Extreme Learning Machine for Regression and Multiclass Classification

Guang-Bin Huang, Senior Member, IEEE, Hongming Zhou, Xiaojian Ding, and Rui Zhang

Abstract—Due to the simplicity of their implementations, least square support vector machine (LS-SVM) and proximal support vector machine (PSVM) have been widely used in binary classification applications. The conventional LS-SVM and PSVM cannot be used in regression and multiclass classification applications directly, although variants of LS-SVM and PSVM have been proposed to handle such cases. This paper shows that both LS-SVM and PSVM can be simplified further and unified learning framework of LS-SVM, PSVM, and other regularization algorithms referred to extreme learning machine (ELM) can be built. ELM works for the “generalized” single-hidden-layer feedforward networks (SLFNs), but the hidden layer (or called feature mapping) in ELM need not be tuned. Such SLFNs include but are not limited to SVM, polynomial network, and the conventional feedforward neural networks. This paper shows the following: 1) ELM provides a unified learning platform with a widespread type of feature mappings and can be applied in regression and multiclass classification applications directly; 2) from the optimization method point of view, ELM has milder optimization constraints compared to LS-SVM and PSVM; 3) in theory, compared to ELM, LS-SVM and PSVM achieve suboptimal solutions and require higher computational complexity; and 4) in theory, ELM can approximate any target continuous function and classify any disjoint regions. As verified by the simulation results, ELM tends to have better scalability and achieve similar (for regression and binary class cases) or much better (for multiclass cases) generalization performance at much faster learning speed (up to thousands times) than traditional SVM and LS-SVM.

Index Terms—Extreme learning machine (ELM), feature mapping, kernel, least square support vector machine (LS-SVM), proximal support vector machine (PSVM), regularization network.

I. INTRODUCTION

In the past two decades, due to their surprising classification capability, support vector machine (SVM) [1] and its variants [2]–[4] have been extensively used in classification and regression applications. SVM has two main learning features: 1) In SVM, the training data are first mapped into a higher dimensional feature space through a nonlinear feature mapping function \( \phi(x) \), and 2) the standard optimization method is then used to find the solution of maximizing the separating margin of two different classes in this feature space while minimizing the training errors. With the introduction of the epsilon-insensitive loss function, the support vector method has been extended to solve regression problems [5].

As the training of SVMs involves a quadratic programming problem, the computational complexity of SVM training algorithms is usually intensive, which is at least quadratic with respect to the number of training examples. It is difficult to deal with large problems using single traditional SVMs [6]; instead, SVM mixtures can be used in large applications [6], [7].

Least square SVM (LS-SVM) [2] and proximal SVM (PSVM) [3] provide fast implementations of the traditional SVM. Both LS-SVM and PSVM use equality optimization constraints instead of inequalities from the traditional SVM, which results in a direct least square solution by avoiding quadratic programming.

SVM, LS-SVM, and PSVM are originally proposed for binary classification. Different methods have been proposed in order for them to be applied in multiclass classification problems. One-against-all (OAA) and one-against-one (OAO) methods are mainly used in the implementation of SVM in multiclass classification applications [8]. OAA-SVM consists of \( m \) SVMs, where \( m \) is the number of classes. The \( i \)th SVM is trained with all of the samples in the \( i \)th class with positive labels and all the other examples from the remaining \( m - 1 \) classes with negative labels. OAO-SVM consists of \( m(m - 1)/2 \) SVMs, where each is trained with the samples from two classes only. Some encoding schemes such as minimal output coding (MOC) [9] and Bayesian coding–decoding schemes [10] have been proposed to solve multiclass problems with LS-SVM. Each class is represented by a unique binary output codeword of \( m \) bits. \( m \) outputs are used in MOC-LS-SVM in order to scale up to \( 2^m \) classes [9]. Bayes’ rule-based LS-SVM uses \( m \) binary LS-SVM plug-in classifiers with its binary class probabilities inferred in a second step within the related probabilistic framework [10]. With the prior multiclass probabilities and the posterior binary class probabilities, Bayes’ rule is then applied \( m \) times to infer posterior multiclass probabilities [10]. Bayes’ rule and different coding scheme are used in PSVM for multiclass problems [11].

The decision functions of binary SVM, LS-SVM, and PSVM classifiers have the same form

\[
f(x) = \text{sign} \left( \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + b \right)
\]
where \( t_i \) is the corresponding target class label of the training data \( x_i \), \( \alpha_i \) is the Lagrange multiplier to be computed by the learning machines, and \( K(\mathbf{u}, \mathbf{v}) \) is a suitable kernel function to be given by users. From the network architecture point of view, SVM, LS-SVM, and PSVM can be considered as a specific type of single-hidden-layer feedforward network (SLFN) (the so-called support vector network termed by Cortes and Vapnik [1]) where the output of the \( i \)th hidden node is \( K(\mathbf{x}, \mathbf{x}_i) \) and the output weight linking the \( i \)th hidden node to the output node is \( \alpha_i t_i \). The term bias \( b \) plays an important role in SVM, LS-SVM, and PSVM, which produces the equality optimization constraints in the dual optimization problems of these methods. For example, the only difference between LS-SVM and PSVM is on how to use the bias \( b \) in the optimization formula while they have the same optimization constraint, resulting in different least square solutions. No learning parameter in the hidden-layer output function (kernel) \( K(\mathbf{u}, \mathbf{v}) \) needs to be tuned by SVM, LS-SVM, and PSVM, although some user-specified parameter needs to be chosen \textit{a priori}.

Extreme learning machine (ELM) [12]–[16] studies a much wider type of “generalized” SLFNs whose hidden layer need not be tuned. ELM has been attracting the attentions from more and more researchers [17]–[22]. ELM was originally developed for the single-hidden-layer feedforward neural networks [12]–[16] and then extended to the “generalized” SLFNs which may not be neuron alike [15], [16]

\[
f(\mathbf{x}) = \mathbf{h}(\mathbf{x})\mathbf{\beta}
\]  

(2)

where \( \mathbf{h}(\mathbf{x}) \) is the hidden-layer output corresponding to the input sample \( \mathbf{x} \) and \( \mathbf{\beta} \) is the output weight vector between the hidden layer and the output layer. One of the salient features of ELM is that the hidden layer need not be tuned. Essentially, ELM originally proposes to apply random computational nodes in the hidden layer, which are independent of the training data. Different from traditional learning algorithms for a neural type of SLFNs [23], ELM aims to reach not only the smallest training error but also the smallest norm of output weights. ELM [12], [13] and its variants [14]–[16], [24]–[28] mainly focus on the regression applications. Latest development of ELM has shown some relationships between ELM and SVM [18], [19], [29].

Suykens and Vandewalle [30] described a training method for SLFNs which applies the hidden-layer output mapping as the feature mapping of SVM. However, the hidden-layer parameters need to be iteratively computed by solving an optimization problem (refer to the last paragraph in Section IV-A1 for details). As Suykens and Vandewalle stated in their work (see [30, p. 907]), the drawbacks of this method are the following: the high computational cost and larger number of parameters in the hidden layer. Liu et al. [18] and Fré Byzay and Verleysen [19] show that the ELM learning approach can be applied to SVRM directly by simply replacing RBF kernel with (random) ELM kernels and better generalization can be achieved. Different from the study of Suykens and Vandewalle [30] in which the hidden layer is parametric, the ELM hidden layer used in the studies of Liu et al. [18] and Fré Byzay and Verleysen [19] is nonparametric, and the hidden-layer parameters need not be tuned and can be fixed once randomly generated. Liu et al. [18] suggest to apply ELM kernel in SVMs and particularly study PSVM [3] with ELM kernel. Later, Frénay and Verleysen [19] show that the normalized ELM kernel can also be applied in the traditional SVM. Their proposed SVM with ELM kernel and the conventional SVM have the same optimization constraints (e.g., both inequality constraints and bias \( b \) are used). Recently, Huang et al. [29] further show the following: 1) SVM’s maximal separating margin property and the ELM’s minimal norm of output weight property are actually consistent, and with ELM framework, SVM’s maximal separating margin property and Barlett’s theory on feedforward neural networks remain consistent, and 2) compared to SVM, ELM requires fewer optimization constraints and results in simpler implementation, faster learning, and better generalization performance. However, similar to SVM, inequality optimization constraints are used in [29]. Huang et al. [29] use random kernels and discard the term bias \( b \) used in the conventional SVM. However, no direct relationship has so far been found between the original ELM implementation [12]–[16] and LS-SVM/PSVM. Whether feedforward neural networks, SVM, LS-SVM, and PSVM can be unified still remains open.

Different from the studies of Huang et al. [29], Liu et al. [18], and Frénay and Verleysen [19], this paper extends ELM to LS-SVM and PSVM and provides a unified solution for LS-SVM and PSVM under equality constraints. In particular, the following contributions have been made in this paper.

1) ELM was originally developed from feedforward neural networks [12]–[16]. Different from other ELM work in literature, this paper manages to extend ELM to kernel learning: It is shown that ELM can use a wide type of feature mappings (hidden-layer output functions), including random hidden nodes and kernels. With this extension, the unified ELM solution can be obtained for feedforward neural networks, RBF network, LS-SVM, and PSVM.

2) Furthermore, ELM, which is with higher scalability and less computational complexity, not only unifies different popular learning algorithms but also provides a unified solution to different practical applications (e.g., regression, binary, and multiclass classifications). Different variants of LS-SVM and SVM are required for different types of applications. ELM avoids such trivial and tedious situations faced by LS-SVM and SVM. In ELM method, all these applications can be resolved in one formula.

3) From the optimization method point of view, ELM and LS-SVM have the same optimization cost function; however, ELM has milder optimization constraints compared to LS-SVM and PSVM. As analyzed in this paper and further verified by simulation results over 36 wide types of data sets, compared to ELM, LS-SVM achieves suboptimal solutions (when the same kernels are used) and has higher computational complexity. As verified by simulations, the resultant ELM method can run much faster than LS-SVM. ELM with random hidden nodes can run even up to tens of thousands times faster than SVM and LS-SVM. Different from earlier ELM works which do not perform well in sparse data sets, the ELM method proposed in this paper can handle sparse data sets well.

4) This paper also shows that the proposed ELM method not only has universal approximation capability (of approximating any target continuous function) but also has classification capability (of classifying any disjoint regions).
II. BRIEF OF SVMs

This section briefs the conventional SVM [1] and its variants, namely, LS-SVM [2] and PSVM [3].

A. SVM

Cortes and Vapnik [1] study the relationship between SVM and multilayer feedforward neural networks and showed that SVM can be seen as a specific type of SLFNs, the so-called support vector networks. In 1962, Rosenblatt [31] suggested that multilayer feedforward neural networks (perceptrons) can be trained in a feature space $Z$ of the last hidden layer. In this feature space, a linear decision function is constructed

$$f(x) = \text{sign} \left( \sum_{i=1}^{L} \alpha_i z_i(x) \right)$$

where $z_i(x)$ is the output of the $i$th neuron in the last hidden layer of a perceptron. In order to find an alternative solution of $z_i(x)$, in 1995, Cortes and Vapnik [1] proposed the SVM which maps the data from the input space to some feature space $Z$ through some nonlinear mapping $\phi(x)$ chosen a priori. Constrained-optimization methods are then used to find the separating hyperplane which maximizes the separating margins of two different classes in the feature space.

Given a set of training data $(x_i, t_i)$, $i = 1, \ldots, N$, where $x_i \in \mathbb{R}^d$ and $t_i \in \{-1, 1\}$, due to the nonlinear separability of these training data in the input space, in most cases, one can map the training data $x_i$ from the input space to a feature space $Z$ through a nonlinear mapping $\phi : x_i \rightarrow \phi(x_i)$. The distance between two different classes in the feature space $Z$ is $2/\|\omega\|$. To maximize the separating margin and to minimize the training errors, $\xi_i$, is equivalent to

Minimize : $L_{P_{SVM}} = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i$

Subject to : $t_i (\omega \cdot \phi(x_i) + b) \geq 1 - \xi_i$, $i = 1, \ldots, N$

$\xi_i \geq 0$, $i = 1, \ldots, N$  \hspace{1cm} (4)

where $C$ is a user-specified parameter and provides a tradeoff between the distance of the separating margin and the training error.

Based on the Karush–Kuhn–Tucker (KKT) theorem [32], to train such an SVM is equivalent to solving the following dual optimization problem:

minimize : $L_{D_{SVM}} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} t_i t_j \alpha_i \alpha_j \phi(x_i) \cdot \phi(x_j)$

$$- \sum_{i=1}^{N} \alpha_i$$

subject to : $\sum_{i=1}^{N} t_i \alpha_i = 0$

$0 \leq \alpha_i \leq C$, $i = 1, \ldots, N$  \hspace{1cm} (5)

where each Lagrange multiplier $\alpha_i$ corresponds to a training sample $(x_i, t_i)$. Vectors $x_i$’s for which $t_i (\omega \cdot \phi(x_i) + b) = 1$ are termed support vectors [1].

Kernel functions $K(u, v) = \phi(u) \cdot \phi(v)$ are usually used in the implementation of SVM learning algorithm. In this case, we have

minimize : $L_{D_{SVM}} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} t_i t_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i$

subject to : $\sum_{i=1}^{N} t_i \alpha_i = 0$

$0 \leq \alpha_i \leq C$, $i = 1, \ldots, N$.  \hspace{1cm} (6)

The SVM kernel function $K(u, v)$ needs to satisfy Mercer’s condition [1]. The decision function of SVM is

$$f(x) = \text{sign} \left( \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + b \right)$$

where $N_s$ is the number of support vectors $x_i$’s.

B. LS-SVM

Suykens and Vandewalle [2] propose a least square version to SVM classifier. Instead of the inequality constraint (4) adopted in SVM, equality constraints are used in the LS-SVM [2]. Hence, by solving a set of linear equations instead of quadratic programming, one can implement the least square approach easily. LS-SVM is proven to have excellent generalization performance and low computational cost in many applications.

In LS-SVM, the classification problem is formulated as

Minimize : $L_{P_{LS-SVM}} = \frac{1}{2} \|w\|^2 + C \frac{1}{2} \sum_{i=1}^{N} \xi_i^2$

Subject to : $t_i (w \cdot \phi(x_i) + b) = 1 - \xi_i$, $i = 1, \ldots, N$.  \hspace{1cm} (8)

Based on the KKT theorem, to train such an LS-SVM is equivalent to solving the following dual optimization problem:

minimize : $L_{D_{LS-SVM}} = \frac{1}{2} \|w\|^2 + C \frac{1}{2} \sum_{i=1}^{N} \xi_i^2$

$$- \sum_{i=1}^{N} \alpha_i \left( t_i (w \cdot \phi(x_i) + b) - 1 + \xi_i \right).$$  \hspace{1cm} (9)

Different from Lagrange multipliers (5) in SVM, in LS-SVM, Lagrange multipliers $\alpha_i$’s can be either positive or negative due to the equality constraints used. Based on the KKT theorem, we can have the optimality conditions of (9) as follows:

$$\frac{\partial L_{D_{LS-SVM}}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i t_i \phi(x_i)$$  \hspace{1cm} (10a)
equations can be equivalently written as
\[ \text{SVM} \ [2], \ \text{PSVM} \ [3], \ \text{ELM} \ [29], \ \text{and TER-ELM} \ [22], \ \text{feature mappings} \ \phi \]

reflected in mathematical expressions that the inequality con-
which classifies data points depending on proximity to either
\[ L_{\text{LS-SVM}} = \frac{1}{2} (\mathbf{w} \cdot \mathbf{w} + b^2) + C \sum_{i=1}^{N} \xi_i^2 \]
Subject to : \[ t_i (\mathbf{w} \cdot \mathbf{x}_i + b) = 1 - \xi_i, \quad i = 1, \ldots, N. \] (14)

The corresponding dual optimization problem is
\[ L_{D_{\text{PSVM}}} = \frac{1}{2} (\mathbf{w} \cdot \mathbf{w} + b^2) + C \sum_{i=1}^{N} \xi_i^2 \]
- \[ \sum_{i=1}^{N} \alpha_i (t_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i). \] (15)

By applying KKT optimality conditions [similar to
(10a)–(10d)], we have
\[ \left( \frac{1}{C} + \Omega_{\text{PSVM}} + \mathbf{T T}^T \right) \alpha = \left( \frac{1}{C} + \mathbf{Z Z}^T + \mathbf{T T}^T \right) = \mathbf{1} \] (16)

where \( \mathbf{Z} = [t_1 x_1, \ldots, t_N x_N]^T \) and \( \Omega_{\text{PSVM}} = \mathbf{Z Z}^T. \)

Similar to LS-SVM, the training data \( \mathbf{x} \) can be mapped
from the input space into a feature space \( \phi : \mathbf{x} \to \phi(\mathbf{x}) \),
and one can obtain the nonlinear version of PSVM: \( \mathbf{Z} = [t_1 \phi(x_1), \ldots, t_N \phi(x_N)]^T. \) As feature mapping \( \phi(\mathbf{x}) \)
is usually unknown, Mercer’s conditions can be applied to
matrix \( \Omega_{\text{PSVM}} : \Omega_{\text{PSVM},ij} = t_i t_j \phi(x_i) \cdot \phi(x_j) = t_i t_j K(x_i, x_j), \) which is the same as LS-SVM’s kernel matrix \( \Omega_{\text{LS-SVM}} \). The decision function of PSVM classifier is
\[ f(\mathbf{x}) = \text{sign}(\sum_{i=1}^{N} \alpha_i t_i K(\mathbf{x}, x_i) + b). \]

III. PROPOSED CONSTRAINED-OPTIMIZATION-BASED ELM

ELM [12]–[14] was originally proposed for the single-
hidden-layer feedforward neural networks and was then ex-
tended to the generalized SLFNs where the hidden layer need
not be neuron alike [15], [16]. In ELM, the hidden layer need
not be tuned. The output function of ELM for generalized
SLFNs (take one output node case as an example) is
\[ f_L(\mathbf{x}) = \sum_{i=1}^{L} \beta_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x}) \beta \] (17)

where \( \beta = [\beta_1, \ldots, \beta_L]^T \) is the vector of the output weights
between the hidden layer of \( L \) nodes and the output node and \( \mathbf{h}(\mathbf{x}) = [h_1(x), \ldots, h_L(x)] \) is the output (row) vector of the
hidden layer with respect to the input \( x \). \( \mathbf{h}(\mathbf{x}) \) actually maps the
data from the \( d \)-dimensional input space to the \( L \)-dimensional
hidden-layer feature space (ELM feature space) \( \mathcal{H} \), and thus,
\( \mathbf{h}(\mathbf{x}) \) is indeed a feature mapping. For the binary classification
applications, the decision function of ELM is
\[ f_L(\mathbf{x}) = \text{sign}(\mathbf{h}(\mathbf{x}) \beta). \] (18)

Different from traditional learning algorithms [23], ELM
tends to reach not only the smallest training error but also
the smallest norm of output weights. According to Bartlett’s
theory [34], for feedforward neural networks reaching smaller
training error, the smaller the norms of weights are, the better
generalization performance the networks tend to have. We conjecture that this may be true to the generalized SLFNs where the hidden layer may not be neuron alike [15], [16]. ELM is to minimize the training error as well as the norm of the output weights [12], [13]

\[
\text{Minimize: } \|H\beta - T\|^2 \text{ and } \|\beta\|
\]  

where \( H \) is the hidden-layer output matrix

\[
H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} h_1(x_1) & \cdots & h_L(x_1) \\ \vdots & \ddots & \vdots \\ h_1(x_N) & \cdots & h_L(x_N) \end{bmatrix}.  
\]

Seen from (18), to minimize the norm of the output weights \( \|\beta\| \) is actually to maximize the distance of the separating margins of the two different classes in the ELM feature space: \( 2/\|\beta\| \).

The minimal norm least square method instead of the standard optimization method was used in the original implementation of ELM [12], [13]

\[
\beta = H^T T
\]

where \( H^T \) is the Moore–Penrose generalized inverse of matrix \( H \) [35], [36]. Different methods can be used to calculate the Moore–Penrose generalized inverse of a matrix: orthogonal projection method, orthogonalization method, iterative method, and singular value decomposition (SVD) [36]. The orthogonal projection method [36] can be used in two cases: when \( H^T H \) is nonsingular and \( H^T = (H^T H)^{-1} H^T \), or when \( HH^T \) is nonsingular and \( H^T = (HH^T)^{-1} \).

According to the ridge regression theory [37], one can add a positive value to the diagonal of \( H^T H \) or \( HH^T \); the resultant solution is stabler and tends to have better generalization performance. Toh [22] and Deng et al. [21] have studied the performance of ELM with this enhancement under the Sigmoid additive type of SLFNs. This section extends such study to generalized SLFNs with a different type of hidden nodes (feature mappings) as well as kernels.

There is a gap between ELM and LS-SVM/PSVM, and it is not clear whether there is some relationship between ELM and LS-SVM/PSVM. This section aims to fill the gap and build the relationship between ELM and LS-SVM/PSVM.

A. Sufficient and Necessary Conditions for Universal Classifiers

1) Universal Approximation Capability: According to ELM learning theory, a widespread type of feature mappings \( h(x) \) can be used in ELM so that ELM can approximate any continuous target functions (refer to [14]–[16] for details). That is, given any target continuous function \( f(x) \), there exists a series of \( \beta_i \)’s such that

\[
\lim_{L \to +\infty} \|f_L(x) - f(x)\| = \lim_{L \to +\infty} \left\| \sum_{i=1}^{L} \beta_i h_i(x) - f(x) \right\| = 0.
\]

With this universal approximation capability, the bias \( b \) in the optimization constraints of SVM, LS-SVM, and PSVM can be removed, and the resultant learning algorithm has milder optimization constraints. Thus, better generalization performance and lower computational complexity can be obtained. In SVM, LS-SVM, and PSVM, as the feature mapping \( \phi(x) \) may be unknown, usually not every feature mapping to be used in SVM, LS-SVM, and PSVM satisfies the universal approximation condition. Obviously, a learning machine with a feature mapping which does not satisfy the universal approximation condition cannot approximate all target continuous functions. Thus, the universal approximation condition is not only a sufficient condition but also a necessary condition for a feature mapping to be widely used. This is also true to classification applications.

2) Classification Capability: Similar to the classification capability theorem of single-hidden-layer feedforward neural networks [38], we can prove the classification capability of the generalized SLFNs with the hidden-layer mapping \( h(x) \) satisfying the universal approximation condition (22).

Definition 3.1: A closed set is called a region regardless whether it is bounded or not.

Lemma 3.1 [38]: Given disjoint regions \( K_1, K_2, \ldots, K_m \) in \( \mathbb{R}^d \) and the corresponding \( m \) arbitrary real values \( c_1, c_2, \ldots, c_m \), and an arbitrary region \( X \) disjointed from any \( K_i \), there exists a continuous function \( f(x) \) such that \( f(x) = c_i \) if \( x \in K_i \) and \( f(x) = c_0 \) if \( x \in X \), where \( c_0 \) is an arbitrary real value different from \( c_1, c_2, \ldots, c_p \).

The classification capability theorem of Huang et al. [38] can be extended to generalized SLFNs which need not be neuron alike.

Theorem 3.1: Given a feature mapping \( h(x) \), if \( h(x)\beta \) is dense in \( C(\mathbb{R}^d) \) or in \( C(M) \), where \( M \) is a compact set of \( \mathbb{R}^d \), then a generalized SLFN with such a hidden-layer mapping \( h(x) \) can separate arbitrary disjoint regions of any shapes in \( \mathbb{R}^d \) or \( M \).

Proof: Given \( m \) disjoint regions \( K_1, K_2, \ldots, K_m \) in \( \mathbb{R}^d \) and their corresponding \( m \) labels \( c_1, c_2, \ldots, c_m \), according to Lemma 3.1, there exists a continuous function \( f(x) \) in \( C(\mathbb{R}^d) \) or in one compact set of \( \mathbb{R}^d \) such that \( f(x) = c_i \) if \( x \in K_i \), and \( f(x) \) is dense in \( C(\mathbb{R}^d) \) or on one compact set of \( \mathbb{R}^d \), then it can approximate the function \( f(x) \), and there exists a corresponding generalized SLFN to implement such a function \( f(x) \). Thus, such a generalized SLFN can separate these decision regions regardless of the shapes of these regions. □

Seen from Theorem 3.1, it is a necessary and sufficient condition that the feature mapping \( h(x) \) is chosen to make \( h(x)\beta \) have the capability of approximating any target continuous function. If \( h(x)\beta \) cannot approximate any target continuous functions, there may exist some shapes of regions which cannot be separated by a classifier with such feature mapping \( h(x) \). In other words, as long as the dimensionality of the feature mapping (number of hidden nodes \( L \) in a classifier) is large enough, the output of the classifier \( h(x)\beta \) is close to the class labels in the corresponding regions as possible.

In the binary classification case, ELM only uses a single-output node, and the class label closer to the output value of ELM is chosen as the predicted class label of the input data. There are two solutions for the multiclass classification case.

1) ELM only uses a single-output node, and among the multiclass labels, the class label closer to the output value of ELM is chosen as the predicted class label of the
input data. In this case, the ELM solution to the binary classification case becomes a specific case of multiclass solution.  
2) ELM uses multioutput nodes, and the index of the output node with the highest output value is considered as the label of the input data.  
For the sake of readability, these two solutions are analyzed separately. It can be found that, eventually, the same solution formula is obtained for both cases.

B. Simplified Constrained-Optimization Problems

1) Multiclass Classifier With Single Output: Since ELM can approximate any target continuous functions and the output of the ELM classifier \( h(x) / \beta \) can be as close to the class labels in the corresponding regions as possible, the classification problem for the proposed constrained-optimization-based ELM with a single-output node can be formulated as

\[
\text{Minimize : } L_{\text{ELM}} = \frac{1}{2} \| \beta \|^2 + C \sum_{i=1}^{N} \xi_i^2 \\
\text{Subject to : } h(x_i) / \beta = t_i - \xi_i, \quad i = 1, \ldots, N. \tag{23}
\]

Based on the KKT theorem, to train ELM is equivalent to solving the following dual optimization problem:

\[
L_{\text{ELM}} = \frac{1}{2} \| \beta \|^2 + C \sum_{i=1}^{N} \xi_i^2 \\
- \sum_{i=1}^{N} \alpha_i \left( h(x_i) / \beta - t_i + \xi_i \right) \tag{24}
\]

where each Lagrange multiplier \( \alpha_i \) corresponds to the \( i \)th training sample. We can have the KKT optimality conditions of (24) as follows:

\[
\frac{\partial L_{\text{ELM}}}{\partial \beta} = 0 \rightarrow \beta = \sum_{i=1}^{N} \alpha_i h(x_i)^T = H^T \alpha \tag{25a}
\]

\[
\frac{\partial L_{\text{ELM}}}{\partial \xi_i} = 0 \rightarrow \alpha_i = C \xi_i, \quad i = 1, \ldots, N \tag{25b}
\]

\[
\frac{\partial L_{\text{ELM}}}{\partial \alpha_i} = 0 \rightarrow h(x_i) / \beta - t_i + \xi_i = 0, \quad i = 1, \ldots, N \tag{25c}
\]

where \( \alpha = [\alpha_1, \ldots, \alpha_N]^T \).

2) Multiclass Classifier With Multioutputs: An alternative approach for multiclass applications is to let ELM have multioutput nodes instead of a single-output node. \( m \)-class of classifiers have \( m \) output nodes. If the original class label is \( p \), the expected output vector of the \( m \) output nodes is \( t = [0, \ldots, 0, 1, 0, \ldots, 0]^T \). In this case, only the \( p \)th element of \( t = [t_{1}, \ldots, t_{m}]^T \) is one, while the rest of the elements are set to zero. The classification problem for ELM with multioutput nodes can be formulated as

\[
\text{Minimize : } L_{\text{ELM}} = \frac{1}{2} \| \beta \|^2 + C \sum_{i=1}^{N} \| \xi_i \|^2 \\
\text{Subject to : } h(x_i) / \beta = t_i^T - \xi_i^T, \quad i = 1, \ldots, N \tag{26}
\]

where \( \xi_i = [\xi_{i,1}, \ldots, \xi_{i,m}]^T \) is the training error vector of the \( m \) output nodes with respect to the training sample \( x_i \). Based on the KKT theorem, to train ELM is equivalent to solving the following dual optimization problem:

\[
L_{D_{\text{ELM}}} = \frac{1}{2} \| \beta \|^2 + C \sum_{i=1}^{N} \| \xi_i \|^2 \\
- \sum_{i=1}^{N} \sum_{j=1}^{m} \alpha_{i,j} \left( h(x_i) / \beta - t_{i,j} + \xi_{i,j} \right) \tag{27}
\]

where \( \beta_i \) is the vector of the weights linking hidden layer to the \( j \)th output node and \( \beta = [ \beta_1, \ldots, \beta_m ] \). We can have the KKT corresponding optimality conditions as follows:

\[
\frac{\partial L_{D_{\text{ELM}}}}{\partial \beta_j} = 0 \rightarrow \beta_j = \sum_{i=1}^{N} \alpha_{i,j} h(x_i)^T \rightarrow \beta = H^T \alpha \tag{28a}
\]

\[
\frac{\partial L_{D_{\text{ELM}}}}{\partial \xi_i} = 0 \rightarrow \alpha_i = C \xi_i, \quad i = 1, \ldots, N \tag{28b}
\]

\[
\frac{\partial L_{D_{\text{ELM}}}}{\partial \alpha_i} = 0 \rightarrow h(x_i) / \beta - t_i^T + \xi_i^T = 0, \quad i = 1, \ldots, N \tag{28c}
\]

where \( \alpha_i = [\alpha_{i,1}, \ldots, \alpha_{i,m}]^T \) and \( \alpha = [\alpha_1, \ldots, \alpha_N]^T \).

It can be seen from (24), (25a)–(25c), (27), and (28a)–(28c) that the single-output node case considered a specific case of multioutput nodes when the number of output nodes is set to one: \( m = 1 \). Thus, we only need to consider the multiclass classifier with multioutput nodes. For both cases, the hidden-layer matrix \( H \) (20) remains the same, and the size of \( H \) is only decided by the number of training samples \( N \) and the number of hidden nodes \( L \), which is irrelevant to the number of output nodes (number of classes).

C. Equality Constrained-Optimization-Based ELM

Different solutions to the aforementioned KKT conditions can be obtained based on the concerns on the efficiency in different size of training data sets.

1) For the Case Where the Number of Training Samples is Not Huge: In this case, by substituting (28a) and (28b) into (28c), the aforementioned equations can be equivalently written as

\[
\left( \frac{I}{C} + HH^T \right) \alpha = T \tag{29}
\]

where

\[
T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix} = \begin{bmatrix} t_{11} & \cdots & t_{1m} \\ \vdots & \ddots & \vdots \\ t_{N1} & \cdots & t_{Nm} \end{bmatrix} \tag{30}
\]

From (28a) and (29), we have

\[
\beta = H^T \left( \frac{I}{C} + HH^T \right)^{-1} T \tag{31}
\]
The output function of ELM classifier is
\[ f(x) = h(x)\beta = h(x)H^T \left( \frac{1}{C} + HH^T \right)^{-1} \beta. \] (32)

1) Single-output node \((m = 1)\): For multiclass classifications, among all the multiclass labels, the predicted class label of a given testing sample is closest to the output of ELM classifier. For binary classification case, ELM needs only one output node \((m = 1)\), and the decision function of ELM classifier is
\[ f(x) = \text{arg} \max_i f_i(x). \] (34)

2) For the Case Where the Number of Training Samples is \(\gg L\): If the number of training data is very large, for example, we have \(N \gg L\). In this case, the output function of ELM classifier is
\[ f(x) = \text{sign} \left( h(x)H^T \left( \frac{1}{C} + HH^T \right)^{-1} T \right). \] (33)

2) Multioutput nodes \((m > 1)\): For multiclass cases, the predicted class label of a given testing sample is the index number of the output node which has the highest output value for the given testing sample. Let \(f_j(x)\) denote the output function of the \(j\)th output node, i.e., \(f(x) = [f_1(x), \ldots, f_m(x)]^T\); then, the predicted class label of sample \(x\) is
\[ \text{label}(x) = \text{arg} \max_i f_i(x). \] (35)

IV. DISCUSSIONS

A. Random Feature Mappings and Kernels

1) Random Feature Mappings: Different from SVM, LS-SVM, and PSVM, in ELM, a feature mapping (hidden-layer output vector) \(h(x) = [h_1(x), \ldots, h_L(x)]\) is usually known to users. According to [15] and [16], almost all nonlinear piecewise continuous functions can be used as the hidden-node output functions, and thus, the feature mappings used in ELM can be very diversified.

For example, as mentioned in [29], we can have
\[ h(x) = [G(a_1, b_1, x), \ldots, G(a_L, b_L, x)] \] (40)
where \(G(a, b, x)\) is a nonlinear piecewise continuous function satisfying ELM universal approximation capability theorems [14]–[16] and \((a_i, b_i)\) are randomly generated according to any continuous probability distribution. For example, such nonlinear piecewise continuous functions can be as follows.

1) Sigmoid function
\[ G(a, b, x) = \frac{1}{1 + \exp \left( - (a \cdot x + b) \right)}. \] (41)

2) Hard-limit function
\[ G(a, b, x) = \begin{cases} 1, & \text{if } a \cdot x - b \geq 0 \\ 0, & \text{otherwise}. \end{cases} \] (42)

3) Gaussian function
\[ G(a, b, x) = \exp \left( - b \| x - a \|^2 \right). \] (43)

4) Multiquadric function
\[ G(a, b, x) = \left( \| x - a \|^2 + b^2 \right)^{1/2}. \] (44)

Sigmoid and Gaussian functions are two of the major hidden-layer output functions used in the feedforward neural networks and RBF networks\(^2\), respectively. Interestingly, ELM with

\(^2\)Readers can refer to [39] for the difference between ELM and RBF networks.
the hidden layer is feature mapping of SVM. However, different from ELM where parametric feature mapping of their SVM implementation is neither.

In this specific case, similar to SVM, LS-SVM, and PSVM, the feature mapping \( h(x) \) need not be known to users; instead, its corresponding kernel \( K(u, v) \) (e.g., \( K(u, v) = \exp(-\gamma||u-v||^2) \)) is given to users. The dimensionality \( L \) of the feature space (number of hidden nodes) need not be given either.

3) Feature Mapping Matrix: In ELM, \( H = [h(x_1)^T, \ldots, h(x_N)^T]^T \) is called the hidden-layer output matrix (or called feature mapping matrix) due to the fact that it represents the corresponding hidden-layer outputs of the given \( N \) training samples. \( h(x_i) \) denotes the output of the hidden layer with regard to the input sample \( x_i \). Feature mapping \( h(x_i) \) maps the data \( x_i \) from the input space to the hidden-layer feature space, and the feature mapping matrix \( H \) is irrelevant to target \( t_i \). As observed from the essence of the feature mapping, it is reasonable to have the feature mapping matrix independent from the target values \( t_i \)'s. However, in both LS-SVM and PSVM, the feature mapping matrix \( Z = [t_1 \phi(x_1)^T, \ldots, t_N \phi(x_N)^T]^T \) (12) is designed to depend on the targets \( t_i \)'s of the training samples \( x_i \)’s.

B. ELM: Unified Learning Mode for Regression, Binary, and Multiclass Classification

As observed from (32) and (38), ELM has the unified solutions for regression, binary, and multiclass classification. The kernel matrix \( \Omega_{ELM} = HH^T \) is only related to the input data \( x_i \) and the number of training samples. The kernel matrix \( \Omega_{ELM} \) is neither relevant to the number of output nodes \( m \) nor to the training target values \( t_i \)'s. However, in multiclass LS-SVM, aside from the input data \( x_i \), the kernel matrix \( \Omega_M \) (52) also depends on the number of output nodes \( m \) and the training target values \( t_i \)'s.

For the multiclass case with \( m \) labels, LS-SVM uses \( m \) output nodes in order to encode multiclasses where \( t_{ij} \) denotes the output value of the \( j \)th output node for the training data \( x_i \) [10]. The \( m \) outputs can be used to encode up to \( 2^m \) different classes. For multiclass case, the primal optimization problem of LS-SVM can be given as [10]

Minimize: \( \ell^{(m)}_{\text{PSVM}} = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{N} \xi_{ij}^2 + C \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{m} (t_i \phi_j(x_i) + b_i) - \xi_{i} \)

Subject to:

\[
\begin{align*}
\ell_{i,1} \left(w_1 \cdot \phi_1(x_i) + b_1\right) & = 1 - \xi_{i,1} \\
\ell_{i,2} \left(w_2 \cdot \phi_2(x_i) + b_2\right) & = 1 - \xi_{i,2} \\
& \vdots \\
\ell_{i,m} \left(w_m \cdot \phi_m(x_i) + b_m\right) & = 1 - \xi_{i,m} \\
\xi_{i,j} & \geq 0, \quad i = 1, \ldots, N, \quad j = 1, \ldots, m. 
\end{align*}
\]

Similar to the LS-SVM solution (11) to the binary classification, with KKT conditions, the corresponding LS-SVM solution for multiclass cases can be obtained as follows:

\[
\begin{bmatrix}
0 \\
T
\end{bmatrix} \Omega_M^{-1} \begin{bmatrix}
\alpha_M^T \\
\Omega_M
\end{bmatrix} = \begin{bmatrix}
0 \\
I
\end{bmatrix}
\]

\( \Omega_M = \text{blockdiag} \left[ \Omega^{(1)} + \frac{I}{C}, \ldots, \Omega^{(m)} + \frac{I}{C} \right] \)

\( \Omega_{kl}^{(j)} = t_{k,j} t_{l,j} K^{(j)}(x_k, x_l) \)

\( b_M = [b_1, \ldots, b_m] \)

\( \alpha_M = [\alpha_1, \ldots, \alpha_N, \ldots, \alpha_1, \ldots, \alpha_{N,m}] \)

\( K^{(j)}(x_k, x_l) = \phi_j(x_k) \cdot \phi_j(x_l) \)

\[
= \exp \left( -\frac{||x_k - x_l||^2}{\sigma_j^2} \right), \quad j = 1, \ldots, N. 
\]
Seems from (52), the multiclass LS-SVM actually uses \( m \) binary-class LS-SVM concurrently for \( m \) class label classifications; each of the \( m \) binary-class LS-SVMs may have different kernel matrix \( \Omega^{(j)} \), \( j = 1, \ldots, m \). However, in any cases, ELM has one hidden layer linking to all the \( m \) output nodes. In multiclass LS-SVM, different kernels may be used in each individual binary LS-SVM, and the \( j \)th LS-SVM uses kernel \( K^{(j)}(u, v) \). Take Gaussian kernel as an example, \( K^{(j)}(u, v) = \exp(-||x_k - x_i||^2/\sigma_j^2) \); from practical point of view, it may be time consuming and tedious for users to choose different kernel parameters \( \sigma_j \); thus, one may set a common value \( \sigma_j = \sigma \) for all the kernels. In multiclass LS-SVM, the size of \( \Omega_M \) is \( N \times Nm \), which is related to the number of output nodes \( m \). However, in ELM, the size of kernel matrix \( \Omega_{ELM} = HH^T \) is \( N \times N \) which is fixed for all the regression, binary, and multiclass classification cases.

### C. Computational Complexity and Scalability

For LS-SVM and PSVM, the main computational cost comes from calculating the Lagrange multipliers \( \alpha \)'s based on (11) and (16). Obviously, ELM computes \( \alpha \) based on a simpler method (29). More importantly, in large-scale applications, instead of \( HH^T \) (size: \( N \times N \)), ELM can get a solution based on (37), where \( H^T \Theta \) (size: \( L \times L \)) is used. As in most applications, the number of hidden nodes \( L \) could be much smaller than the number of training samples: \( L \ll N \), the computational cost reduces dramatically. For the case \( L \ll N \), ELM can use \( H^T H \) (size: \( L \times L \)). Compared with LS-SVM and PSVM which use \( HH^T \) (size: \( N \times N \)), ELM has much better computational scalability with regard to the number of training samples \( N \). (cf. Fig. 1 for example.)

In order to reduce the computational cost of LS-SVM in large-scale problems, fixed-size LS-SVM has been proposed by Suykens et al. \[40\]-\[44\]. Fixed-size LS-SVM uses an \( M \)-sample subset of the original training data set (\( M \ll N \)) to compute a finite dimensional approximation \( \phi(x) \) to the feature map \( h(x) \). However, different from the fixed-size LS-SVM, if \( L \ll N \), \( L \times L \) solution of ELM still uses the entire \( N \) training samples. In any case, the feature map \( h(x) \) of ELM is not approximated. In fact, the feature map \( h(x) \) of ELM is randomly generated and independent of the training samples (if random hidden nodes are used). The kernel matrix of the fixed-size LS-SVM is built with the subset of size \( M \ll N \), while the kernel matrix of ELM is built with the entire data set of size \( N \) in all cases.

### D. Difference From Other Regularized ELMs

Toh \[22\] and Deng et al. \[21\] proposed two different types of weighted regularized ELMs. The total error rate (TER) ELM \[22\] uses \( m \) output nodes for \( m \) class label classification applications. In TER-ELM, the counting cost function adopts a quadratic approximation. The OAA method is used in the implementation of TER-ELM in multiclass classification applications. Essentially, TER-ELM consists of \( m \) binary TER-ELM, where \( j \)th TER-ELM is trained with all of the samples in the \( j \)th class with positive labels and all the other examples from the remaining \( m - 1 \) classes with negative labels. Suppose that there are \( m_j \) number of positive category patterns and \( m_i \) number of negative category patterns in the \( j \)th binary TER-ELM. We have a positive output \( y_j^+ = (\tau + \eta)1_j^+ \) for the \( j \)th class of samples and a negative class output \( y_j^- = (\tau - \eta)1_j^- \) for all the non-\( j \)th class of samples, where \( 1_j^+ = [1, \ldots, 1]^T \in \mathbb{R}^{m_j} \) and \( 1_j^- = [1, \ldots, 1]^T \in \mathbb{R}^{m_j} \). A common setting for threshold \( \tau \) and bias \( \eta \) will be set for all the \( m \) outputs. The output weight vector \( \beta_j \) in \( j \)th binary TER-ELM is calculated as

\[
\beta_j = \left( \frac{1}{m_j} H_j^{-T} H_j + \frac{1}{m_j} H_j^{+T} H_j + \right)^{-1} \cdot \left( \frac{1}{m_j} H_j^{-T} y_j^- + \frac{1}{m_j} H_j^{+T} y_j^+ \right) \tag{54}
\]

where \( H_j^+ \) and \( H_j^- \) denote the hidden-layer matrices of the \( j \)th binary TER-ELM corresponding to the positive and negative samples, respectively. By defining two class-specific diagonal weighting matrices \( W_j^+ = \text{diag}(0, \ldots, 0, 1/m_j^+, \ldots, 1/m_j^+) \) and \( W_j^- = \text{diag}(1/m_j^-, \ldots, 1/m_j^-, 0, \ldots, 0) \), the solution formula (54) of TER-ELM can be written as

\[
\beta_j = \left( \frac{1}{C} + H_j^T W_j H_j \right)^{-1} H_j^T W_j y_j \tag{55}
\]

where \( W_j = W_j^+ + W_j^- \) and the elements of \( H_j \) and \( y_j \) are ordered according to the positive and negative samples of the two classes (\( j \)th class samples and all the non-\( j \)th class samples). In order to improve the stability of the learning, \( I/C \) is introduced in the aforementioned formula. If the dimensionality of the hidden layer is much larger than the number of the training data (\( L \gg N \)), an alternative solution suggested in [22] is

\[
\beta_j = H_j^T \left( \frac{1}{C} + W_j H_j H_j^T \right)^{-1} W_j y_j \tag{56}
\]

Kernels and generalized feature mappings are not considered in TER-ELM.
Deng et al. [21] mainly focus on the case where \( L < N \), and (37) of ELM and the solution formula of Deng et al. [21] look similar to each other. However, different from the ELM solutions provided in this paper, Deng et al. [21] do not consider kernels and generalized feature mappings in their weighted regularized ELM. In the proposed solutions of ELM, \( L \) hidden nodes may have a different type of hidden-node output function \( h_i(x) : h(x) = [h_1(x), \ldots, h_L(x)] \), while in [21], all the hidden nodes use the Sigmoid type of activation functions. Deng et al. [21] do not handle the alternative solution (31).

Seen from (37), multivariate polynomial model [45] can be considered as a specific case of ELM.

The original solutions (21) of ELM [12], [13], [26], TER-ELM [22], and the weighted regularized ELM [21] are not able to apply kernels in their implementations. With the new suggested approach, kernels can be used in ELM [cf. (49)].

### E. Milder Optimization Constraints

In LS-SVM, as the feature mapping \( \phi(x) \) is usually unknown, it is reasonable to think that the separating hyperplane in LS-SVM may not necessarily pass through the origin in the LS-SVM feature space, and thus, a term bias \( b \) is required in their optimization constraints:

\[
t_i((w \cdot \phi(x_i)) + b) = 1 - \xi_i.
\]

The corresponding KKT condition (necessary condition) [cf. (10b)] for the conventional LS-SVM is

\[
\sum_{i=1}^{N} \alpha_i t_i = 0.
\]

Poggio et al. [46] prove in theory that the term bias \( b \) is not required in positive definite kernel and that it is not incorrect to have the term bias \( b \) in the SVM model. Different from the analysis of Poggio et al. [46], Huang et al. [29] show that, from the practical and universal approximation point of view, the term bias \( b \) should not be given in the ELM learning.

According to ELM theories [12]–[16], almost all nonlinear piecewise continuous functions as feature mappings can make ELM satisfy universal approximation capability, and the separating hyperplane of ELM basically tends to pass through the origin in the ELM feature space. There is no term bias \( b \) in the optimization constraint of ELM, \( h(x_i)\beta = t_i - \xi_i \), and thus, different from LS-SVM, ELM does not need to satisfy the condition

\[
\sum_{i=1}^{N} \alpha_i t_i = 0.
\]

Although LS-SVM and ELM have the same primal optimization formula, ELM has milder optimization constraints than LS-SVM, and thus, compared to ELM, LS-SVM obtains a suboptimal optimization.

The differences and relationships among ELM, LS-SVM/PSVM, and SVM can be summarized in Table I.

### V. PERFORMANCE VERIFICATION

This section compares the performance of different algorithms (SVM, LS-SVM, and ELM) in real-world benchmark regression, binary, and multiclass classification data sets. In order to test the performance of the proposed ELM with various feature mappings in supersmall data sets, we have also tested ELM on the XOR problem.

#### A. Benchmark Data Sets

In order to extensively verify the performance of different algorithms, wide types of data sets have been tested in our simulations, which are of small sizes, low dimensions, large sizes, and/or high dimensions. These data sets include 12 binary classification cases, 12 multiclassification cases, and 12 regression cases. Most of the data sets are taken from UCI Machine Learning Repository [47] and Statlib [48].

1) **Binary Class Data Sets**: The 12 binary class data sets (cf. Table II) can be classified into four groups of data:

- data sets with relatively small size and low dimensions, e.g., Pima Indians diabetes, Bupa Liver disorders [47], and Banana [49];
- data sets with relatively small size and high dimensions, e.g., leukemia data set [50] and colon microarray data set [51];
- data sets with relatively large size and low dimensions, e.g., Stat/Galaxy-Bright data set [52], Galaxy Dim data set [52], and mushroom data set [47];
- data sets with large size and high dimensions, e.g., adult data set [47].

The leukemia data set was originally taken from a collection of leukemia patient samples [53]. The data set consists of 72 samples: 25 samples of AML and 47 samples of ALL. Each sample of leukemia data set is measured over 7129 genes (cf. Leukemia in Table II). The colon microarray data set consists of 22 normal and 40 tumor tissue samples. In this data set, each sample of colon microarray data set contains 2000 genes (cf. Colon in Table II).

Performances of the different algorithms have also been tested on both leukemia data set and colon microarray data set after the minimum-redundancy–maximum-relevance feature selection method [54] being taken (cf. Leukemia (Gene Sel) and Colon (Gene Sel) in Table II).

2) **Multiclass Data Sets**: The 12 multiclass data sets (cf. Table III) can be classified into four groups of data as well:

- data sets with relatively small size and low dimensions, e.g., Iris, Glass Identification, and Wine [47];
- data sets with relatively medium size and medium dimensions, e.g., Vowel Recognition, Statlog Vehicle Silhouettes, and Statlog Image Segmentation [47];
- data sets with relatively large size and medium dimensions, e.g., letter and shuttle [47];
- data sets with large size and/or large dimensions, e.g., DNA, Satimage [47], and USPS [50].

3) **Regression Data Sets**: The 12 regression data sets (cf. Table IV) can be classified into three groups of data:

- data sets with relatively small size and low dimensions, e.g., Basketball, Strike [48], Cloud, and Autoprice [47];
TABLE I  
FEATURE COMPARISONS AMONG ELM, LS-SVM, AND SVM

<table>
<thead>
<tr>
<th>ELM</th>
<th>LS-SVM and PSVM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature mapping</td>
<td>i) $h(x)$ is usually known to users (Wide type of nonlinear piecewise continuous functions (with randomly generated parameters) can be used [14]-[16]; e.g., additive, RBF, trigonometric, threshold, fully complex, high-order, ridge polynomial, etc) ii) If $h(x)$ is unknown, kernels can be used in ELM then</td>
<td>$\phi(x)$ is usually known to users and kernels are often used.</td>
</tr>
<tr>
<td>Kernel matrix</td>
<td>$\Omega_{ELM} = [\Omega_{ij}]$ where $\Omega_{ij} = K(x_i, x_j)$</td>
<td>$\Omega_{LS-SVM} = \Omega_{PSVM} = [\Omega_{ij}]$ where $\Omega_{ij} = t_i x_j K(x_i, x_j)$</td>
</tr>
<tr>
<td>Universal approximation capability</td>
<td>Ensured for almost all type of nonlinear piecewise random hidden nodes and often used kernels</td>
<td>Not ensured. It depends on the kernel to be used</td>
</tr>
<tr>
<td>Support vectors based</td>
<td>No ($\alpha_i$ are proportional to the training errors)</td>
<td>No ($\alpha_i$ are proportional to the training errors)</td>
</tr>
<tr>
<td>Conditions on $\alpha_i$</td>
<td>No conditions on $\alpha_i$</td>
<td>For LS-SVM: $\sum_{i=1}^{N} t_i \alpha_i = 0$; For PSVM: $\sum_{i=1}^{N} t_i \alpha_i = b$</td>
</tr>
<tr>
<td>Multi classes</td>
<td>Single ELM for $m$ classes</td>
<td>$m$ binary LS-SVM for $m$ classes, depending on coding/decoding scheme</td>
</tr>
<tr>
<td>Regression</td>
<td>ELM unified for both regression and binary/multi-classification problems</td>
<td>Different variants required for regression and classification</td>
</tr>
</tbody>
</table>

TABLE III  
SPECIFICATION OF MULTICLASS CLASSIFICATION PROBLEMS

<table>
<thead>
<tr>
<th>Datasets</th>
<th># train</th>
<th># test</th>
<th># features</th>
<th># classes</th>
<th>Random Perm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>100</td>
<td>50</td>
<td>4</td>
<td>3</td>
<td>Yes</td>
</tr>
<tr>
<td>Glass</td>
<td>142</td>
<td>72</td>
<td>9</td>
<td>6</td>
<td>Yes</td>
</tr>
<tr>
<td>Wine</td>
<td>118</td>
<td>60</td>
<td>13</td>
<td>3</td>
<td>Yes</td>
</tr>
<tr>
<td>Ecoli</td>
<td>224</td>
<td>112</td>
<td>7</td>
<td>8</td>
<td>Yes</td>
</tr>
<tr>
<td>Vowel</td>
<td>528</td>
<td>462</td>
<td>10</td>
<td>11</td>
<td>No</td>
</tr>
<tr>
<td>Vehicle</td>
<td>564</td>
<td>282</td>
<td>18</td>
<td>4</td>
<td>Yes</td>
</tr>
<tr>
<td>Segment</td>
<td>1540</td>
<td>770</td>
<td>19</td>
<td>7</td>
<td>Yes</td>
</tr>
<tr>
<td>Sattimage</td>
<td>2000</td>
<td>36</td>
<td>6</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>DNA</td>
<td>2000</td>
<td>1186</td>
<td>180</td>
<td>33</td>
<td>No</td>
</tr>
<tr>
<td>Letter</td>
<td>13333</td>
<td>6667</td>
<td>16</td>
<td>26</td>
<td>Yes</td>
</tr>
<tr>
<td>Shuttle</td>
<td>43500</td>
<td>14500</td>
<td>9</td>
<td>7</td>
<td>No</td>
</tr>
<tr>
<td>USPS</td>
<td>7291</td>
<td>2007</td>
<td>256</td>
<td>10</td>
<td>No</td>
</tr>
</tbody>
</table>

2) data sets with relatively small size and medium dimensions, e.g., Pyrim, Housing [47], Bodyfat, and Cleveland [48];
3) data sets with relatively large size and low dimensions, e.g., Balloon, Quake, Space-ga [48], and Abalone [47].

Column “random perm” in Tables II–IV shows whether the training and testing data of the corresponding data sets are reshuffled at each trial of simulation. If the training and testing data of the data sets remain fixed for all trials of simulations, it is marked “No.” Otherwise, it is marked “Yes.”

B. Simulation Environment Settings

The simulations of different algorithms on all the data sets except for Adult, Letter, Shuttle, and USPS data sets are carried out in MATLAB 7.0.1 environment running in Core 2 Quad, 2.66-GHZ CPU with 2-GB RAM. The codes used for SVM and LS-SVM are downloaded from [55] and [56], respectively. Simulations on large data sets (e.g., Adult, Letter, Shuttle, and USPS data sets) are carried out in a high-performance computer with 2.52-GHZ CPU and 48-GB RAM. The symbol “∗” marked in Tables VI and VII indicates that the corresponding data sets are tested in such a high-performance computer.
C. User-Specified Parameters

The popular Gaussian kernel function \( K(u, v) = \exp(-\gamma \|u - v\|^2) \) is used in SVM, LS-SVM, and ELM. ELM performance is also tested in the cases of Sigmoid type of additive hidden node and multiquadric RBF hidden node.

In order to achieve good generalization performance, the cost parameter \( C \) and kernel parameter \( \gamma \) of SVM, LS-SVM, and ELM need to be chosen appropriately. We have tried a wide range of \( C \) and \( \gamma \). For each data set, we have used 50 different values of \( C \) and 50 different values of \( \gamma \), resulting in a total of 2500 pairs of \((C, \gamma)\). The 50 different values of \( C \) and \( \gamma \) are \( \{2^{-24}, 2^{-23}, \ldots, 2^{24}, 2^{25}\} \).

It is known that the performance of SVM is sensitive to the combination of \((C, \gamma)\). Similar to SVM, the generalization performance of LS-SVM and ELM with Gaussian kernel depends closely on the combination of \((C, \gamma)\) as well (see Fig. 2 for the performance sensitivity of LS-SVM and ELM with Gaussian kernel on the user-specified parameters \((C, \gamma)\)). The best generalization performance of SVM, LS-SVM, and ELM with Gaussian kernel is usually achieved in a very narrow range of such combinations. Thus, the best combination of \((C, \gamma)\) of SVM, LS-SVM, and ELM with Gaussian kernel needs to be chosen for each data set.

For ELM with Sigmoid additive hidden node and multiquadric RBF hidden node, \( h(x) = \left[ G(a_1, b_1, x), \ldots, G(a_L, b_L, x) \right] \), where \( G(a, b, x) = 1/(1 + \exp(-(a \cdot x + b))) \) for Sigmoid additive hidden node or \( G(a, b, x) = (\|x - a\|^2 + b^2)^{1/2} \) for multiquadric RBF hidden node. All the hidden-node parameters \((a_i, b_i)\) are randomly generated based on uniform distribution. The user-specified parameters are \((C, L)\), where \( C \) is chosen from the range \( \{2^{-24}, 2^{-23}, \ldots, 2^{24}, 2^{25}\} \). Seen from Fig. 3, ELM can achieve good generalization performance as long as the number of hidden nodes \( L \) is large enough. In all our simulations on ELM with Sigmoid additive hidden node and multiquadric RBF hidden node, \( L = 1000 \). In other words, the performance of ELM with Sigmoid additive hidden node and multiquadric RBF hidden node is not sensitive to the number of hidden nodes \( L \). Moreover, \( L \) need not be specified by users; instead, users only need to specify one parameter: \( C \).

Fifty trials have been conducted for each problem. Simulation results, including the average testing accuracy, the corresponding standard deviation (Dev), and the training times, are given in this section.
The user-specified parameters chosen in our simulations are given in Table V.

D. Performance Comparison on XOR Problem

The performance of SVM, LS-SVM, and ELM has been tested in the XOR problem which has two training samples in each class. The aim of this simulation is to verify whether ELM can handle some rare cases such as the cases with extremely few training data sets. Fig. 4 shows the boundaries of different classifiers in XOR problem. It can be seen that, similar to SVM and LS-SVM, ELM is able to solve the XOR problem well.

E. Performance Comparison on Real-World Benchmark Data sets

Tables VI–VIII show the performance comparison of SVM, LS-SVM, and ELM with Gaussian kernel, random Sigmoid hidden nodes, and multiquadric RBF nodes. It can be seen that ELM can always achieve comparable performance as SVM and LS-SVM with much faster learning speed. Seen from Tables VI–VIII, different output functions of ELM can be used in different data sets in order to have efficient implementation.
in different size of data sets, although any output function can be used in all types of data sets.

Take Shuttle (large number of training samples) and USPS (medium size of data set with high input dimensions) data sets in Table VII as examples.

1) For Shuttle data sets, ELM with Gaussian kernel and random multiquadric RBF nodes runs 6 and 4466 times faster than LS-SVM, respectively.

2) For USPS data sets, ELM with Gaussian kernel and random multiquadric RBF nodes runs 6 and 65 times faster than LS-SVM, respectively, and runs 1342 and 13 832 times faster than SVM, respectively.

On the other hand, different from LS-SVM which is sensitive to the combinations of parameters \((C, \gamma)\), ELM with random multiquadric RBF nodes is not sensitive to the unique user-specified parameter \(C\) [cf. Fig. 3(b)] and is ease of use in the respective implementations.

Tables VI–VIII particularly highlight the performance comparison between LS-SVM and ELM with Gaussian kernel, and among the comparisons of these two algorithms, apparently, better test results are given in boldface. It can be seen that ELM with Gaussian kernel achieves the same generalization performance in almost all the binary classification and regression cases as LS-SVM at much faster learning speeds; however, ELM usually achieves much better generalization performance in multiclass classification cases (cf. Table VII) than LS-SVM.

Fig. 5 shows the boundaries of different classifiers in Banana case. It can be seen that ELM can classify different classes well.

VI. CONCLUSION

ELM is a learning mechanism for the generalized SLFNs, where learning is made without iterative tuning. The essence of ELM is that the hidden layer of the generalized SLFNs should not be tuned. Different from traditional learning theories on learning, ELM learning theory [14]–[16] shows that if SLFNs \(f(x) = \mathbf{h}(x)^T \mathbf{H}^T \left( \frac{1}{C} + \mathbf{H}^T \mathbf{H} \right)^{-1} \mathbf{T} \) can approximate any target continuous

<table>
<thead>
<tr>
<th>Datasets</th>
<th>SVM</th>
<th>LSSVM</th>
<th>Extreme Learning Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gaussian Kernel</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rate (%)</td>
</tr>
</tbody>
</table>

TABLE VI

Performance Comparison of SVM, LS-SVM, and ELM: Binary Class Data Sets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Testing</th>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rate (%)</td>
<td>Dev (%)</td>
</tr>
</tbody>
</table>

TABLE VII

Performance Comparison of SVM, LS-SVM, and ELM: Multiclass Data Sets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Testing</th>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rate (%)</td>
<td>Dev (%)</td>
</tr>
</tbody>
</table>
functions, tuning is not required in the hidden layer then. All the hidden-node parameters which are supposed to be tuned by conventional learning algorithms can be randomly generated according to any continuous sampling distribution [14]–[16]. This paper has shown that both LS-SVM and PSVM can be simplified by removing the term bias $b$ and the resultant learning algorithms are unified with ELM. Instead of different variants requested for different types of applications, ELM can be applied in regression and multiclass classification applications directly. More importantly, according to ELM theory [14]–[16], ELM can work with a widespread type of feature mappings (including Sigmoid networks, RBF networks, trigonometric networks, threshold networks, fuzzy inference systems, fully complex neural networks, high-order networks, ridge polynomial networks, etc).

ELM requires less human intervention than SVM and LS-SVM/PSVM. If the feature mappings $h(x)$ are known to users, in ELM, only one parameter $C$ needs to be specified by users. The generalization performance of ELM is not sensitive to the dimensionality $L$ of the feature space (the number of hidden nodes) as long as $L$ is set large enough (e.g., $L \geq 1000$ for all the real-world cases tested in our simulations). Different from SVM, LS-SVM, and PSVM which usually request two parameters ($C, \gamma$) to be specified by users, single-parameter setting makes ELM be used easily and efficiently.

If feature mappings are unknown to users, similar to SVM, LS-SVM, and PSVM, kernels can be applied in ELM as well. Different from LS-SVM and PSVM, ELM does not have constraints on the Lagrange multipliers $\alpha_i$’s. Since LS-SVM and ELM have the same optimization objective functions and LS-SVM has some optimization constraints on Lagrange multipliers $\alpha_i$’s, in this sense, LS-SVM tends to obtain a solution which is suboptimal to ELM.

As verified by the simulation results, compared to SVM and LS-SVM ELM achieves similar or better generalization performance for regression and binary class classification cases, and much better generalization performance for multiclass classification cases. ELM has better scalability and runs at much faster learning speed (up to thousands of times) than traditional SVM and LS-SVM.

This paper has also shown that, in theory, ELM can approximate any target continuous function and classify any disjoint regions.

**ACKNOWLEDGMENT**

The authors would like to thank L. Ljung from Linköpings Universitet, Sweden, for reminding us of possible relationships between ELM and LS-SVM. The authors would also like to thank H. White from the University of California, San Diego, for his constructive and inspiring comments and suggestions on our research on ELM.

**REFERENCES**

Guang-Bin Huang (M’98–SM’04) received the B.Sc degree in applied mathematics and M.Eng degree in computer engineering from Northeastern University, Shenyang, China, in 1991 and 1994, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 1999.

During undergraduate period, he also concurrently studied in the Applied Mathematics Department and Wireless Communication Department, Northeastern University, China. From June 1998 to May 2001, he was a Research Fellow with Singapore Institute of Manufacturing Technology (formerly known as Gintic Institute of Manufacturing Technology), where he has led/implemented several key industrial projects (e.g., Chief Designer and Technical Leader of Singapore Changi Airport Cargo Terminal Upgrading Project, etc). Since May 2001, he has been an Assistant Professor and Associate Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His current research interests include machine learning, computational intelligence, and extreme learning machines.

Dr. Huang serves as an Associate Editor of Neurocomputing and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B.

Hongming Zhou received the B.Eng. degree from Nanyang Technological University, Singapore, in 2009, where he is currently working toward the Ph.D. degree in the School of Electrical and Electronic Engineering.

His research interests include extreme learning machines, neural networks, and support vector machines.

Xiaojian Ding received the B.Sc. degree in applied mathematics and the M.Sc. degree in computer engineering from Xi’an University of Technology, Xi’an, China, in 2003 and 2006, respectively, and the Ph.D. degree from the School of Electronic and Information Engineering, Xi’an Jiaotong University, Xi’an, in 2010. He studied in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, under the award from the Chinese Scholarship Council of China from August 2009 to August 2010.

His research interests include extreme learning machines, neural networks, pattern recognition, and machine learning.

Rui Zhang received the B.Sc. and M.Sc. degrees in mathematics from Northwest University, Xi’an, China, in 1994 and 1997, respectively. She is currently working toward the Ph.D. degree in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

From August 2004 to January 2005, she was a Visiting Scholar with the Department of Mathematics, University of Illinois at Champaign–Urbana, Urbana. Since 1997, she has been with the Department of Mathematics, Northwest University, where she is currently an Associate Professor. Her current research interests include extreme learning machines, machine learning, and neural networks.