that likely incorporates effective feature extraction or optimization techniques tailored for each datasets unique characteristics. Furthermore, this enhancement could be attributed to various factors inherent in the design of AS3ELM. For instance, it may utilize advanced algorithms or architectures that allow for more efficient processing of input features or leverage ensemble methods that combine multiple learning strategies effectively. Such improvements are essential for developing machine learning models capable of adapting to new challenges while maintaining high accuracy levels. In summary, the findings presented in Table 7 provide compelling evidence supporting the efficacy of the proposed method over existing alternatives within this study framework. This reinforces the notion that ongoing research into innovative modeling approaches can yield significant advancements in classification tasks across different domains and datasets.

5.5. Parameter sensitivity analysis

In general, parameters correspond to optimal output results. In order to investigate the effect of the parameters λ_1 and λ_1 on the performance of the AS3ELM algorithm, this section conducts a series of experiments on four image datasets. Empirical values of $L = 500$ were used for the hidden layer nodes, which were fixed in the experiments, and the parameters λ_1 and λ_1 were varied over the ranges $\lambda_1 \in \{10^{-}2, 10^{-}1, 10^1, 10^2, 10^3\}$ and $\lambda_2 \in \{10^{-}4, 10^{-}3, 10^{-}2, 10^{-}1, 10^0, 10^1, 10^2, 10^3, 10^4, 10^5\}$. The experimental results are presented in Figure 3. As shown in Figure 3, it can be seen that the proposed algorithm AS3ELM does not fluctuate much in classification accuracy under different parameters and is relatively stable. This indicates that the proposed method is insensitive to parameters and has good generalization performance. The graph in Figure 3(d) displays significant fluctuations, which can be attributed to the adaptive learning of these parameters during the process of evaluating the risk associated with unlabeled samples and constructing the most localized structural graph in our proposed method. This adaptability allows our model to dynamically adjust its parameters based on the characteristics of each sample, resulting in a more accurate representation of the underlying data distribution. By continuously updating and refining these parameters, our method is able to capture intricate patterns and relationships within the dataset. The fluctuations observed in the graph reflect this ongoing learning process as it iteratively refines its understanding of geometric distribution information from both labeled and unlabeled samples. This acquisition of geometric distribution information plays a crucial role in improving the performance of our proposed method. By incorporating this knowledge into our model's decision-making process, we are able to make more informed predictions about unseen data points. This not only enhances accuracy but also enables better generalization capabilities when applied to real-world scenarios. Furthermore, by constructing a localized structural graph that adapts to each sample's risk evaluation, we ensure that important local dependencies are captured effectively. This localization helps prevent overgeneralization or underestimation by considering specific neighborhood structures for each data point. In summary, through adaptive parameter learning and the construction of a localized structural graph, our proposed method acquires valuable geometric distribution information from both labeled and unlabeled samples. These fluctuations observed in Figure 3(d) represent an ongoing refinement process that ultimately leads to improved accuracy and generalization capabilities for predicting unseen data points.



(a) USPST

(b) YaleB

(c) ORL

(d) COIL20

Figure 3. Performance analysis of the algorithm AS3ELM under different parameters.

6. Conclusions

In this paper, we embark on the construction of an innovative adaptive feature-weighted local structure optimal graph. This graph not only facilitates adaptive learning of graph weights during both model training and prediction phases but also adeptly extracts pertinent features from individual data points through automatic weighting mechanisms. This process effectively prunes redundant features, thereby accelerating the learning process, mitigating the risk of dimensional catastrophe, and enhancing the overall efficiency of the learning task. Furthermore, we introduce an adaptive risk measure, which meticulously quantifies the degree of unlabeled risk in relation to the inherent uncertainty and potential hazards associated with unlabeled samples. This measure provides a nuanced understanding of the risk landscape, enabling more informed decision-making during the learning process. Subsequently, we develop and implement an adaptive safety semi-supervised learning framework tailored to a specific model. Within this framework, we propose an adaptive safety semi-supervised limit learning machine tailored for pattern classification tasks. This machine leverages an alternating iteration approach to solve the model, ensuring convergence and stability. Experimental evaluations conducted on diverse datasets demonstrate the competitiveness and efficacy of our proposed method in comparison to other related algorithms. However, it is acknowledged that in terms of time efficiency, our model does not exhibit a marked advantage over its counterparts. This limitation primarily stems from the alternating iteration strategy employed, which necessitates iterative updates of four variables, thereby increasing computational overhead. To address this challenge, future research endeavors should focus on devising efficient algorithms that can solve the model without compromising on classification accuracy. Such endeavors hold the promise of enhancing the practical applicability and scalability of our adaptive safety semi-supervised learning framework.

Author contributions

Jun Ma, Junjie Li, Jiachen Sun: The responsibilities included algorithm development, software creation, numerical example preparation, original draft writing, and review and editing of the manuscript. All authors contributed equally to this work and have read and approved the final version of the manuscript for publication.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Ningxia Provincial (No. 2024AAC05055), in part by the Fundamental Research Funds for the Central Universities of North Minzu University (No. 2023ZRLG01, No. 2021JCYJ107), in part by the National Natural Science Foundation of China (No. 62366001, No. 12361062), in part by the Key Research and Development Program of Ningxia (Introduction of Talents Project)(No. 2022BSB03046), and in part by the Natural Science Foundation of Ningxia Provincial (No. 2023AAC02053).

Conflict of interest

The authors declare that they have no conflict of interest and no relevant financial or non-financial interests to disclose.

References

1. M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, *J. Mach. Learn. Res.*, **7** (2006), 2399–2434.
2. O. Chapelle, B. Schlkopf, A. Zien, *Semi-supervised learning*, Handbook on Neural Information Processing, Springer Berlin Heidelberg, 2013.

3. Y. Wang, Y. Meng, Y. Li, S. Chen, Z. Fu, H. Xue, Semi-supervised manifold regularization with adaptive graph construction, *Pattern Recog. Lett.*, **98** (2017), 90–95. https://doi.org/10.1016/j.patrec.2017.09.004

4. Z. Kang, H. Pan, S. C. Hoi, Z. Xu, Robust graph learning from noisy data, *IEEE T. Cybernetics*, **50** (2019), 1833–1843. https://doi.org/10.1109/TCYB.2018.2887094

5. Z. Kang, C. Peng, Q. Cheng, X. Liu, X. Peng, Z. Xu, et al., Structured graph learning for clustering and semi-supervised classification, *Pattern Recog.*, **110** (2021), 107627. https://doi.org/10.1016/j.patcog.2020.107627

6. Z. Kang, X. Lu, Y. Lu, C. Peng, W. Chen, Z. Xu, Structure learning with similarity preserving, *Neural Networks*, **129** (2020), 138–148. https://doi.org/10.1016/j.neunet.2020.05.030

7. T. Yang, C. E. Priebe, The effect of model misspecification on semi-supervised classification, *IEEE T. Pattern Anal.*, **33** (2011), 2093–2103. https://doi.org/10.1109/TPAMI.2011.45

8. Y. F. Li, Z. H. Zhou, Towards making unlabeled data never hurt, *IEEE T. Pattern Anal.*, **37** (2015), 175–188. https://doi.org/10.1109/TPAMI.2014.2299812

9. Y. F. Li, Z. H. Zhou, *Improving semi-supervised support vector machines through unlabeled instances selection*, In: Proceedings of the AAAI Conference on Artificial Intelligence, **25** (2011), 386–391. https://doi.org/10.1609/aaai.v25i1.7920

10. Y. Wang, S. Chen, Z. H. Zhou, New semi-supervised classification method based on modified cluster assumption, *IEEE T. Neur. Net. Lear.*, **23** (2012), 689–702. https://doi.org/10.1109/TNNLS.2012.2186825

11. Y. Wang, S. Chen, Safety-aware semi-supervised classification, *IEEE T. Neur. Net. Lear.*, **24** (2013), 1763–1772. https://doi.org/10.1109/TNNLS.2013.2263512

12. M. Kawakita, J. Takeuchi, Safe semi-supervised learning based on weighted likelihood, *Neural Networks*, **53** (2014), 146–164. https://doi.org/10.1016/j.neunet.2014.01.016

13. H. T. Gan, Z. Z. Luo, M. Meng, Y. Ma, Q. She, A risk degree-based safe semi-supervised learning algorithm, *Int. J. Mach. Learn. Cyb.*, **7** (2015), 1–10. https://doi.org/10.1007/s13042-015-0416-8

14. H. T. Gan, Z. Luo, Y. Sun, X. Xi, N. Sang, R. Huang, Towards designing risk-based safe Laplacian regularized least squares, *Expert Syst. Appl.*, **45** (2016), 1–7. https://doi.org/10.1016/j.eswa.2015.09.017

15. H. T. Gan, Z. Li, Y. Fan, Z. Luo, Dual learning-based safe semi-supervised learning, *IEEE Access*, **6** (2017), 2615–2621. https://doi.org/10.1109/ACCESS.2017.2784406

16. H. T. Gan, Z. Li, W. Wu, Z. Luo, R. Huang, Safety-aware graph-based semi-supervised learning, *Expert Syst. Appl.*, **107** (2018), 243–254. https://doi.org/10.1016/j.eswa.2018.04.031

17. N. Sang, H. T. Gan, Y. Fan, W. Wu, Z. Yang, Adaptive safety degree-based safe semi-supervised learning, *Int. J. Mach. Learn. Cyb.*, **10** (2018), 1101–1108. https://doi.org/10.1007/s13042-018-0788-7

18. Y. Wang, Y. Meng, Z. Fu, H. Xue, Towards safe semi-supervised classification: Adjusted cluster assumption via clustering, *Neural Process. Lett.*, **46** (2017), 1031–1042. https://doi.org/10.1007/s11063-017-9607-5

19. H. T. Gan, G. Li, S. Xia, T. Wang, A hybrid safe semi-supervised learning method, *Expert Syst. Appl.*, **149** (2020), 1–9. https://doi.org/10.1016/j.eswa.2020.113295

20. Y. T. Li, J. T. Kwok, Z. H. Zhou, *Towards safe semi-supervised learning for multivariate performance measures*, In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16), AAAI Press, **30** (2016), 1816–1822. https://doi.org/10.1609/aaai.v30i1.10282

21. G. Huang, S. Song, J. N. D. Gupta, C. Wu, Semi-supervised and unsupervised extreme learning machines, *IEEE T. Cybernetics*, **44** (2017), 2405–2417. https://doi.org/10.1109/TCYB.2014.2307349

22. Q. She, B. Hu, H. Gan, Y. Fan, T. Nguyen, T. Potter, et al., Safe semi-supervised extreme learning machine for EEG signal classification, *IEEE Access*, **6** (2018), 49399–49407. https://doi.org/10.1109/ACCESS.2018.2868713

23. H. Xu, X. Wang, J. Huang, F. Zhang, F. Chu, Semi-supervised multi-sensor information fusion tailored graph embedded low-rank tensor learning machine under extremely low labeled rate, *Inform. Fusion*, **105** (2024), 102222. https://doi.org/10.1016/j.inffus.2023.102222

24. J. Huang, F. Zhang, B. Safaei, Z. Qin, F. Chu, The flexible tensor singular value decomposition and its applications in multisensor signal fusion processing, *Mech. Syst. Signal Pr.*, **220** (2024), 111662. https://doi.org/10.1016/j.ymssp.2024.111662

25. G. B. Huang, Q. Y. Zhu, C. K. Siew, Extreme learning machine: Theory and applications, *Neurocomputing*, **70** (2006), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

26. G. B. Huang, X. J. Ding, H. M. Zhou, Optimization method based extreme learning machine for classification, *Neurocomputing*, **74** (2010), 155–163. https://doi.org/10.1016/j.neucom.2010.02.019

27. Z. Liu, Z. Lai, W. Ou, K. Zhang, R. Zheng, Structured optimal graph based sparse feature extraction for semi-supervised learning, *Signal Process.*, **170** (2020), 107456. https://doi.org/10.1016/j.sigpro.2020.107456

28. M. Luo, F. Nie, X. Chang, Y. Yang, A. G. Hauptmann, Q. Zheng, Adaptive unsupervised feature selection with structure regularization, *IEEE T. Neur. Net. Lear.*, **29** (2018), 944–956. https://doi.org/10.1109/TNNLS.2017.2650978

29. J. S. Wu, M. X. Song, W. Min, J. H. Lai, W. S. Zheng, Joint adaptive manifold and embedding learning for unsupervised feature selection, *Pattern Recog.*, **112** (2020), 107742. https://doi.org/10.1016/j.patcog.2020.107742

30. F. Nie, W. Zhu, X. Li, Structured graph optimization for unsupervised feature selection, *IEEE T. Knowl. Data En.*, **33** (2019), 1210–1222. https://doi.org/10.1109/TKDE.2019.2937924

31. F. Nie, S. J. Shi, X. Li, Semi-supervised learning with auto-weighting feature and adaptive graph, *IEEE T. Knowl. Data En.*, **32** (2019), 1167–1178. https://doi.org/10.1109/TKDE.2019.2901853

32. Q. Li, L. Jing, J. Yu, *Adaptive graph constrained NMF for semi-supervised learning*, In: Iapr International Workshop on Partially Supervised Learning, Springer, Berlin, Heidelberg, 2013, 36–48. https://doi.org/10.1007/978-3-642-40705-5_4

33. Y. Yuan, X. Li, Q. Wang, F. Nie, A semi-supervised learning algorithm via adaptive Laplacian graph, *Neurocomputing*, **426** (2020), 162–173. https://doi.org/10.1016/j.neucom.2020.09.069

34. Z. Liu, K. Shi, K. Zhang, W. Ou, L. Wang, Discriminative sparse embedding based on adaptive graph for dimension reduction, *Eng. Appl. Artif. Intel.*, **94** (2020), 103758. https://doi.org/10.1016/j.engappai.2020.103758