Generalized Variant Support Vector Machine

Mohammadi, Majid; Mousavi, S. Hamid; Effati, Sohrab

DOI
10.1109/TSMC.2019.2917019

Publication date
2021

Document Version
Accepted author manuscript

Published in
IEEE Transactions on Systems, Man, and Cybernetics: Systems

Citation (APA)
Mohammadi, M., Mousavi, S. H., & Effati, S. (2021). Generalized Variant Support Vector Machine. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*(5), 2798-2809. Article 8730505. https://doi.org/10.1109/TSMC.2019.2917019

Generalized Variant Support Vector Machine

Majid Mohammadi, S. Hamid Mousavi, and Sohrab Effati

***Abstract*—With the advancement in information technology, datasets with an enormous amount of data are available. The classification task on these datasets is more time- and memory-consuming as the number of data increases. The support vector machine (SVM), which is arguably the most popular classification technique, has disappointing performance in dealing with large datasets due to its constrained optimization problem. To deal with this challenge, the variant SVM (VSVM) has been utilized which has the fraction $(1/2)b^2$ in its primal objective function, where b is the bias of the desired hyperplane. The VSVM has been solved with different optimization techniques in more time- and memory-efficient fashion. However, there is no guarantee that its optimal solution is the same as the standard SVM. In this paper, we introduce the generalized VSVM (GVSVM) which has the fraction $(1/2t)b^2$ in its primal objective function, for a fixed positive scalar t. Further, we present the thorough theoretical insights that indicate the optimal solution of the GVSVM tends to the optimal solution of the standard SVM as $t \to \infty$. One vital corollary is to derive a closed-form formula to obtain the bias term in the standard SVM. Such a formula obviates the need of approximating it, which is the modus operandi to date. An efficient neural network is then proposed to solve the GVSVM dual problem, which is asymptotically stable in the sense of Lyapunov and converges globally exponentially to the exact solution of the GVSVM. The proposed neural network has less complexity in architecture and needs fewer computations in each iteration in comparison to the existing neural solutions. Experiments confirm the efficacy of the proposed recurrent neural network and the proximity of the GVSVM and the standard SVM solutions with more significant values of t.**

***Index Terms*—Convex programming, exponential convergence, generalized VSVM (GVSVM), recurrent neural network (RNN), support vector machine (SVM).**

I. Introduction

THE SUPPORT vector machine (SVM) is arguably the most popular classification approach in the realm of pattern recognition and machine learning. It has proved promising performance in various fields, including but not limited to image processing [1]–[3], geoscience [4], [5], bioinformatics [6]–[8], and biomedical [9]–[11].

Let $\{x_i \in \mathbb{R}^n\}_{i=1}^l$ be a set of data points and $y_i \in \{-1, 1\}$ be the corresponding label for x_i. The SVM goal is to divide these data points into two disjoint groups by a hyperplane such that it has the maximum margin of both classes. In addition, this hyperplane must separate data of a similar class in the same group. When data are linearly separable, the desired hyperplane $w^T x + b = 0$ can be obtained by solving the following convex optimization problem:

$$\begin{aligned} \min_{w,b} \ & \frac{1}{2} w^T w + c \sum_{i=1}^{l} \xi_i \\ \text{s.t.} \ & y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l \\ & \xi_i \geq 0, \quad i = 1, \ldots, l \end{aligned} \tag{1}$$

where w is an $n \times 1$ vector, $b \in \mathbb{R}$ is the bias term, $c \geq 0$ is a regularization parameter for the tradeoff between model complexity and training error, and ξ_i measures the difference between $w^T x_i + b$ and y_i.

When data are not linearly separable, they are transformed to another high-dimensional space $\phi(.)$, where they can be linearly separable, and then the constraints of the minimization (1) can be rewritten as

$$\begin{aligned} & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l \\ & \xi_i \geq 0, \quad i = 1, \ldots, l. \end{aligned} \tag{2}$$

Since the desired space $\phi(.)$ is unknown, solving the problem (1) subject to the constraints (2) becomes more complicated. To address this problem, the dual of the SVM is presented as follows:

$$\begin{aligned} \max_{u} \ & \sum_{i=1}^{l} u_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} u_i u_j y_i y_j K(x_i, x_j) \\ \text{s.t.} \ & \sum_{i=1}^{l} u_i y_i = 0 \\ & 0 \leq u_i \leq c, \quad i = 1, \ldots, l \end{aligned} \tag{3}$$

where $u = (u_1, \ldots, u_l)$ is the Lagrangian multiplier and K is a kernel function satisfying $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$.

Salient features of the standard SVM, including the convexity and the robustness against noises, have caused it to be more popular among other classifiers. However, finding its either primal or dual solution becomes challenging when large datasets are available. Hence, a high number of methods have been proposed to find the SVM solution more efficiently [12]–[16].

Manuscript received July 21, 2018; revised January 31, 2019; accepted May 4, 2019. Date of publication June 4, 2019; date of current version April 15, 2021. This paper was recommended by Associate Editor L. Wang. *(Corresponding author: Majid Mohammadi.)*

M. Mohammadi is with the Faculty of Technology, Policy and Management, Delft University of Technology, 2628BX Delft, The Netherlands (e-mail: m.mohammadi@tudelft.nl).

S. H. Mousavi is with the Department of Medical Physics and Acoustics, Carl von Ossietzky University of Oldenburg, D-26129 Oldenburg, Germany, and also with the Cluster of Excellence Hearing4all, Carl von Ossietzky University of Oldenburg, D-26129 Oldenburg, Germany (e-mail: hamid.mousavi@uni-oldenburg.de).

S. Effati is with the Department of Applied Mathematics, Ferdowsi University of Mashhad, Mashhad 9177948974, Iran (e-mail: s-effati@um.ac.ir).

Chapelle [12] considered the primal problem and proposed a method for the linear and nonlinear SVM. *LibLinear* is also the library for the linear classifiers which contain a fast solver for the linear classification using SVM [15]. Similarly, a Laplacian-based method was developed for the primal SVM which also suits in the semisupervised case [13]. A fur-ther semisupervised version of the SVM has been developed recently [17], [18].

Along with these algorithms, which solve the minimization (1) or (3), several methods find the desired hyperplane by utilizing the variant SVM (VSVM), which has the square of the bias term in its objective function, i.e.,

$$\begin{aligned} \min_{w,b} \ & \frac{1}{2}\left(w^T w + b^2\right) + c\sum_{i=1}^{l}\xi_i \\ \text{s.t.} \ & y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l \\ & \xi_i \geq 0, \qquad\qquad\qquad\quad i = 1, \ldots, l \end{aligned} \tag{4}$$

and its dual form is

$$\begin{aligned} \max_{u} \ & \sum_{i=1}^{l} u_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} u_i u_j y_i y_j K(x_i, x_j) \\ & - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} u_i u_j y_i y_j \\ \text{s.t.} \ & 0 \leq u_i \leq c, \quad i = 1, \ldots, l. \end{aligned} \tag{5}$$

In contrast to the standard SVM in which the bias term is not acquirable directly from solving the dual problem, this term can be obtained directly after solving the dual problem of the VSVM.

Another advantage of the VSVM is that its dual problem (5) does not have the equality constraint. This property enables us to apply effective matrix splitting methods in a straightforward manner. Mangasaria and Musicant [19] proposed a method based on the successive over-relaxation (SOR) to find the solution of the VSVM. As the SOR handles one point at a time, it can be applied to large datasets because only one sample resides in memory at a time. The Lagrangian SVM (LSVM) is yet another effort to accelerate the speed of convergence [20]. In this algorithm, the solution of the VSVM is obtained via applying a simple iterative method on the fixed point problem derived from the VSVM. More methods utilizing the VSVM can be found in [21]–[26].

Recently the same approach as the VSVM has applied to the Laplacian SVM [27]. The squared bias term is added to the objective function of the primal Laplacian SVM, and its solution is obtained using SOR.

However, the solution of the VSVM is not necessarily equivalent to the solution of the standard SVM since their objective functions are distinct. Hence, there is no guarantee for utilizing the VSVM to obtain the maximum-margin separating hyperplane and to achieve the desired bias term in the standard SVM (the same statement is correct for the Laplacian SVM). In this paper, the generalized VSVM (GVSVM) is introduced, in which the fraction $(1/2t)$ of the squared bias is added to the objective function of the primal SVM in lieu of the square bias itself, where t is a fixed positive scalar. The primary motivation of this paper is to prove that the solution of the GVSVM converges to the solution of the standard SVM as $t \to \infty$. Further, the bias term, which is obtained directly by the GVSVM, is assured to tend to the bias term of the standard SVM. By using the GVSVM, as a result, the distinguishing feature of the VSVM is inherited while the identity of its solution with the standard SVM is guaranteed. The experiments confirm assigning the biggest possible value for t to acquire the exactness of the solutions of the standard SVM and GVSVM.

One avenue to solve the optimization problems is to use the recurrent neural network (RNN) [28]–[30]. The use of neural networks for solving optimization problems has several salient advantages. First, the structure of the neural network can be implemented using VLSI technology so that they can be used in real-time data processing. Second, the differential equation of the continuous neural network can be efficiently solved using numerical methods on digital computers. In this regard, we show that the GVSVM can be solved in a more efficient way than the standard SVM by first developing a neural solution and then juxtaposing it with the existing neural networks for the standard SVM. The proposed neural network is proved to be asymptotically stable in the sense of Lyapunov and is globally exponentially convergent to the solution of GVSVM. In contrast to the existing neural solution whose convergence is reliant on the given dataset, the proposed neural network is convergent to the optimal solution of the GVSVM regardless of the given dataset. Finally, the proposed neural network theoretically and empirically compared with the existing neural networks for the standard SVM in terms of their complexity and convergence rate, during which its superiority is demonstrated.

In a nutshell, the contributions of this paper can be summarized as follows.

1) The GVSVM is introduced and it is demonstrated that its solution is identical to the standard SVM under certain circumstances.
2) We derive a formula to directly compute the bias term of the desired hyperplane and will demonstrate that it is identical to that in minimization (1).
3) An efficient RNN is proposed to solve the GVSVM dual problem. The proposed neural network is assured to converge globally exponentially to its equilibrium.
4) This neural network is shown to have less complexity in architecture and computations in each iteration and also converges exponentially to the solution of the GVSVM.

This paper is structured as follows. In Section II, the GVSVM is introduced, and its optimal solution is demonstrated to converge to the solution of the standard SVM for the significant values of t in Section III. An efficient neural network is proposed in Section IV and its convergence is meticulously analyzed. The experimental results are presented in Section V, and the main points and conclusion are presented in Sections VI and VII, respectively.

II. Generalized Variant SVM

In this section, the GVSVM problem is introduced, and its dual form is obtained accordingly. Further, it is proved that the solution of GVSVM tends to the solution of the standard SVM under certain circumstances.

A. Formulation

As mentioned above, the term $(1/2t)b^2$ is added to the objective function of the standard SVM to derive the GVSVM problem. Thus, let $z = (w^T, b)^T$, the matrix form of the GVSVM primal is presented as

$$\begin{aligned} P_t: \quad \min \ & P_t(z,\xi) = \frac{1}{2} z^T Q_t z + C^T \xi \\ \text{s.t.} \ & \mathbf{1}_{l\times 1} - Az - \xi \le 0 \\ & \xi \ge 0 \end{aligned} \tag{6}$$

where C is a vector with elements c, $\mathbf{1}_{l\times 1}$ denotes an $l \times 1$ vector with elements 1, and Q_t and A are the matrices with definitions

$$Q_t = \begin{bmatrix} I_{n\times n} & 0 \\ 0 & \frac{1}{t} \end{bmatrix}, \ A = \begin{bmatrix} y_1 & 0 & \dots & 0 \\ 0 & y_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & y_l \end{bmatrix} \times \begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_l^T & 1 \end{bmatrix}$$

and $x_i = (x_{i1}, x_{i2}, \dots, x_{in})^T$.

The GVSVM (6) is equivalent to the standard SVM and VSVM for $(1/t)$ being 0 and 1, respectively. In addition, the objective function of the GVSVM (and also VSVM) is strictly convex since the matrix Q_t is positive definite, while the objective function of the standard SVM is just convex because of Q_t being positive semidefinite.

In addition, note that the GVSVM with $\bar{b} = (b/\sqrt{t})$ is identical to the VSVM problem and hence there is a one-to-one relationship between these problems. However, the GVSVM with bigger values for t will be proved to have more proximity to the standard SVM rather than the VSVM, in which $t = 1$. In further sections, the complete theoretical insights for this claim are discussed.

B. GVSVM Dual Problem

Considering the non-negative Lagrangian multipliers $u = (u_1, \dots, u_l)$ and $v = (v_1, \dots, v_l)$, the augmented objective function for the GVSVM can be written as

$$\begin{aligned} L(w,b,\xi,u,v) &= \frac{1}{2} z^T Q_t z + C^T \xi + u^T(\mathbf{1}_{l\times 1} - Az - \xi) - v^T \xi \\ &= \frac{1}{2} w^T w + \frac{1}{2t} b^2 + c \sum_{i=1}^{l} \xi_i \\ &\quad - \sum_{i=1}^{l} u_i \big(y_i \big(w^T x_i + b\big) - 1 + \xi_i\big) - \sum_{i=1}^{l} v_i \xi_i. \end{aligned} \tag{7}$$

According to the necessary and sufficient Karush–Kuhn–Tucker (KKT) optimality conditions [31], $(w^*, b^*, \xi^*, u^*, v^*)$ are optimal for primal and dual GVSVM problems if and only if

$$\begin{cases} w^* = \sum_{i=1}^{l} u_i^* y_i x_i \\ b^* = t \sum_{i=1}^{l} u_i^* y_i \\ c - u_i^* - v_i^* = 0, \quad i = 1, \dots, l \\ u_i^* \big(1 - y_i \big(w^{*T} x_i + b^*\big) - \xi_i^*\big) = 0, \quad i = 1, \dots, l \\ v_i^* \xi_i^* = 0, \quad i = 1, \dots, l. \end{cases}$$

Substituting above equalities in (7), the dual to GVSVM problem is presented as

$$\begin{aligned} D_t: \quad \max_{u} \ & \sum_{i=1}^{l} u_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} u_i u_j y_i y_j x_i x_j \\ & - \frac{t}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} u_i u_j y_i y_j \\ \text{s.t.} \ & 0 \le u_i \le c, \quad i = 1, \dots, l. \end{aligned} \tag{8}$$

Note that in the nonlinear case, the data will be transformed into a higher-dimensional space, and $\phi(x_i)$ will replace x_i. In this case, kernel trick $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ will be applied.

Solving the minimization (8), one can obtain u^* directly, and compute w^* and b^*, respectively, by

$$w^* = \sum_{i=1}^{l} u_i^* y_i x_i \quad b^* = t \sum_{i=1}^{l} u_i^* y_i. \tag{9}$$

The following self-evident proposition will clarify the optimal value of ξ.

Proposition 1: Let (z^*, ξ^*) and u^* be the optimal solutions to (6) and (8), respectively. Then for any $i = 1, \dots, l$

$$\xi_i^* = \begin{cases} 0 & u_i^* = 0 \\ 1 - y_i \big(w^{*T} x_i + b^*\big) & u_i^* \ne 0. \end{cases} \tag{10}$$

III. Theoretical Study of the GVSVM Solution

This section aims to study the properties of the solution of the GVSVM and discuss its different features. Note that different values of $t > 0$ in the GVSVM results in different problems. Therefore, an infinite number of problems are available based on the value of t. Although the objective functions of the GVSVM and the standard SVM are similar for a large value of t, there is no guarantee that their solutions tend to each other since even a subtle change in the objective function and its corresponding gradient can lead to totally distinct solutions. On top of that the bias term b is obtained by the GVSVM, which needs to be proved that it is the same bias as the standard SVM. A by-product of the forthcoming proofs is a closed-form solution for the bias term b.

We will first show that there exists an optimal solution for all these problems. To do so, let the standard SVM be reformulated as

$$\begin{aligned} P_\infty: \quad \min \ & P_\infty(z,\xi) = \frac{1}{2} z^T Q_\infty z + C^T \xi \\ \text{s.t.} \ & \mathbf{1}_{l\times 1} - Az - \xi \le 0 \\ & \xi \ge 0 \end{aligned} \tag{11}$$

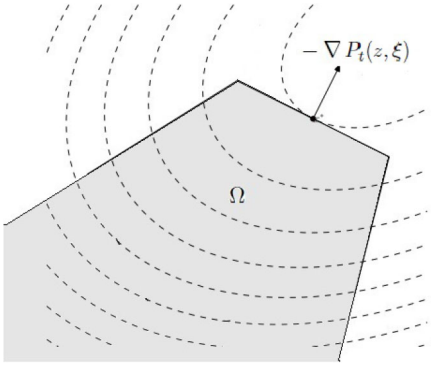


Fig. 1. Descent direction of the optimal solution.

where Q_∞ is the matrix Q_t in which the term $(1/t)$ is replaced by zero.

Theorem 1: Let Ω be the set of all feasible solutions (z, ξ). For any $t > 0$, there exists a unique solution for GVSVM. Moreover, the optimal solutions to the problems (6) and (11) are achieved on the boundaries.

Proof: According to [32], minimization (11) has a unique answer on the Ω. Now, let $(z^{*T}, \xi^{*T})^T$ be the optimal solution of the problem (11), then for any $(z, \xi) \in \Omega$ we have

$$\begin{aligned} 0 \leq \frac{1}{2} z^{*T} Q_\infty z^* + C^T \xi^* &\leq \frac{1}{2} z^T Q_\infty z + C^T \xi \\ &\leq \frac{1}{2} z^T Q_t z + C^T \xi. \end{aligned}$$

It shows that the objective value of the problem (6) is bounded from below for any $t \in \mathcal{R}$. Hence, the optimal value is finite, and there exists an optimal solution for such an optimization. On the other hand, since Q_t is positive definite for any $t \in \mathcal{R}$, the objective function of the minimization (6) is strictly convex. Therefore, the optimal solution is unique.

Moreover, we have $(Q_\infty z, C) \neq 0$ and $(Q_t z, C) \neq 0$ for any arbitrary $(z, \xi) \in \text{int}(\Omega)$. Thus, there exist descend directions $-(Q_\infty z, C)$ and $-(Q_t z, C)$. Moving in these directions, the objective values of problems (11) and (6) will decrease until they take their optimal solutions on the boundaries (see Fig. 1). ■

This theorem not only guarantees the optimal solution of the GVSVM but also indicates its uniqueness. Taking advantage of this theorem, we construct a convergent net with respect to the solutions of the problem (6) for different values of $t > 0$. Then, it will be demonstrated that these nets tend to the solution of the minimization (1) as $t \longrightarrow \infty$. First, several basic definitions about nets are required.

Definition 1 (Directed Set [33]): A directed set is a set D ordered by a preorder relation $\leq$ (a reflexive, transitive binary relation), such that every two elements in D have an upper bound in D. That is,

$$\forall a, b \in D, \exists k \in D;\ \ a \leq k,\ \ b \leq k.$$

For instance, the set $\mathcal{R} = (\mathbb{R}^{>0}, \leq)$ is a directed set on which our desired nets are built.

Definition 2 (Net [33]): A net in a set X is a function $f : D \longrightarrow X$, where D is a directed set. We write $x_d = f(d)$ for all $d \in D$ and denote the net as $(x_d)_{d \in D}$. Furthermore, the net $(x_d)_{d \in D}$ is said to be convergent to a point $a \in D$, and is written $(x_d)_{d \in D} \longrightarrow a$, if

$$\forall \epsilon > 0, \exists s \in D \ \ \forall e \geq s \Longrightarrow x_e \in B_\epsilon(a)$$

where $B_\epsilon(a)$ is the open ϵ-ball of a.

Theorem 2 (Monotone Convergence [34]): Every increasing net in $\mathbb{R}$ which is bounded from above is convergent to its supremum. Moreover, the limit of a net is unique.

As it can be readily seen, the concepts of nets are more general than sequences, where sequences are defined on a countable set $\mathbb{N}$ while nets are defined on a directed set which can be uncountable. Indeed, every sequence can be considered as a net and most of their properties can be expanded for nets. Now, we may proceed to build the convergent nets corresponding to the solutions of the GVSVM for different values of t.

Lemma 1: Let $(w_t^T, b_t, \xi_t^T)^T$ be the optimal solutions to problems (6) and (11) for any $t \in \mathcal{R}$ and $(w^{*T}, b^*, \xi^{*T})^T$, respectively. Then, $(w_t)_{t \in \mathcal{R}}$, $(b_t)_{t \in \mathcal{R}}$, and $(\xi_t)_{t \in \mathcal{R}}$ are convergent nets such that:

1) $(b_t)_{t \in \mathcal{R}} \longrightarrow b^*$;
2) $(w_t)_{t \in \mathcal{R}} \longrightarrow w^*$;
3) $(\xi_t)_{t \in \mathcal{R}} \longrightarrow \xi^*$.

Proof: Suppose $t_1, t_2 \in \mathcal{R}$ and $t_1 < t_2$. Let (z_{t_1}, ξ_{t_1}) and (z_{t_2}, ξ_{t_2}) be the optimal solutions to problems P_{t_1} and P_{t_2}, respectively. Since (z_{t_1}, ξ_{t_1}) is optimal for P_{t_1} and (z_{t_2}, ξ_{t_2}) is a feasible solution, then

$$\begin{aligned} \frac{1}{2} z_{t_1}^T Q_{t_1} z_{t_1} + C^T \xi_{t_1} &= \frac{1}{2}\left(w_{t_1}^T w_{t_1} + \frac{1}{t_1} b_{t_1}^2 \right) + C^T \xi_{t_1} \\ &< \frac{1}{2} z_{t_2}^T Q_{t_1} z_{t_2} + C^T \xi_{t_2} \\ &= \frac{1}{2}\left(w_{t_2}^T w_{t_2} + \frac{1}{t_1} b_{t_2}^2 \right) + C^T \xi_{t_2}. \end{aligned} \tag{12}$$

Similarly, for the problem P_{t_2}, we have

$$\begin{aligned} \frac{1}{2} z_{t_2}^T Q_{t_2} z_{t_2} + C^T \xi_{t_2} &= \frac{1}{2}\left(w_{t_2}^T w_{t_2} + \frac{1}{t_2} b_{t_2}^2 \right) + C^T \xi_{t_2} \\ &< \frac{1}{2} z_{t_1}^T Q_{t_2} z_{t_1} + C^T \xi_{t_1} \\ &= \frac{1}{2}\left(w_{t_1}^T w_{t_1} + \frac{1}{t_2} b_{t_1}^2 \right) + C^T \xi_{t_1}. \end{aligned} \tag{13}$$

The following inequality can be obtained by adding above inequalities:

$$\left(\frac{1}{t_1} - \frac{1}{t_2} \right) b_{t_1}^2 < \left(\frac{1}{t_1} - \frac{1}{t_2} \right) b_{t_2}^2.$$

From this inequality, it can be deduced that $b_{t_1}^2 < b_{t_2}^2$, which implies $(b_t^2)_{t \in \mathcal{R}}$ to be a nonmonotonous increasing net. It is obvious that b^{*2} is an upper bound for this net because otherwise assume that there exists $t_i \in \mathcal{R}$ which $b^{*2} < b_{t_i}^2$. Since $(1/2) w^{*T} w^* + C^T \xi^* < (1/2) w_{t_i}^T w_{t_i} + C^T \xi_{t_i}$ then

$$\frac{1}{2}\left(w^{*T} w^* + \frac{1}{t_i} b^{*2} \right) + C^T \xi^* < \frac{1}{2}\left(w_{t_i}^T w_{t_i} + \frac{1}{t_i} b_{t_i}^2 \right) + C^T \xi_{t_i}$$

and it contradicts with the optimality of (z_{t_i}, ξ_{t_i}) for P_{t_i}. As $(b_t^2)_{t \in \mathcal{R}}$ is nonmonotonous increasing and is bounded from

above, it is convergent to its supremum $\bar{b}^2$ according to Lemma 2. Next, we will prove that $\bar{b} = b^*$.

For $\bar{b}$, there are $\bar{w}$ and $\bar{\xi}$ such that $(\bar{z}^T, \bar{\xi}^T)^T \in \Omega$ is the optimal solution of the problem (6) as $t \longrightarrow \infty$. Since $(z^{*T}, \xi^{*T})^T \in \Omega$, according to [31, Th. 3.4.3], we have

$$\begin{pmatrix} Q_t\bar{z} \\ C \end{pmatrix}^T \begin{pmatrix} z^* - \bar{z} \\ \xi^* - \bar{\xi} \end{pmatrix} \geq 0$$

or equivalently

$$\begin{pmatrix} \bar{w} \\ \frac{1}{t}\bar{b} \\ C \end{pmatrix}^T \begin{pmatrix} w^* - \bar{w} \\ b^* - \bar{b} \\ \xi^* - \bar{\xi} \end{pmatrix} \geq 0. \tag{14}$$

Similarly, (z^*, ξ^*) is optimal for the minimization (11), then

$$\begin{pmatrix} w^* \\ 0 \\ C \end{pmatrix}^T \begin{pmatrix} \bar{w} - w^* \\ \bar{b} - b^* \\ \bar{\xi} - \xi^* \end{pmatrix} \geq 0. \tag{15}$$

By adding (14) and (15), it is obtainable that

$$-\left\|w^* - \bar{w}\right\|^2 \geq \frac{1}{t}\bar{b}\big(\bar{b} - b^*\big). \tag{16}$$

By $t \longrightarrow \infty$, we get

$$\frac{1}{t}\bar{b}\big(\bar{b} - b^*\big) \longrightarrow 0, \implies \left\|w^* - \bar{w}\right\| \longrightarrow 0 \tag{17}$$

$$\implies w^* = \bar{w}. \tag{18}$$

Further, substitute $w^* = \bar{w}$ in (14) and (15), then

$$\frac{1}{t}\bar{b}\big(b^* - \bar{b}\big) + C^T\big(\xi^* - \bar{\xi}\big) \geq 0, \quad C^T\big(\bar{\xi} - \xi^*\big) \geq 0 \tag{19}$$

hence

$$\frac{1}{t}\bar{b}\big(b^* - \bar{b}\big) \geq C^T\big(\bar{\xi} - \xi^*\big) \geq 0. \tag{20}$$

As a result, we get $\xi^* = \bar{\xi}$ by $t \longrightarrow \infty$.

Now, by contradiction, suppose $\bar{b} \neq b^*$, since $\|\bar{w}\|^2 = \|w^*\|^2$. Then (z^*, ξ^*) and $(\bar{z}, \bar{\xi})$ are both optimal for problem (11) which contradicts with the uniqueness of the standard SVM solution. Therefore, $\bar{b} = b^*$ and $(b_t^2)_{t\in\mathcal{R}} \longrightarrow b^{*2}$.

Now, we prove that $(b_t)_{t\in\mathcal{R}} \longrightarrow b^*$. Since for every $t \in \mathcal{R}$, (z_t, ξ_t) satisfies (16), we get $(1/t)\bar{b}^2 \leq \bar{b}b^*$. It shows that the sign of all b_t, $t \in \mathcal{R}$, and b^* are the same. Hence, $(b_t^2)_{t\in\mathcal{R}} \longrightarrow b^{*2}$ implies $(b_t)_{t\in\mathcal{R}} \longrightarrow b^*$ as $t \longrightarrow \infty$.

2), 3) From (12) and (13), it is obtainable that $(1/2)\|w_{t_1}\|^2 + C^T\xi_{t_1} > (1/2)\|w_{t_2}\|^2 + C^T\xi_{t_2}$, which implies a nonmonotonous decreasing net $((1/2)\|w_t\|^2 + C^T\xi_t)_{t\in\mathcal{R}}$.

The non-negativity of $(1/2)\|w_t\|^2$ and $C^T\xi_t$, caused the net $((1/2)\|w_t\|^2 + C^T\xi)_{t\in\mathcal{R}}$ to be separated into two nonmonotonous decreasing nets $(\|w_t\|)_{t\in\mathcal{R}}$ and $(\xi_t)_{t\in\mathcal{R}}$.

Since (16) and (20) are valid for (z_t, ξ_t) as $t \longrightarrow \infty$, then $\|w_t - w^*\| \longrightarrow 0$ and $\|\xi_t - \xi^*\| \longrightarrow 0$. It indicates that $(w_t)_{t\in\mathcal{R}} \longrightarrow w^*$ and $(\xi_t)_{t\in\mathcal{R}} \longrightarrow \xi^*$, which completes the proof. ■

Although this lemma proves the convergence of the variables of the GVSVM solution to the optimum of the standard SVM, the rates of convergence might be different. Furthermore, the convergence of net $(z_t, \xi_t)_{t\in\mathcal{R}}$ is still needed to be generalized. Next theorem guarantees that the solution of the minimization (6) tends to the solution of the problem (11) as $t \longrightarrow \infty$.

Theorem 3: Let $(z_t^T, \xi_t^T)^T$ for any $t \in \mathcal{R}$ and $(z^{*T}, \xi^{*T})^T$ be the optimal solutions to problems (6) and (11), respectively. Then, $(z_t, \xi_t)_{t\in\mathcal{R}} \longrightarrow (z^*, \xi^*)$ as $t \longrightarrow \infty$. Moreover, $\lim_{t\longrightarrow\infty} \inf_\Omega P_t(z, \xi) = \inf_\Omega P_\infty(z, \xi)$.

Proof: According to Lemma 1, solutions $(z_t, \xi_t)_{t\in\mathcal{R}}$ are convergent nets in $\mathbb{R}^{n+1} \times \mathbb{R}^n$. We prove that $(z_t, \xi_t)_{t\in\mathcal{R}} \longrightarrow (z^*, \xi^*)$. Let $\epsilon > 0$ be arbitrary, since $(b_t)_{t\in\mathcal{R}} \longrightarrow b^*$, then

$$\exists N_1 \in \mathcal{R} \;\; \forall t \geq N_1, \quad \left|b_t - b^*\right| < \frac{\epsilon}{3}.$$

Similarly, by $(w_t)_{t\in\mathcal{R}} \longrightarrow w^*$ and $(\xi_t)_{t\in\mathcal{R}} \longrightarrow \xi^*$, we have

$$\exists N_2 \in \mathcal{R} \;\; \forall t \geq N_2, \quad \left\|w_t - w^*\right\| < \frac{\epsilon}{3}$$

and

$$\exists N_3 \in \mathcal{R} \;\; \forall t \geq N_3, \quad \left\|\xi_t - \xi^*\right\| < \frac{\epsilon}{3}.$$

Now let $N = \max\{N_1, N_2, N_3\}$, then for any $t \geq N$

$$\begin{aligned} \left\|\big(z_t - z^*, \xi_t - \xi^*\big)\right\| &\leq \left\|w_t - w^*\right\| + \left|b_t - b^*\right| + \left\|\xi_t - \xi^*\right\| \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon \end{aligned}$$

which indicates $(z_t, \xi_t)_{t\in\mathcal{R}} \longrightarrow (z^*, \xi^*)$. Note that the first inequality is valid because $\sqrt{a^2 + b^2} \leq |a| + |b|$, for any $a, b \in \mathbb{R}$. Moreover

$$\begin{aligned} \lim_{t\longrightarrow\infty} \inf_\Omega(P_t(z, \xi)) &= \lim_{t\longrightarrow\infty} \frac{1}{2} z_t^T Q_t z_t + C^T\xi_t \\ &= \lim_{t\longrightarrow\infty} \frac{1}{2} w_t^T w_t + \frac{1}{2t} b_t^2 + C^T\xi_t \\ &= \frac{1}{2} w^{*T} w^* + C^T\xi^* = \inf_\Omega(P_\infty(z, \xi)). \end{aligned}$$

■

Corollary 1: Based on the foregoing theorem, $(z_t, \xi_t)_{t\in\mathcal{R}} \longrightarrow (z^*, \xi^*)$ as $t \longrightarrow \infty$. Hence, the larger t would result in more proximity of the GVSVM solution to the standard SVM.

Taking t to be sufficiently large will guarantee the analogy of the GVSVM and the standard SVM solutions. To discuss this property, the performance of the GVSVM for different values of t is examined in the forthcoming sections. The empirical results also illustrate the same outcome as what the theoretical studies suggest.

IV. Efficient Neural Network and Its Convergence

In this section, an efficient RNN is proposed to solve the GVSVM dual problem, and is proved to be asymptotically stable in the sense of Lyapunov and is globally exponentially convergent to the solution of GVSVM. We further juxtapose the proposed neural network with the existing ones and demonstrate that it is more efficient in terms of architecture and complexity.

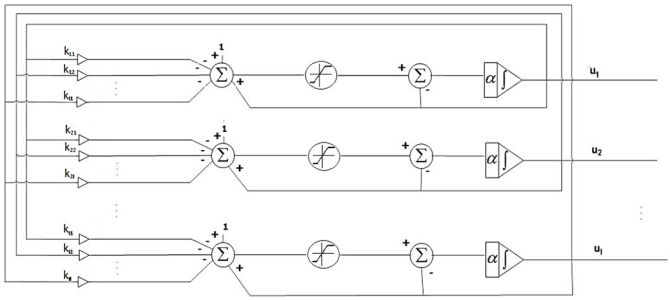


Fig. 2. Block diagram of the proposed RNN (23).

A. Neural Network With One-Layer Architecture

Consider the GVSVM dual problem with a kernel function $k(.,.)$

$$\begin{aligned} \min_{u} \ & \frac{1}{2}u^T\hat{K}u - e^T u \\ \text{s.t.} \ & 0 \le u \le ce \end{aligned} \tag{21}$$

where $e \in R^l$ is a vector whose elements are 1, and $\hat{K}$ is a matrix with elements $\hat{K}_{ij} = y_i y_j (K(x_i, x_j) + t/2)$. It is evident that matrix $\hat{K}$ is positive definite since the kernel function is positive semidefinite and $t > 0$. Based on this formulation, the following theorem is obtained.

Theorem 4: u^* is the optimal solution of the minimization (21) if and only if

$$P_\Omega\Big(u - \Big(\hat{K}u - e\Big)\Big) = u \tag{22}$$

where P_Ω is an element-wise operator defined as

$$(P_\Omega(\gamma))_i = \begin{cases} c, & \gamma_i > c \\ \gamma_i, & 0 \le \gamma_i \le c \\ 0 & \gamma_i < 0. \end{cases}$$

Proof: Equation (22) is easily obtained from the KKT conditions of the minimization (21). ■

Based on this theorem, an RNN is proposed whose dynamical equation is

$$\frac{du}{dt} = \alpha\Big(-u + P_\Omega\Big(u - \Big(\hat{K}u - e\Big)\Big)\Big) \tag{23}$$

where $\alpha > 0$ is a scaling parameter. The RNN can be restated in the element form as

$$\frac{du_i}{dt} = -\alpha u_i + \alpha P_\Omega\Big(u_i - \Big(\hat{K}^i u - 1\Big)\Big)$$

where $\hat{K}^i$ is the ith row of the matrix $\hat{K}$. This dynamical system can be easily recognized as a single-layer RNN depicted in Fig. 2.

We first guarantee the convergence and stability of the dynamical system (23), and then it is contrasted with other RNNs for the standard SVM.

B. Convergence Analysis

In this section, the proposed RNN is first proved to be asymptotically stable in the sense of Lyapunov. It is further investigated that it is globally exponentially convergent to the solution of the GVSVM, and the rate of convergence is reliant on the scaling parameter α. We first begin with several basic definitions, which are the building blocks of the upcoming proofs.

Definition 3: A continuous-time neural network is globally convergent if its trajectory tends to an equilibrium point for any given arbitrary initial point. A dynamic system du/dt is globally exponentially convergent to a point u^*, if for any initial point

$$\big\|u(t) - u^*\big\| \le \beta_1 e^{\gamma(t-t_0)} \quad \forall t \ge t_0 \tag{24}$$

Lemma 2 [35]: For the closed convex set $\Omega \in R^N$, we have

$$\begin{aligned} &(i) \quad (v - P_\Omega(v))^T (P_\Omega(v) - x) \ge 0, && w \in R^N, x \in \Omega \\ &(ii) \quad \|P_\Omega(u) - P_\Omega(v)\| \le \|u - v\|, && u, v \in R^N. \end{aligned}$$

Lemma 3: There exists a unique continuous solution for the neural network (23) for an arbitrary initial point. Further, its equilibrium point solves the GVSVM dual problem (21).

Proof: According to Lemma 2, P_Ω is Lipschitz continuous, so is the right-hand side of the system (23). Hence, there is a unique continuous solution $u(t)$ according to the Peano's theorem [36]. Moreover, the equilibrium point of the neural network (23) solves the problem (21), thanks to Lemma 4. ■

Theorem 5: The proposed neural network (23) with the arbitrary initial point u_0 is asymptotically stable in the sense of Lyapunov and globally converges to the solution of GVSVM.

Proof: Consider the following Lyapunov function:

$$V(u) = G(u)^T F(u) - \frac{1}{2}\|F(u)\|^2 + \frac{1}{2}\big\|u - u^*\big\|$$

where u^* is the equilibrium of the dynamical system (23), and

$$G(u) = Qu + 1, \qquad F(u) = -u + P_\Omega(u - (Qu + 1)).$$

We first investigate essential inequalities for the projection operator $P_\Omega(.)$. In the first inequality of Lemma 2, let $w = u - G(u)$ and $x = u^*$, then

$$\begin{aligned} &(-F(u) - G(u))^T\big(F(u) - u - u^*\big) \ge 0 \\ &\Rightarrow -G(u)^T\big(u - u^*\big) - \|F(u)\|^2 \ge F(u)^T\big(G(u) + u - u^*\big). \end{aligned} \tag{25}$$

Having this inequality under the belt, the derivation of the Lyapunov function with respect to u is obtained as [37]

$$\frac{dV}{du} = G(u) - (\nabla G(u) - I)F(u) + \big(u - u^*\big) \tag{26}$$

where I denotes the identity matrix and $\nabla G(u) = \hat{K}$. It follows:

$$\begin{aligned} \frac{dV(u)}{dt} &= \left(\frac{dV(u)}{du}\right)^T \frac{du}{dt} \\ &= \alpha\big(G(u) - (\nabla G(u) - I)F(u) + \big(u - u^*\big)\big)^T F(u) \\ &\le \alpha\big(G(u) + u - u^*\big)^T F(u) + \alpha\|F(u)\|^2 \\ &\quad - \alpha F(u)^T \nabla G(u) F(u) \\ &\le^{(1)} -\alpha G(u)^T\big(u - u^*\big) - \alpha F(u)^T \nabla G(u) F(u) \\ &<^{(2)} 0 \end{aligned}$$

where (1) is deduced by (25) and (2) is correct since $\nabla G = \hat{K}$ is positive definite and $G(u)^T(u - u^*) \ge 0$. Therefore, the

dynamical system (23) is asymptotically stable in the sense of Lyapunov.

For the global convergence, consider again the first inequality of Lemma 2 with $w = u - \alpha G(u)$ and $x = u$. Then

$$G(u)^T F(u) \leq -\|F(u)\|^2$$

which follows:

$$V(u) \geq \frac{1}{2}\|u - u^*\|^2 + \frac{1}{2}\|F(u)\|^2 \geq \frac{1}{2}\|u - u^*\|^2. \quad (27)$$

Thus, the trajectory of solution is bounded for any given initial point. According to the invariant set theorem [38], all trajectories of the system (23) converge to a largest invariant set Ψ where $dV(u)/dt = 0$. We need to show that $dV(u)/dt = 0$ if and only if $du/dt = 0$. If $du/dt = 0$, then

$$\frac{d}{dt}V(u) = \left(\frac{dV}{du}\right)^T \frac{du}{dt} = 0. \quad (28)$$

For $\hat{u} \in \psi$, $dV/dt = 0$ implies

$$G(\hat{u})^T(u - u^*) + F(\hat{u})^T \nabla G(\hat{u}) F(\hat{u}) = 0. \quad (29)$$

Both terms in this equation are non-negative, hence the equality holds if and only if both are zero. Thus

$$F(\hat{u})^T \nabla G(\hat{u}) F(\hat{u}) = 0 \overset{\nabla G \geq 0}{\Longrightarrow} F(\hat{u}) = 0$$
$$\Longrightarrow \frac{d\hat{u}}{dt} = \hat{u} - P_\Omega(\hat{u} - (Q\hat{u} + 1)) = F(\hat{u}) = 0. \quad (30)$$

Therefore, the dynamical system (23) converges globally to the solution of the GVSVM dual problem. ■

Theorem 6: The proposed neural network (23) is exponentially convergent to the optimal solution of the GVSVM. Further, the rate of convergence is commensurate with α.

Proof: Considering the same Lyapunov function in Theorem 5, the following inequality is obtainable:

$$\|F(u)\|\|I + Q\| \geq \|Q\|\|u - u^*\| \quad (31)$$

where u^* is the equilibrium point of the system (23). Based on the proof of Theorem 5, we have

$$\frac{dV}{dt} = \left(\frac{dV}{du}\right)^T \frac{du}{dt} \leq -\alpha F(u)^T \nabla G(u) F(u)$$
$$\overset{\text{integration}}{\Longrightarrow} E(u(t)) \leq E(u(t_0)) - \alpha \int_{t_0}^{t} F(u(s))^T Q F(u(s)) ds$$
$$\leq E(u(t_0)) - \alpha\|\hat{K}\| \int_{t_0}^{t} \|F(u(s))\|^2 ds$$
$$\leq E(u(t_0)) - \alpha \frac{\|\hat{K}\|^3}{\|I + \hat{K}\|^2} \int_{t_0}^{t} \|u - u^*\|^2$$
$$\leq E(u(t_0)) - \frac{\alpha\|\hat{K}\|^3}{\|I + \hat{K}\|^2} \int_{t_0}^{t} \|u - u^*\|^2$$
$$\leq 2E(u(t_0))e^{-\rho(t-t^0)}, \quad \rho = \frac{\alpha\|\hat{K}\|^3}{\|I + \hat{K}\|^2}$$

where the last inequality is obtained from the Gronwall inequality [38]. It follows that:

$$\|u(t) - u^*\| \leq 2E(u(t)) \leq 4E(u(t_0))e^{-\rho(t-t^0)}. \quad (32)$$

Thus, the proposed RNN is exponentially convergent to the solution of the GVSVM. The convergence rate can be increased merely by increasing α since it is commensurate with ρ. ■

C. Comparison With Other Neural Networks

The proposed RNN is now compared with the existing neural solutions for the standard SVM. The first model considered here is proposed in [39] and [40], and is a two-layer RNN for the standard SVM with the dynamical system being given by

$$\frac{d}{dt}\begin{pmatrix} u \\ \mu \end{pmatrix} = \alpha\begin{pmatrix} -ee^T\hat{u} + (I + \tilde{K})(P_X(\tilde{u} - (\tilde{K}u + e\mu - y)) - \tilde{u}) \\ -e^T P_X(\tilde{u} - (\tilde{K}\tilde{u} + e\mu - y)) \end{pmatrix} \quad (33)$$

where $\tilde{K}$ is a matrix with elements $\tilde{K}_{ij} = y_i y_j K(x_i, x_j)$, μ is an auxiliary variable, $\tilde{u}_i = y_i u_i$, and $P_X(.)$ is a projection function with the definition

$$P_X(a_i) = \begin{cases} d_i, & a_i < d_i \\ a_i, & d_i \leq a_i \leq h_i \\ h_i, & a_i > h_i \end{cases}$$

where $d_i = -c(\text{sign}(1 - y_i))$, $h_i = c(\text{sign}(1 + y_i))$, and $X = \{a \in \mathbb{R}^l | d \leq a \leq h\}$. The dynamical system (33) is recognized by a neural network with a two-layer structure.

Nazemi and Dehghan [41] proposed another neural network for the training of the SVM. The dynamic equation of their proposed network is

$$\frac{d}{dt}\begin{pmatrix} u \\ v \\ \mu \end{pmatrix} = \alpha\begin{pmatrix} \tilde{K}u + \mathbf{1} + H^T(v + Hu - g)^+ + 1\mu \\ (v + Hu - g)^+ - v \\ e^T u \end{pmatrix} \quad (34)$$

where $(x)^+ = \max(0, x)$, $H = (I - I)^T \in R^{2l \times l}$, $g = (0\mathbf{1}c) \in R^{2l}$, and $v \in R^{2l}$. This neural solution is guaranteed to converge globally to its equilibrium.

A simpler one-layer neural solution is proposed in [42] with the dynamical equation

$$\frac{d}{dt}\begin{pmatrix} u \\ \mu \end{pmatrix} = \alpha\begin{pmatrix} P_X(u - (\tilde{K}u + e\mu - y)) - u \\ e^T u \end{pmatrix}. \quad (35)$$

This is a simple one-layer network which is proved to converge globally exponentially to the optimal solution provided that $\tilde{K}$ is positive definite.

Yang *et al.* [43] developed another neural solution whose convergence is not reliant on the positive definiteness of the kernel function. The dynamic system describing their neural network is

$$\frac{d}{dt}\begin{pmatrix} u \\ \mu \end{pmatrix} = \alpha\begin{pmatrix} \mathbf{1} - \tilde{K}u - \mu y + D(P_X(u) - u) \\ e^T u \end{pmatrix} \quad (36)$$

where D is a positive diagonal matrix.

The above neural networks are now compared with the proposed neural network (23) concerning their structures, their convergence, and the number of operations in each iteration. The proposed neural network needs l^2 multiplications and $l^2 + 2l$ additions/subtractions in each iteration. The neural networks (33) and (34) require $3l^2 + 2l$ and $5l^2 + 2l$ multiplications, and $3l^2 + 4l$ and $7l^2 + 2l - l$ additions/subtractions, respectively. Similarly, the neural networks

TABLE I
NUMBER OF OPERATIONS REQUIRED IN EACH ITERATION OF THE PROPOSED NEURAL NETWORK ALONG WITH FOUR OTHER NETWORKS IN THE LITERATURE

Method	Multiplication	Additions/Subtractions
Tan et al. [40]	$3l^2+2l$	$3l^2+4l$
Xia et al. [42]	l^2+2l	l^2+4l
Nazemi et al. [41]	$5l^2+2l$	$7l^2+3l-1$
Yang et al. [43]	$2l^2+2l$	$2l^2+2l-1$
Proposed network	l^2	l^2+2l

TABLE II
NUMBER OF COMPONENTS REQUIRED FOR THE CIRCUIT IMPLEMENTATION OF THE PROPOSED NEURAL NETWORK ALONG WITH FOUR OTHER NETWORKS IN THE LITERATURE

Method	Summers	Integerators	Activation	Weight connections
Tan et al. [40]	$4l+1$	$2l$	$2l$	l^2+5l
Xia et al. [42]	$2l+1$	$l+1$	l	$l(l+2)$
Nazemi et al. [41]	$5l+1$	$3l+1$	$3l$	$4l^2$
Yang et al. [43]	$2l+1$	$l+1$	l	$l(l+3)$
Proposed network	$2l$	l	l	l^2

with the dynamic systems (35) and (36) need $2l^2+4l$ and $2l^2+2l-1$ additions, and l^2+2l and $2l^2+2l$ multiplications, respectively.

The structure of the proposed neural solution can be implemented by $2l$ summers, l integrator, l piecewise activation functions, and l^2 weight connections. In contrast, the model in (33) needs $2L$ integrator, $2l$ piecewise activation function, and $l(l+3)+l(2l+1)$ summers and connection weights. The neural network in (34) can be implemented by $5l+1$ summers, $3l+1$ integrators, $3l$ activation functions, and $4l^2$ weight connections. By the same token, the neurodynamic model in (35) requires $l+1$ integrator, l piecewise activation function, $l(l+3)$ summers, and $l(l+2)$ weight connections. The network in (36) has the same requirement with having required l more summers. Therefore, the proposed neural network is superior to those in (33) and (35) from the structural complexity and the computations in each iteration.

Tables I and II tabulate the number of operations in each iteration and the components required for circuit implementation, respectively. According to this table, the proposed neural network has a simpler architecture and is more time efficient since it needs fewer operations in each iteration.

Regarding the convergence rate, the proposed system in (23) is globally exponentially convergent while the models in (33), (34), and (36) are globally convergent but not exponentially. The system (35) is also promised to converge exponentially provided that the kernel function is positive definite. However, the conjecture could be violated if the kernel function is positive semidefinite, or there exist repetitive data points in the dataset. Hence, the exponential convergence of the neural network is not guaranteed. The proposed system (23) is globally exponentially convergent, regardless of what kernel function is utilized.

Last but not least, the bias term can be obtained directly from the solution of our neural network while other models would need to approximate it.

V. EXPERIMENTS

The experiments regarding the proposed neural network are investigated in this section. First, the convergence of the neural network in (23) is empirically examined, and it is followed by a toy example scrutinizing the closeness of the standard SVM and the GVSVM solutions for various values of t. Then, the classification of real datasets is performed by different standard SVM solvers, and the related results are reported.

A. Empirical Convergence Analysis

As a complement to the theoretical study in Section IV-B, we inspect the convergence of the neural network in practice. To this end, the *wine* benchmark is used which consists of 178 samples in three different classes. The samples corresponding to two classes are selected, and the classification using the proposed model is performed.

The convergence must be probed into by different initializations. This is done by taking the initial point as a vector of zero, one, and a randomly generated vector. Fig. 3 displays the transient behavior of the proposed neural network with different initial points and $\alpha = 10$. The x-axis of this figure represents the iterations and y-axis is the value of elements in the vector u. It is evident that the trajectory of the neural network converges to the same values regardless of the initial point. This corroborates the global convergence of the neural network in that the initial point is of no matter.

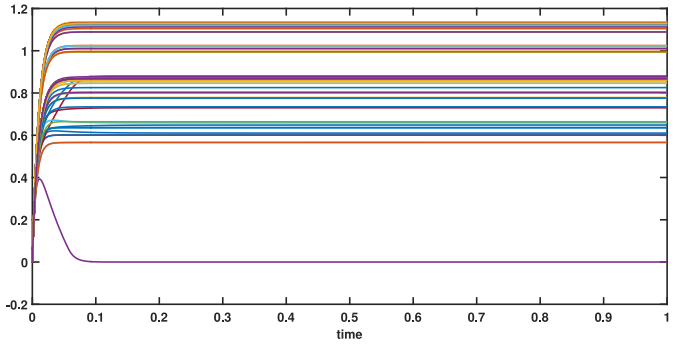
The convergence rate of the system (23) is further investigated via the *energy error*. The energy error of the proposed neural network with respect to the state u is defined as

$$\mathrm{ER}(u) = \left\| u - P_\Omega(u - (\hat{K}u - 1)) \right\|^2.$$

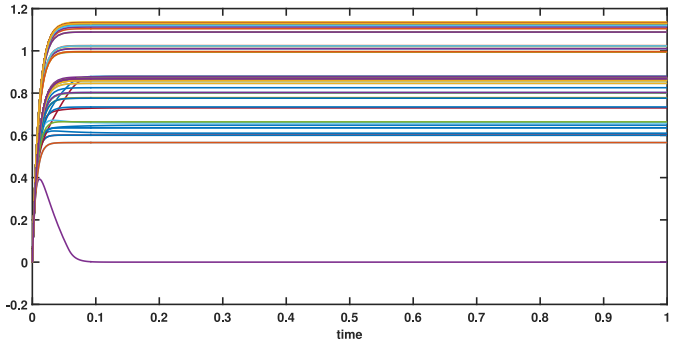
According to the discussions in Section IV-B, $\mathrm{ER}(u^*) = 0$ if and only if u^* is an optimal solution. We repeat the experiment over the *wine* benchmark in which the values of α are set to be 10, 15, and 20. Fig. 4 displays the transient behavior of the energy error with three values of α. It is trivial that the bigger values of α will increase the convergence rate of the neural network. Thus, the energy error swiftly tends to zero for larger α which reinforces the dependency of the convergence rate to α.

B. Toy Example

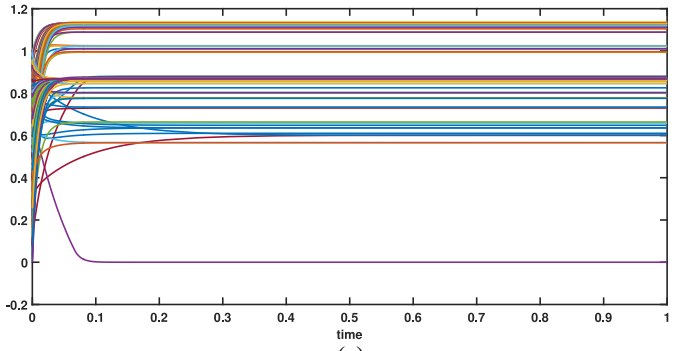
In this section, the proximity of solutions of the standard SVM and GVSVM is empirically explored. To do so, the Fisher's Iris dataset is selected, and the standard SVM and GVSVM are applied to this classification task. The Fisher's Iris dataset includes 150 data points of three different classes. To better visualize the results, we take two linearly separable features with the data points of two classes. Figs. 5 and 6 plot the desired hyperplanes obtained by the standard SVM and GVSVM with different values of t. It is readily seen that the solution of GVSVM tends to the solution of standard SVM as the value of t increases. For $t = 1000$ in Fig. 5 and $t = 10$ in Fig. 6, the solutions of GVSVM and the standard SVM are precisely the same, and their corresponding separating hyperplanes lie on each other.



(a)



(b)



(c)

Fig. 3. Empirical convergence of the neural network in (23) with distinct initializations and $\alpha = 10$. (a) With the initialization $u = 1$. (b) With the initialization $u = 0$. (c) With the random initialization. The x-axis is the number of iteration and y-axis is the value of an element of u.

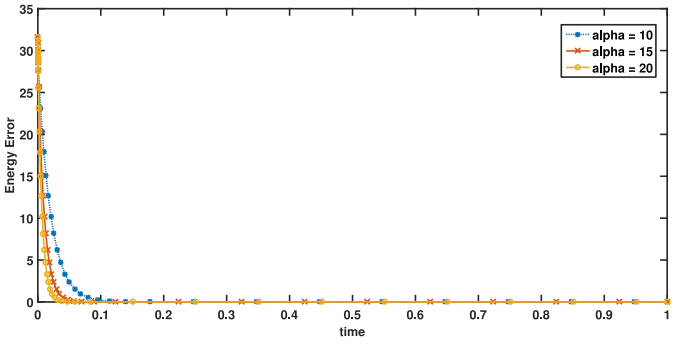


Fig. 4. Behavior of the proposed network in (23) in terms of the energy error on the *wine* benchmark for three different values of α.

On important point in these figures is the magnitude of t in each figure. The value 10 is seemingly big for the first case while the quantity 1000 is viewed as large enough for

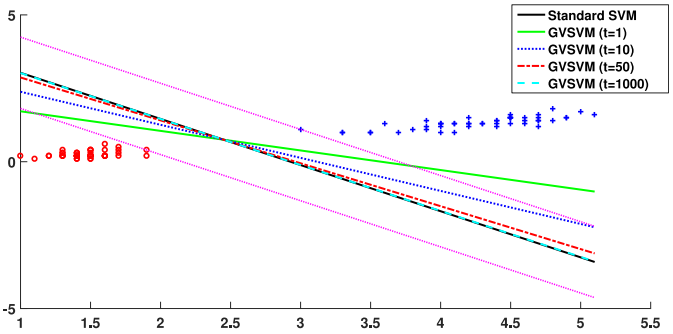


Fig. 5. Separating hyperplane of the standard SVM and GVSVM over the Fisher's Iris dataset (features 1, 3) for $t = 1, 10, 50$, and 1000. For $t = 1000$, the separating hyperplane of the standard SVM and GVSVM lie on each other.

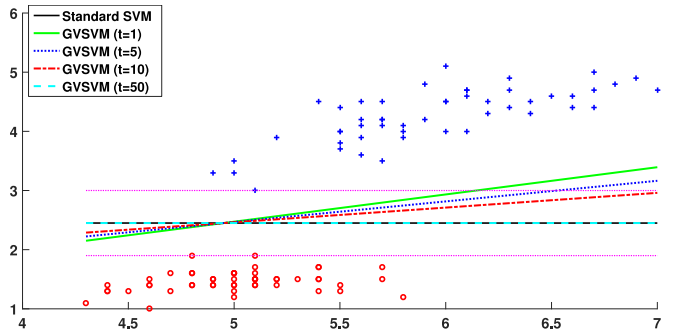


Fig. 6. Separating hyperplane of the standard SVM and GVSVM over the Fisher's Iris dataset (features 3, 4) for $t = 1, 5, 10$, and 50. For $t = 50$, the separating hyperplane of the standard SVM and GVSVM lie on each other.

TABLE III
AMOUNT OF $\|w_t\|$ AND $|b_t|$ FOR DIFFERENT VALUES OF t ON THE IRIS AND FISHER'S IRIS DATASETS FOR DIFFERENT FEATURES

	$\|w_t\|$		$\|b_t\|$	
dataset	F. Iris	Iris	F. Iris	Iris
t = 1	2.2890	0.2345	0.3560	0.0038
t = 5	2.1275	0.2342	1.4573	0.0188
t = 10	2.0101	0.2339	2.3671	0.0375
t = 50	1.8182	0.2316	4.4545	0.1830
t = 100	1.8182	0.2289	4.4545	0.3559
t = 1000	1.8182	0.2010	4.4545	2.3760
t = 10000	1.8182	0.1818	4.4560	4.4546
Stand. SVM	1.8182	0.1818	4.4566	4.4546

the second case. These experiments confirm the fact that the optimal value of t is highly related to the dataset under study. The safer way of selection t is to use the biggest value possible for the machine.

Further, the standard SVM and the GVSVM are applied to the data points of classes 1 and 2 of Fisher's Iris, and the convergence of the GVSVM is investigated. The data points are linearly separable so that the primal SVM is used for training. As a result, the optimal value of the bias term is also in hand, which makes the comparison possible. Fig. 7 shows the difference between the optimal solutions to the standard SVM and the GVSVM by various values of t. It is plain to see that the difference between their solutions is imperceptible when the value of t increases.

Moreover, Tables III presents a comparison between the optimal solutions to the standard SVM and the GVSVM. In

TABLE IV
COMPARISON OF THE PROPOSED RNN, LSVM [20], TAN RNN [40], XIA RNN [42], NAZEMI RNN [41], AND YANG RNN [43] IN TERMS OF THE ACCURACY, THE AVERAGE NUMBER OF ITERATION IN TENFOLD CROSS-VALIDATION, AND THE AVERAGE EXECUTION TIME OF PERFORMING TENFOLD CROSS-VALIDATION ON EACH BENCHMARK. THE DATASETS ARE OBTAINED FROM THE UCI REPOSITORY

Method	size	LSVM [20]			Tan RNN [40]			Xia Rnn [42]			Proposed RNN			Nazemi RNN [41]			Yang RNN [43]		
Dataset		Acc.	Iter.	Time	Acc.	Iter.	Time	Acc.	Iter.	Time	Acc.	Iter.	Time	Acc.	Iter.	Time	Acc.	Iter.	Time
scene	2407×294	93.24	13	256	90.87	70	807	92.64	85	512	95.12	9	263	91.24	99	1930	92.64	93	562
sick euthyriod	3163×42	90.73	24	310	88.33	69	833	90.73	68	530	90.73	12	458	88.33	88	3025	90.73	121	830
tyroid sick	3772×52	93.87	21	330	93.87	69	823	93.87	75	629	93.87	13	404	90.27	35	4306	93.87	72	604
ozone level	2536×72	97.12	25	412	97.12	66	970	97.12	71	736	98.2	20	530	97.12	82	2830	97.12	51	630
solar flare	1389×32	95.10	10	125	95.60	149	502	95.10	108	271	96.90	11	76	95.10	290	1152	95.10	77	106
german numer	1000×25	74.20	12	107	70.00	62	480	70.00	168	234	74.20	7	132	70.00	93	925	70.00	206	286
svmguide	3089×5	85.88	33	77	68.63	73	172	64.74	142	112	85.88	13	93	64.53	66	368	64.74	253	286
Adult	48842×14	84.5	78	410	-	-	-	84.4	232	1112	84.5	73	355	-	-	-	-	-	-
MNIST	70000×784	99.3	430	4850	-	-	-	99.3	232	4400	99.3	1415	8320	-	-	-	-	-	-

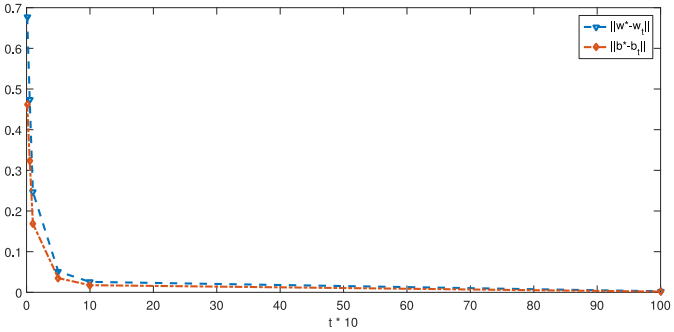


Fig. 7. Behavior of $\|w^* - w_t\|$ and $|b^* - b_t|$ on the Fisher's Iris dataset (features 3, 4).

this table, the amount of $\|w_t\|$ and $|b_t|$ appear for different values of t. For larger values of t, the solution of the GVSVM tends to the solution of the standard SVM, as our theoretical study suggested.

Figs. 5–7 and Table III imply the fact that the solution of the VSVM, in which $t = 1$, is significantly different from the standard SVM. Therefore, the utilization of the GVSVM is crucial in order to obtain a solution identical to the standard SVM.

C. Real Datasets

As the final experiment, the proposed neural network is applied to several classification tasks and its performance is compared with the LSVM [20] and neural networks in [40]–[43]. LSVM solves the VSVM with the least square loss function; thus, the resulting problem is more straightforward since it entails finding the solution of a linear system. Other neural networks are modeled based on the standard SVM. Another important point for the real problems is the selection of the kernel function. The type of kernel is reliant on the type of data we have. However, if there is no prior knowledge on the features of the given datasets, then the selection of the kernel function is not straightforward. Since the selected datasets for this experiment are well known to be nonlinearly separable, we use the radial basis function for all solvers. The RBF kernel function is defined as

$$K(x, y) = \exp\left(\frac{-\|x - y\|^2}{2\sigma^2}\right)$$

where σ is the width of the function. The optimal parameters for σ and c are obtained by the techniques in [45], and are identical for all algorithms of the SVM training. Further, the scaling parameter α is set to 10 for all neural networks.

For seven datasets, we use the tenfold cross-validation and gauge the test accuracy of six foregoing algorithms. The first seven rows of Table IV tabulates the accuracy of each algorithm over seven datasets obtained for the UCI repository, and their average number of iterations to converge to the optima over different folds.

We also consider two big datasets: 1) Adult and 2) MNIST. For MNIST, we considered the classification of the digit 1 with other digits. Since the training and test partitions of these datasets are determined, we do not conduct tenfold cross-validation for these datasets. The RBF parameters for all classifiers are set to 0.05 and 0.02 for *Adult* and *MNIST* classifier, respectively. Except for LSVM, Xia RNN, and the proposed neural network, other neural solutions failed to produce acceptable results in a reasonable time (<24 h). The neural network and LSVM have similar results in terms of the accuracy, but the proposed neural network is more time efficient with respect to LSVM.

It is plain to grasp that the proposed neural network significantly outperforms other neural solutions from both the accuracy and the average number of iteration views. The result of the neural network is also competitive with LSVM from both perspectives. This table illustrates that the GVSVM bears a reasonable result in real-world scenarios, and the proposed neural network is an efficient solver for it.

VI. DISCUSSION

The GVSVM has shown acceptable performance in the classification. In comparison to the standard SVM, it has a fraction of the square bias term in the objective function of the primal minimization. However, this small change can significantly impact on the procedure for solving the optimization problem. Using the GVSVM needs the adjusting a fixed parameter t, which needs to be large enough to guarantee the closeness of the GVSVM to the standard SVM. We recommend setting t as the largest number a machine can address. The consequence of using the GVSVM is that it can directly compute the bias term after solving the minimization, and the corresponding neural network has simple architecture and is timewise efficient due to the fewer operations in each iteration and the exponential convergence of the neural model.

VII. Conclusion

This paper introduced the GVSVM and elaborated the equivalence of its solution to the standard SVM. The difference between the GVSVM and the standard SVM is that the GVSVM has the term $(1/2t)b^2$ in its objective function, where t is a positive scalar. In the GVSVM, the bias term is directly obtained and is suitable when large datasets are available. As the GVSVM is different from the standard SVM, there is no guarantee that its solution is equivalent to the standard SVM. This paper illustrated that as $t \to \infty$, the optimal solution of GVSVM tends to the optimal solution of the standard SVM. The GVSVM solution implies a closed-from formula for the bias term of the standard SVM which obviates the need of an approximation for it. We further proposed an efficient neural network to solve the GVSVM dual problem. It is demonstrated that the neural network is asymptotically stable and is globally exponentially convergent to the solution of the GVSVM. The experimental results illustrated the efficacy of the proposed neural network and confirmed that separating hyperplane found by the GVSVM with a larger t is analogous to the separating hyperplane of the standard SVM.

References

[1] Y.-T. Hu, Y.-Y. Lin, H.-Y. Chen, K.-J. Hsu, and B.-Y. Chen, "Matching images with multiple descriptors: An unsupervised approach for locally adaptive descriptor selection," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5995–6010, Dec. 2015.

[2] S. Zhang, S. Zhao, Y. Sui, and L. Zhang, "Single object tracking with fuzzy least squares support vector machine," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5723–5738, Dec. 2015.

[3] L. Zhang, L. Wang, and W. Lin, "Semisupervised biased maximum margin analysis for interactive image retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2294–2308, Apr. 2012.

[4] X. Li, X. Jia, L. Wang, and K. Zhao, "On spectral unmixing resolution using extended support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 4985–4996, Sep. 2015.

[5] P. Insom *et al.*, "A support vector machine-based particle filter method for improved flooding classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 12, no. 9, pp. 1943–1947, Sep. 2015.

[6] D. Kleftogiannis, K. Theofilatos, S. Likothanassis, and S. Mavroudi, "Yamipred: A novel evolutionary method for predicting pre-mirnas and selecting relevant features," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 5, pp. 1183–1192, Sep./Oct. 2015.

[7] T. Mehmood, J. Bohlin, and L. Snipen, "A partial least squares based procedure for upstream sequence classification in prokaryotes," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 3, pp. 560–567, May/Jun. 2015.

[8] F. Chu and L. Wang, "Applications of support vector machines to cancer classification with microarray data," *Int. J. Neural Syst.*, vol. 15, no. 6, pp. 475–484, 2005.

[9] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule extraction from support vector machines using ensemble learning approach: An application for diagnosis of diabetes," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 2, pp. 728–734, Mar. 2015.

[10] B.-W. Chen, C.-Y. Chen, and J.-F. Wang, "Smart homecare surveillance system: Behavior identification based on state-transition support vector machines and sound directivity pattern analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 6, pp. 1279–1289, Nov. 2013.

[11] L. Wang, B. Liu, and C. Wan, "Classification using support vector machines with graded resolution," in *Proc. IEEE Int. Conf. Granular Comput.*, vol. 2. Beijing, China, 2005, pp. 666–670.

[12] O. Chapelle, "Training a support vector machine in the primal," *Neural Comput.*, vol. 19, no. 5, pp. 1155–1178, May 2007.

[13] S. Melacci and M. Belkin, "Laplacian support vector machines trained in the primal," *J. Mach. Learn. Res.*, vol. 12, pp. 1149–1184, Feb. 2011.

[14] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for SVM," *Math. Program.*, vol. 127, no. 1, pp. 3–30, 2011.

[15] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[16] E. Hazan, T. Koren, and N. Srebro, "Beating SGD: Learning SVMS in sublinear time," in *Proc. Adv. Neural Inf. Process.*, 2011, pp. 1233–1241.

[17] J. Wang, D. Yang, W. Jiang, and J. Zhou, "Semisupervised incremental support vector machine learning based on neighborhood kernel estimation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 10, pp. 2677–2687, Oct. 2017.

[18] B. Fan, X. Lu, and H.-X. Li, "Probabilistic inference-based least squares support vector machine for modeling under noisy environment," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 12, pp. 1703–1710, Dec. 2016.

[19] O. L. Mangasarian and D. R. Musicant, "Successive overrelaxation for support vector machines," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1032–1037, Sep. 1999.

[20] O. L. Mangasarian and D. R. Musicant, "Lagrangian support vector machines," *J. Mach. Learn. Res.*, vol. 1, pp. 161–177, Mar. 2001.

[21] G. Fung and O. L. Mangasarian, "Proximal support vector machine classifiers," in *Proc. KDD Knowl. Disc. Data Min.*, 2001, pp. 77–86.

[22] Y.-J. Lee and O. L. Mangasarian, "RSVM: Reduced support vector machines," in *Proc. SDM*, vol. 1, 2001, pp. 325–361.

[23] O. L. Mangasarian and D. R. Musicant, "Active support vector machine classification," in *Proc. NIPS*, 2000, pp. 577–583.

[24] G. Fung and O. L. Mangasarian, "Finite Newton method for Lagrangian support vector machine classification," *Neurocomputing*, vol. 55, nos. 1–2, pp. 39–55, 2003.

[25] P. Zhong, M. Li, K. Mu, J. Wen, and Y. Xue, "Image steganalysis in high-dimensional feature spaces with proximal support vector machine," *Int. J. Digit. Crime Forensics*, vol. 11, no. 1, pp. 78–89, 2019.

[26] K. Wang, H. Pei, X. Ding, and P. Zhong, "Robust proximal support vector regression based on maximum correntropy criterion," *Sci. Program.*, vol. 2019, Jan. 2019, Art. no. 7102946.

[27] Z. Qi, Y. Tian, and Y. Shi, "Successive overrelaxation for Laplacian support vector machine," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 674–683, Apr. 2015.

[28] Y. Xia, G. Feng, and J. Wang, "A recurrent neural network with exponential convergence for solving convex quadratic program and related linear piecewise equations," *Neural Netw.*, vol. 17, no. 7, pp. 1003–1015, 2004.

[29] M. Mohammadi and A. Mansoori, "A projection neural network for identifying copy number variants," *IEEE J. Biomed. Health Inform.*, to be published.

[30] M. Mohammadi, Y.-H. Tan, W. Hofman, and S. H. Mousavi, "A novel one-layer recurrent neural network for the l_1-regularized least square problem," *Neurocomputing*, vol. 315, pp. 135–144, Nov. 2018.

[31] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming, Theory and Algorithms*. Hoboken, NJ, USA: Wiley Intersci., 2006.

[32] C. J. C. Burges and D. J. Crisp, "Uniqueness of the SVM solution," in *Proc. NIPS*, vol. 12, 2000, pp. 223–229.

[33] S. Willard, *General Topology* (Dover Books on Mathematics). Newburyport, MA, USA: Dover, 2012.

[34] W. Rudin, *Principles of Mathematical Analysis*. New York, NY, USA: McGraw-Hill, 1976.

[35] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, vol. 31. New York, NY, USA: SIAM, 1980.

[36] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, vol. 99. New York, NY, USA: Springer, 2013.

[37] Y. Xia, H. Leung, and J. Wang, "A projection neural network and its application to constrained optimization problems," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 49, no. 4, pp. 447–458, Apr. 2002.

[38] R. A. Dory, "Ordinary differential equations," *Comput. Phys.*, vol. 3, no. 5, pp. 88–91, 1989.

[39] D. Anguita and A. Boni, "Improved neural network for SVM learning," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1243–1244, Sep. 2002.

[40] Y. Tan, Y. Xia, and J. Wang, "Neural network realization of support vector methods for pattern classification," in *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw. (IJCNN)*, vol. 6, 2000, pp. 411–416.

[41] A. Nazemi and M. Dehghan, "A neural network method for solving support vector classification problems," *Neurocomputing*, vol. 152, pp. 369–376, Mar. 2015.

[42] Y. Xia and J. Wang, "A one-layer recurrent neural network for support vector machine learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 2, pp. 1261–1269, Apr. 2004.
[43] Y. Yang, Q. He, and X. Hu, "A compact neural network for training support vector machines," *Neurocomputing*, vol. 86, pp. 193–198, Jun. 2012.
[44] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, U.K.: MIT Press, 2001.
[45] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 131–159, 2002.

Majid Mohammadi received the B.Sc. degree in software engineering and the M.Sc. degree in artificial intelligence from the Ferdowsi University of Mashhad, Mashhad, Iran, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Information and Communication Technology Group, Department of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands, focusing on the use of machine learning and Bayesian statistics for enable interoperability in logistics.

His current research interests include semantic interoperability, machine learning, information theoretic learning, and neurodynamic optimization.

S. Hamid Mousavi was born in Mashhad, Iran, in 1988. He received the B.Sc. degree in pure mathematics and the M.Sc. degree in applied mathematics (focused on control and optimization problems) from the Ferdowsi University of Mashhad (FUM), Mashhad, in 2011 and 2015, respectively. He is currently pursuing the doctoral degree with the Carl von Ossietzky University of Oldenburg, Oldenburg, Germany, focusing on the probabilistic reasoning and sparse coding with applications to voice recognition.

In 2015, he joined the Machine Learning Group, Carl von Ossietzky University of Oldenburg. His current research interests include optimization and probabilistic algorithms.

Sohrab Effati received the B.S. degree in applied mathematical from Birjand University, Birjand, Iran, in 1992, the M.S. degree in applied mathematics from the Tarbiat Moallem University of Tehran, Tehran, Iran, in 1995, and the Ph.D. degree in control systems from the Ferdowsi University of Mashhad, Mashhad, Iran, in 2000.

Since 2005, he has been an Associate Professor with the Department of Applied Mathematics, Ferdowsi University of Mashhad. His current research interests include control systems, optimization, ordinary differential equation and partial differential equations, and neural networks and their applications in optimization problems.