Applied Soft Computing 127 (2022) 109390

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A kernel-free fuzzy reduced quadratic surface ν-support vector machine with applications

Zheming Gao [a], Yiwen Wang [a], Min Huang [a,*], Jian Luo [b], Shanshan Tang [a]

[a] State Key Laboratory of Synthetical Automation for Process Industries, College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110819, China
[b] School of Management, Hainan University, Haikou, Hainan 570228, China

ARTICLE INFO

Article history:
Received 15 January 2022
Received in revised form 29 June 2022
Accepted 19 July 2022
Available online 26 July 2022

Keywords:
Data mining
ν-SVM
Kernel-free SVM
Fuzzy SVM
Binary classification

ABSTRACT

The kernel-free support vector machine (SVM) models are recently developed and studied to overcome some drawbacks induced by the kernel-based SVM models. To further improve the classification accuracy and computational efficiency of existing kernel-free quadratic surface support vector machine (QSSVM) models, a novel kernel-free ν-fuzzy reduced QSSVM model is proposed. The proposed model utilizes a reduced quadratic surface for nonlinear binary classification as well as reducing the effect of outliers in the data set. Some theoretical properties are rigorously studied, especially, the effects of the parameter ν on the dual feasibility and the number of support vectors. Computational experiments are conducted on some public benchmark data sets to indicate the superior performance of the proposed model over some well-known binary classification models. The numerical results also favors the higher training efficiency of the proposed model over those of other kernel-free SVM models. Moreover, the proposed model is successfully applied to the prodromal detection of Alzheimer's Disease with good performance, by using the data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

1. Introduction

Data classification is an essential research direction of machine learning and has significantly influenced many fields. As one of the most powerful classification technique, the SVM model was originally proposed by Cortes and Vapnik [1], and has been well developed and applied to many real-world problems [2,3]. For a binary classification task, the ordinary linear SVM model separates the two classes of data with a hyperplane, while the between-class margin is maximized and the mis-classified data points are penalized. By employing kernel functions, the kernel-based SVM is able to generate nonlinear separation surfaces for certain data sets. The idea is to map the data into a higher dimensional feature space and then classify the mapped data points linearly. Scholköpf et al. [4] proposed the ν-SVM, in which the parameter ν controls the number of support vectors. Even though the ν-SVM produces the same classification results as the ordinary SVM model does under some specific conditions [4–6], tuning the parameter is more convenient when using ν-SVM as the parameter ν is bounded [5]. The ν-SVM can also be equipped with kernels to handle nonlinear cases [7].

Recently, the kernel-free SVM models were proposed [8], which separate the two classes of data by directly utilizing nonlinear surfaces instead of mapping data into high dimensional feature spaces. No pre-selection of kernels is required and no kernel parameters needs to be tuned, which saves much computational time and efforts in the training process. As the most typical kernel-free SVM model, the soft QSSVM (SQSSVM) [9] directly generates a quadratic separation surface for classification while maximizing the relative geometric margin between classes and minimizing the mis-classification errors. Various kernel-free QSSVM models were proposed in literature afterwards. Bai et al. [10] proposed a least squares based kernel-free QSSVM and applied it to disease classification. Tian et al. [11] proposed a fuzzy QSSVM for reject inference in credit scoring. Mousavi et al. [12] investigated the sparsity of SQSSVM by adding an ℓ_1 regularization to the objective function.

However, kernel-free QSSVM models may have disadvantages as well. First, the number of variables in SQSSVM increases exponentially after the equivalent reformulation [9,12,13], which slows down the training efficiency of SQSSVM. Besides, the best penalty parameter of the SQSSVM model is relatively large [9,14], which brings the difficulty when tuning parameters with the grid method. To overcome these drawbacks, in this paper, we propose a state-of-the-art fuzzy reduced quadratic surface support vector machine model for nonlinear binary classification, which is denoted as (ν-FRQSSVM). Certain theoretical properties

[*] Corresponding author.
E-mail address: mhuang@mail.neu.edu.cn (M. Huang).

https://doi.org/10.1016/j.asoc.2022.109390
1568-4946/© 2022 Elsevier B.V. All rights reserved.

are rigorously studied for the proposed model. Computational experiments are also conducted to investigate the classification accuracy, the training efficiency and the parameter sensibility of the proposed model.

The SVM models have been successfully applied to many real-life problems, one of which is the disease diagnosis. As an essential field of machine learning aided disease diagnosis, the prodromal detection of the Alzheimer's Disease (AD) [15–17] attracts much attention. Since the cause of AD has not been fully understood by human beings, the clinical diagnosis of AD can only be made at a very late stage after confirming the clinic hallmarks of AD [18]. Hence, an early detection of potential AD patients is helpful for doctors to provide them with early treatments and save unnecessary medical costs. The data in disease diagnosis is often small-scaled with many features [19], so it can be relatively well-handled by SVM models among the machine learning methods. Recently, some kernel-free QSSVM models [10] have also been applied to disease diagnosis. However, due to the large number of features in the MRI-based AD data sets, the kernel-free QSSVM models proposed in literature might not be suitable for the detection of AD. Some state-of-the-art machine learning models [20,21] are also proposed and applied to the AD diagnosis. In this paper, by using the AD data sets obtained from ADNI database, the proposed model is shown to successfully handle the prodromal detection of AD.

The main contributions of this paper are summarized as the following:

1. The proposed (ν-FRQSSVM) model generates the separation quadratic surface without considering the cross terms in the quadratic form so that it is much more efficient than other tested kernel-free QSSVM models in terms of computational time. In addition, the proposed model is equipped with fuzzy membership, which is applied to each training data point and helps reduce the relative contribution of outliers or noise in generating the optimal separation surface. The employment of the ν-SVM idea further expedite the training process of the proposed model.
2. The proposed (ν-FRQSSVM) model is theoretically investigated. The dual formulation is derived and the relationship between its feasibility and the range of parameter ν is rigorously shown. We also show how the number of support vectors is controlled by the value of the parameter ν.
3. To the best of our knowledge, this is the first work of applying a kernel-free SVM to Alzheimer's Disease forecasting. There are hundreds of features in the data of the Alzheimer's Disease forecasting problem, which limits the applicability of well-known kernel-free QSSVM models but not impact the applicability of the proposed model. From the computational results, the good performance of the proposed model for Alzheimer's Disease forecasting, indicates the potential of the kernel-free proposed model in well handling real-world problems where the data has the large number of features.

The rest of this paper is organized as follows. We first bring some related research works of SVM models for binary classification in Section 2. Then the proposed (ν-FRQSSVM) model is introduced in Section 3. Some theoretical properties are also analyzed. In Section 4, computational experiments are conducted on some public benchmark data sets and the AD data from ADNI database. Section 5 concludes the paper.

2. Preliminaries

Some preliminary knowledge, including a brief review of some related SVM models for binary classification, is introduced in this section.

2.1. Notations

Some mathematical notations are introduced in this section. Throughout this paper, we use lower case letters to denote scalars, bold lower case letters to denote vectors, and bold upper case letters to denote matrices. The n-dimensional Euclidean space is denoted by $\mathbb{R}^n$, and its non-negative orthant is denoted by $\mathbb{R}^n_+$. Denote the set of n-dimensional symmetric matrices by $\mathbb{S}^n$, and denote the set of n-dimensional diagonal matrices by $\mathbb{D}^n$. For any matrix $\mathbf{B} \in \mathbb{S}^n$, write $\mathbf{B} \succeq 0$ if it is positive semi-definite, and $\mathbf{B} \succ 0$ if it is positive definite. For any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, denote its ith row by a row vector $\mathbf{A}_{i\bullet}$, and its jth column by a column vector $\mathbf{A}_{\bullet j}$.

For further convenience, we introduce a vectorization technique for symmetric matrices, which is widely used in the reformulations of kernel-free SVM models [9,14]. For any $\mathbf{B} \in \mathbb{S}^n$, define the half-vectorization of $\mathbf{B}$ as the following:

$$\mathrm{hvec}(\mathbf{B}) \triangleq [B_{11}, \dots, B_{1n}, B_{22}, \dots, B_{2n}, \dots, B_{n-1,n-1}, B_{n-1,n}, B_{nn}]^T \in \mathbb{R}^{n(n+1)/2}. \tag{1}$$

If $\mathbf{B} \in \mathbb{D}^n$, define the diag-vectorization as the following:

$$\mathrm{dvec}(\mathbf{B}) \triangleq [B_{11}, B_{22}, \dots, B_{n-1,n-1}, B_{nn}]^T \in \mathbb{R}^n. \tag{2}$$

In addition, for any vector $\boldsymbol{a} = [a_1, \dots, a_n]^T \in \mathbb{R}^n$, define the quadratic vector qvec($\boldsymbol{a}$) as

$$\mathrm{qvec}(\boldsymbol{a}) \triangleq \left[\frac{1}{2}a_1^2, \frac{1}{2}a_2^2, \dots, \frac{1}{2}a_n^2\right]^T \in \mathbb{R}^n \tag{3}$$

and the quadratic vector lvec($\boldsymbol{a}$) with cross terms as

$$\mathrm{lvec}(\boldsymbol{a}) \triangleq \left[\frac{1}{2}a_1^2, a_1a_2, \dots, a_1a_n, \frac{1}{2}a_2^2, a_2a_3, \dots, a_2a_n, \dots, \frac{1}{2}a_n^2\right]^T \in \mathbb{R}^{n(n+1)/2} \tag{4}$$

Remark. With the definitions above, it is possible to linearize the quadratic forms of a symmetric or a diagonal matrix. In order words, for any $\boldsymbol{x} \in \mathbb{R}^n$, $\mathbf{B} \in \mathbb{S}^n$ and $\mathbf{D} \in \mathbb{D}^n$.

$$\frac{1}{2}\boldsymbol{x}^T\mathbf{B}\boldsymbol{x} = \mathrm{lvec}(\boldsymbol{x})^T\mathrm{hvec}(\mathbf{B}), \qquad \frac{1}{2}\boldsymbol{x}^T\mathbf{D}\boldsymbol{x} = \mathrm{qvec}(\boldsymbol{x})^T\mathrm{dvec}(\mathbf{D}).$$

Throughout this article, define $\mathcal{Q} : \mathbb{S}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ with

$$\mathcal{Q}(\mathbf{W}, \boldsymbol{b}, c) \triangleq \{\boldsymbol{x} \in \mathbb{R}^n | \frac{1}{2}\boldsymbol{x}^T\mathbf{W}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c = 0\} \subset \mathbb{R}^n. \tag{5}$$

Notice that, $\mathcal{Q}(\mathbf{W}, \boldsymbol{b}, c)$ represents a quadratic surface parameterized by tuple $(\mathbf{W}, \boldsymbol{b}, c)$ in $\mathbb{R}^n$. When $\mathbf{W}$ is a diagonal matrix, the corresponding quadratic surface is called a *reduced quadratic surface* and is denoted as $\mathcal{Q}_d(\mathbf{W}, \boldsymbol{b}, c)$.

We also define the binary data set as the following:

$$\mathcal{D} = \left\{ \left(\boldsymbol{x}^{(i)}, y^{(i)}\right)_{i=1,\dots,N} \;\middle|\; \boldsymbol{x}^{(i)} \in \mathbb{R}^n,\ y^{(i)} \in \{-1, 1\} \right\}, \tag{6}$$

where N is the number of data points. Each data point $\boldsymbol{x}^{(i)} = [x_1^{(i)}, \dots, x_n^{(i)}]^T \in \mathbb{R}^n$ is a vector of n feature values, and $y^{(i)}$ is the label of $\boldsymbol{x}^{(i)}$. We bring the definition of a trivial feature in data set $\mathcal{D}$ as the following:

Definition 2.1 (*Trivial Feature*)**.** Given data $\mathcal{D}$ as defined by (6), the kth feature is trivial if $x_k^{(1)} = x_k^{(2)} = \cdots = x_k^{(N)}$.

For any binary data set $\mathcal{D}$ used throughout this paper, we make the following assumption without loss of generality:

Assumption 1. All the features of the data set $\mathcal{D}$ are non-trivial.

Intuitively speaking, a feature with unique value does not contribute to the classification result, and we can always drop it when preprocessing the data set.

We also define the index sets of the positive and the negative class as $I_+ \triangleq \{i|y^{(i)} = +1\}$ and $I_- \triangleq \{i|y^{(i)} = -1\}$, respectively. Moreover, denote the cardinalities of I_+ and I_- by N_+ and N_-, respectively. Denote the index set of the smaller class by $I_{\min}$, i.e.,

$$\begin{cases} I_+, & N_+ \leqslant N_-, \\ I_-, & \text{otherwise.} \end{cases} \tag{7}$$

Let $N_{\min} \triangleq |I_{\min}|$ and notice that $N_{\min} = \min(N_+, N_-)$. Without loss of generality, we assume $N_{\min} > 0$ for this article.

2.2. Related binary classification SVM models

In this subsection, we provide a brief review of some related SVM models. For a binary classification task, the idea of SVM is to find a separation hyperplane while maximizing the margin of separation [1]. When the data is not linearly separable, the soft-margin is employed by introducing the slack vector $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^T \in \mathbb{R}^N_+$. The soft-margin linear SVM is formulated as the following:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geqslant 1 - \xi_i \quad \forall i = 1, \ldots, N \\ & \boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N_+. \end{aligned} \tag{SVM}$$

where $C > 0$ is a given parameter. Each ξ_i is a slack variable that measures the mis-classification error for each $\boldsymbol{x}^{(i)}$. For those cases that the linear hyperplanes cannot characterize the nonlinear structure of the data sets, the kernel-based SVM models were proposed. The idea is to map the data points into a higher dimensional feature space using via a nonlinear feature map $\phi : \mathbb{R}^n \to \mathbb{R}^l$ $(l > n)$, and then separate the mapped data with a hyperplane in the feature space. The kernel-based SVM is formulated as the following:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\boldsymbol{v}\|_2^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y^{(i)}\left(\boldsymbol{v}^T\phi(\boldsymbol{x}^{(i)}) + d\right) \geqslant 1 - \xi_i \quad \forall i = 1, \ldots, N \\ & \boldsymbol{v} \in \mathbb{R}^l, d \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N_+. \end{aligned} \tag{KSVM}$$

where $C > 0$ is the given parameter. Denote $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \phi(\boldsymbol{x}^{(i)})^T\phi(\boldsymbol{x}^{(j)})$ as the kernel function of $\boldsymbol{x}^{(j)}$ and $\boldsymbol{x}^{(j)}$. Various kernel functions have been proposed in literature, including the frequently used RBF (radial basis function) kernel and the quadratic (second order polynomial) kernel [12]. Notice that (KSVM) reduces to (SVM) when $\phi(\boldsymbol{x}) = \boldsymbol{x}$.

$$\begin{aligned} &\text{(RBF kernel)} \quad K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \exp\left(\frac{-\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_2^2}{2\gamma^2}\right) \\ &\text{(quadratic kernel)} \quad K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \left(\boldsymbol{x}^{(i)T}\boldsymbol{x}^{(j)} + r\right)^2 \end{aligned} \tag{8}$$

Since (KSVM) is a convex quadratic programming (QP) problem, so it is meaningful to study its dual problem:

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})y^{(i)}y^{(j)}\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N}\alpha_i y^{(i)} = 0 \\ & 0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, N. \end{aligned} \tag{DKSVM}$$

where $C > 0$ is a given parameter.

Similar to the SVM models, Schölkopf et al. proposed the ν-SVM model for binary classification in [4]. The kernel-based ν-SVM model is formulated as the following:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\boldsymbol{v}\|_2^2 - \nu\rho + \frac{1}{N}\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y^{(i)}\left(\boldsymbol{v}^T\phi(\boldsymbol{x}^{(i)}) + d\right) \geqslant \rho - \xi_i \quad \forall i = 1, \ldots, N \\ & \boldsymbol{v} \in \mathbb{R}^l, d \in \mathbb{R}, \rho \in \mathbb{R}_+, \boldsymbol{\xi} \in \mathbb{R}^N_+. \end{aligned} \tag{ν-KSVM}$$

The only parameter in (ν-KSVM) is $\nu \in [0, 1]$, which yields a more efficient process for tuning the parameter. Moreover, parameter ν provides a more effective control of the number of support vectors [4]. From the view of optimization, the (ν-KSVM) is still a convex QP problem and its dual problem can be formulated as the following:

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})y^{(i)}y^{(j)}\alpha_i\alpha_j \\ \text{s.t.} \quad & \sum_{i=1}^{N}\alpha_i y^{(i)} = 0 \\ & \sum_{i=1}^{N}\alpha_i \geqslant \nu \\ & 0 \leqslant \alpha_i \leqslant \frac{1}{N}, \quad i = 1, \ldots, N. \end{aligned} \tag{ν-DKSVM}$$

where ν is the given parameter. Some theoretical properties have been investigated in [4–6], including the following lemma:

Lemma 2.1. *(ν-DKSVM) is feasible if and only if $\nu \leqslant \nu^*$, where*

$$\nu^* \triangleq \frac{2N_{\min}}{N}$$

where $N_{\min}$ is defined at the beginning of Section 2.2.

Remark. Notice that $\nu^* \leqslant 1$. The upper bound of ν provided by Lemma 2.1 is significant in training (ν-DKSVM) as it reduces the range of ν from $[0, 1]$ to $[0, \nu^*]$. The proof can be found in [5].

Recently, multiple the kernel-free SVM models have been proposed and developed for nonlinear classification [8,9,12,14]. Instead of mapping the data points into a higher-dimensional feature space, the kernel-free SVM models separate the data sets by directly generating the nonlinear separation surfaces in the original space. A typical kernel-free SVM model is the following soft quadratic surface SVM (SQSSVM) model [9], which maximizes the summation of relative geometrical margins of data points while penalizing the margin errors:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N}\|\mathbf{W}\boldsymbol{x}^{(i)} + \boldsymbol{b}\|_2^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y^{(i)}\left(\frac{1}{2}\boldsymbol{x}^{(i)T}\mathbf{W}\boldsymbol{x}^{(i)} + \boldsymbol{x}^{(i)T}\boldsymbol{b} + c\right) \geqslant 1 - \xi_i, \quad i = 1, \ldots, N, \\ & \mathbf{W} \in \mathbb{S}^n, \ \boldsymbol{b} \in \mathbb{R}^n, \ c \in \mathbb{R}, \ \boldsymbol{\xi} \in \mathbb{R}^N_+. \end{aligned} \tag{SQSSVM}$$

where C is a given parameter. The (SQSSVM) model can be equivalently reformulated into the following model for easier implementation:

$$\begin{aligned}\min\quad & \frac{1}{2}\boldsymbol{z}^T\mathbf{G}\boldsymbol{z}+C\sum_{i=1}^{N}\xi_i\\ s.t.\quad & y^{(i)}\left({\boldsymbol{s}^{(i)}}^T\boldsymbol{z}+c\right)\geqslant 1-\xi_i,\quad i=1,\ldots,N,\\ & \boldsymbol{z}\in\mathbb{R}^{n(n+1)/2},\ c\in\mathbb{R},\ \boldsymbol{\xi}\in\mathbb{R}^N_+.\end{aligned}\tag{SQSSVM$'$}$$

where $\mathbf{G}$ and $\boldsymbol{s}^{(i)}(\forall i=1,\ldots N)$ are defined in [9].

And the dual problem of (SQSSVM′) [9] is formulated as the following:

$$\begin{aligned}\min\quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^{(i)}y^{(j)}{\boldsymbol{s}^{(i)}}^T\mathbf{G}^{-1}\boldsymbol{s}^{(i)}-\sum_{i=1}^{N}\alpha_i\\ s.t.\quad & \sum_{i=1}^{N}\alpha_i y^{(i)}=0\\ & 0\leqslant\alpha_i\leqslant C,\ i=1,\ldots,N.\end{aligned}\tag{DSQSSVM$'$}$$

In some real-life applications, the data sets may contain noise or outliers, which may decrease the classification accuracy. To reduce the influence from outliers, Lin and Wang [22] equipped the SVM model with the fuzzy property. Each data point $\boldsymbol{x}^{(i)}$ is assigned a fuzzy membership $\theta_i\in(0,1]$ as a weight to represent its importance for the classification. If $\boldsymbol{x}^{(i)}$ is more like an outlier, θ_i will be close to 0, which will reduce its contribution to the optimal classifier.

Given a data set $\mathcal{D}$ as defined by (6), the fuzzy membership θ_i of $\boldsymbol{x}^{(i)}$ is defined as the following:

$$\theta_i=\begin{cases}1-\dfrac{\|\boldsymbol{x}_+-\boldsymbol{x}^{(i)}\|_2}{r_++\epsilon} & i\in I_+\\[2mm] 1-\dfrac{\|\boldsymbol{x}_--\boldsymbol{x}^{(i)}\|_2}{r_-+\epsilon} & i\in I_-\end{cases}\tag{9}$$

where $\epsilon>0$ is a small perturbation to avoid zero fuzzy membership value. The $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ are the means of the positive and the negative classes, respectively. Also, r_+ and r_-, as defined below, are the radii of the two classes, respectively.

$$r_+=\max_{i\in I_+}\{\|\boldsymbol{x}_+-\boldsymbol{x}^{(i)}\|_2\},\qquad r_-=\max_{i\in I_-}\{\|\boldsymbol{x}_--\boldsymbol{x}^{(i)}\|_2\}.\tag{10}$$

The linear fuzzy SVM (FSVM) [22] is formulated as the following:

$$\begin{aligned}\min\quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2+C\sum_{i=1}^{N}\theta_i\xi_i\\ s.t.\quad & y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b\right)\geqslant 1-\xi_i\quad\forall i=1,\ldots,N\\ & \boldsymbol{w}\in\mathbb{R}^n,\ b\in\mathbb{R},\ \boldsymbol{\xi}\in\mathbb{R}^N_+.\end{aligned}\tag{FSVM}$$

where $C>0$ is the given parameter and $\theta_i\in(0,1]$ is the fuzzy membership associated with data point $\boldsymbol{x}^{(i)}$ $(\forall i=1,\ldots,N)$. Based on the value of θ_i, we may control the importance of the data point $\boldsymbol{x}^{(i)}$ in the case since a bigger value of θ_i yields a relative important data point $\boldsymbol{x}^{(i)}$ for classification. Moreover, the dual problem of (FSVM) is formulated as the following:

$$\begin{aligned}\min\quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}{\boldsymbol{x}^{(i)}}^T\boldsymbol{x}^{(j)}y^{(i)}y^{(j)}\alpha_i\alpha_j-\sum_{i=1}^{N}\alpha_i\\ s.t.\quad & \sum_{i=1}^{N}\alpha_i y^{(i)}=0\\ & 0\leqslant\alpha_i\leqslant\theta_i C,\quad i=1,\ldots,N.\end{aligned}\tag{DFSVM}$$

where C is the given parameter. Notice that the FSVM models can be equipped with kernels for nonlinear classification in a similar way. Various fuzzy-based SVM models were proposed in literature [11,23], which have been applied to solving different types of real-life problems.

3. The ν-fuzzy reduced quadratic surface support vector machine model

In this section, we propose the ν-FRQSSVM model for binary classification. In Section 3.1, we derive the proposed model from the SQSSVM model step by step, and Section 3.2 brings the dual of (ν-FRQSSVM) as well as some theoretical properties of the proposed model.

3.1. Model derivation

We first introduce the following ν-SQSSVM model:

$$\begin{aligned}\min\quad & \sum_{i=1}^{N}\|\mathbf{W}\boldsymbol{x}^{(i)}+\boldsymbol{b}\|_2^2-\nu\rho+\frac{1}{N}\sum_{i=1}^{N}\xi_i\\ s.t.\quad & y^{(i)}\left(\frac{1}{2}{\boldsymbol{x}^{(i)}}^T\mathbf{W}\boldsymbol{x}^{(i)}+{\boldsymbol{x}^{(i)}}^T\boldsymbol{b}+c\right)\geqslant\rho-\xi_i,\quad i=1,\ldots,N,\\ & \mathbf{W}\in\mathbb{S}^n,\ \boldsymbol{b}\in\mathbb{R}^n,\ c\in\mathbb{R},\ \rho\in\mathbb{R}_+,\ \boldsymbol{\xi}\in\mathbb{R}^N_+.\end{aligned}\tag{ν-SQSSVM}$$

where ν is a given parameter. Similarly, (ν-SQSSVM) becomes the following (ν-SQSSVM′) after the equivalent reformulation used when formulating (SQSSVM′):

$$\begin{aligned}\min\quad & \frac{1}{2}\boldsymbol{z}^T\mathbf{G}\boldsymbol{z}-\nu\rho+\frac{1}{N}\sum_{i=1}^{N}\xi_i\\ s.t.\quad & y^{(i)}\left({\boldsymbol{s}^{(i)}}^T\boldsymbol{z}+c\right)\geqslant\rho-\xi_i,\quad i=1,\ldots,N,\\ & \boldsymbol{z}\in\mathbb{R}^{\frac{n(n+1)}{2}+n},\ c\in\mathbb{R},\ \rho\in\mathbb{R}_+,\ \boldsymbol{\xi}\in\mathbb{R}^N_+.\end{aligned}\tag{ν-SQSSVM$'$}$$

where ν is the given parameter.

Since variable $\boldsymbol{z}$ in (ν-SQSSVM′) is the vectorization of matrix variable $\mathbf{W}$ in model (ν-SQSSVM), the number of variables associated with the coefficients of the separation surface is in the order of $\mathcal{O}(n^2)$. Consequently, the computational efficiency slows down quickly as the dimension of the data set increases.

Recall that all the upper triangular elements of $\mathbf{W}$ are recorded in $\boldsymbol{z}$ when formulating (SQSSVM′). The elements on the diagonal of $\mathbf{W}$ are the coefficients of quadratic terms while the rest are the coefficients of the cross terms. In order to taking the advantage of the kernel-free QSSVM model while reducing the computational complexity induced by a large number of features, we only record the elements on the diagonal of $\mathbf{W}$. In other words, the idea is to separate the two classes of data by utilizing a reduced quadratic surface, whose quadratic coefficient matrix is diagonal. To reach this goal, we formulate the following reduced ν-SQSSVM model:

$$\begin{aligned}\min\quad & \sum_{i=1}^{N}\|\boldsymbol{\Sigma}\boldsymbol{x}^{(i)}+\boldsymbol{b}\|_2^2-\nu\rho+\frac{1}{N}\sum_{i=1}^{N}\xi_i\\ s.t.\quad & y^{(i)}\left(\frac{1}{2}{\boldsymbol{x}^{(i)}}^T\boldsymbol{\Sigma}\boldsymbol{x}^{(i)}+{\boldsymbol{x}^{(i)}}^T\boldsymbol{b}+c\right)\geqslant\rho-\xi_i,\quad i=1,\ldots,N,\\ & \boldsymbol{\Sigma}\in\mathbb{D}^n,\ \boldsymbol{b}\in\mathbb{R}^n,\ c\in\mathbb{R},\ \rho\in\mathbb{R}_+,\ \boldsymbol{\xi}\in\mathbb{R}^N_+.\end{aligned}\tag{ν-RSQSSVM}$$

where ν is the given parameter.

(ν-RSQSSVM) can be equivalently reformulated as the following model by using the same technique when formulating

(SQSSVM′). Let $\boldsymbol{v} \triangleq \mathrm{dvec}(\boldsymbol{\Sigma})$, and $\forall i = 1, \dots, N$, $\boldsymbol{r}^{(i)} \triangleq \mathrm{qvec}(\boldsymbol{x}^{(i)})$. The reduced ν-SQSSVM′ can be formulated as the following:

$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{v}^T \bar{\mathbf{G}} \boldsymbol{v} - \nu\rho + \frac{1}{N}\sum_{i=1}^{N} \xi_i \\ s.t. \quad & y^{(i)}\left(\boldsymbol{r}^{(i)^T}\boldsymbol{v} + c\right) \geqslant \rho - \xi_i, \quad i = 1, \dots, N, \\ & \boldsymbol{v} \in \mathbb{R}^{2n}, \;\; c \in \mathbb{R}, \;\; \rho \in \mathbb{R}_+, \;\; \boldsymbol{\xi} \in \mathbb{R}^N_+. \end{aligned} \tag{ν-RSQSSVM′}$$

where ν is the given parameter. Denote the optimal solution to (ν-RSQSSVM) as $(\boldsymbol{\Sigma}^*, \boldsymbol{b}^*, c^*)$, the decision function produced by (ν-RSQSSVM) is

$$f(\boldsymbol{x}) = \mathrm{sign}\left(\frac{1}{2}\boldsymbol{x}^{(i)^T}\boldsymbol{\Sigma}^*\boldsymbol{x}^{(i)} + \boldsymbol{x}^{(i)^T}\boldsymbol{b}^* + c^*\right) \tag{11}$$

Remark. Recall that an orthogonal transformation preserves the inner product, i.e., the length of vectors and the angles between vectors. Therefore, the separability of a given data set $\mathcal{D}$, intuitively, does not change after an orthogonal transformation. In fact, utilizing a quadratic surface $(\mathbf{W}, \boldsymbol{b}, c)$ to classify a data set $\mathcal{D}$ is equivalent to utilizing a reduced quadratic surface $(\boldsymbol{\Sigma}, \hat{\boldsymbol{b}}, c)$ to classify a data set obtained by applying an orthogonal transformation to $\mathcal{D}$. We leave the detailed derivation in Appendix A.

By assigning the fuzzy membership (9) to each data point $\boldsymbol{x}^{(i)}$, we propose the following (ν-FRQSSVM) model:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} \theta_i \|\boldsymbol{\Sigma}\boldsymbol{x}^{(i)} + \boldsymbol{b}\|_2^2 - \nu\rho + \frac{1}{N}\sum_{i=1}^{N} \theta_i\xi_i \\ s.t. \quad & y^{(i)}\left(\frac{1}{2}\boldsymbol{x}^{(i)^T}\boldsymbol{\Sigma}\boldsymbol{x}^{(i)} + \boldsymbol{x}^{(i)^T}\boldsymbol{b} + c\right) \geqslant \rho - \xi_i, \quad i = 1, \dots, N, \\ & \boldsymbol{\Sigma} \in \mathbb{D}^n, \;\; \boldsymbol{b} \in \mathbb{R}^n, \;\; c \in \mathbb{R}, \;\; \rho \in \mathbb{R}_+, \;\; \boldsymbol{\xi} \in \mathbb{R}^N_+. \end{aligned} \tag{ν-FRQSSVM}$$

where ν is the given parameter. Denote the optimal solution to (ν-FRQSSVM) as $(\boldsymbol{\Sigma}^*, \boldsymbol{b}^*, c^*)$, the decision function produced by (ν-FRQSSVM) is also defined by (11).

Remark. The slack variable $\xi^{(i)}$ measures the error of misclassification, and the fuzzy membership θ_i represents a relative in-class importance of data $\boldsymbol{x}^{(i)}$. Recall that the fuzzy membership is embedded in the two terms in the objective function of the (ν-FRQSSVM) model. Each term $\theta_i\xi^{(i)}$ and $\theta_i\|\boldsymbol{\Sigma}\boldsymbol{x}^{(i)} + \boldsymbol{b}\|_2^2$ can be regarded as a measure of $\xi^{(i)}$ and $\|\boldsymbol{\Sigma}\boldsymbol{x}^{(i)} + \boldsymbol{b}\|_2^2$ with a weight θ_i, respectively. When $\boldsymbol{x}^{(i)}$ is likely to be an outlier, it is supposed to be less important for the classification. Hence, its corresponding weight θ_i is expected to be small to reduce the effect of $\xi^{(i)}$ and $\|\boldsymbol{\Sigma}\boldsymbol{x}^{(i)} + \boldsymbol{b}\|_2^2$ on the separation surface produced by the proposed model.

By using the same reformulation technique, the proposed (ν-FRQSSVM) model can be equivalently reformulated as the following:

$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{v}^T \hat{\mathbf{G}} \boldsymbol{v} - \nu\rho + \frac{1}{N}\sum_{i=1}^{N} \theta_i\xi_i \\ s.t. \quad & y^{(i)}\left(\boldsymbol{r}^{(i)^T}\boldsymbol{v} + c\right) \geqslant \rho - \xi_i, \quad i = 1, \dots, N, \\ & \boldsymbol{v} \in \mathbb{R}^{2n}, \;\; c \in \mathbb{R}, \;\; \rho \in \mathbb{R}_+, \;\; \boldsymbol{\xi} \in \mathbb{R}^N_+. \end{aligned} \tag{ν-FRQSSVM′}$$

where ν is the given parameter. The calculation of matrix $\hat{\mathbf{G}}$ can be find in Appendix A.

3.2. The dual and some theoretical properties of ν-FRQSSVM

Lemma 3.1 (*Positive Definiteness of Matrix* $\hat{\mathbf{G}}$)**.** *Given any data set as defined in* (6), *the matrix* $\hat{\mathbf{G}} \succ 0$ *under* Assumption 1.

The proof can be found in Appendix C.

Notice that (ν-FRQSSVM′) is a convex QP problem, and the Lagrangian function is

$$\begin{aligned} L(\boldsymbol{z}, c, \rho, \boldsymbol{\xi}, \boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) = & \frac{1}{2}\boldsymbol{z}^T\hat{\mathbf{G}}\boldsymbol{z} - \nu\rho + \frac{1}{N}\sum_{i=1}^{N}\theta_i\xi_i \\ & + \sum_{i=1}^{N}\alpha_i\left[\rho - \xi_i - y^{(i)}\left(\boldsymbol{z}^T\boldsymbol{r}^{(i)} + c\right)\right] \\ & - \beta\rho - \sum_{i=1}^{N}\gamma_i\xi_i \end{aligned} \tag{12}$$

Take the partial derivative of L on each variable and we have

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{z}} = \hat{\mathbf{G}}\boldsymbol{z} - \displaystyle\sum_{i=1}^{N}\alpha_i y^{(i)}\boldsymbol{r}^{(i)} \\ \dfrac{\partial L}{\partial c} = -\displaystyle\sum_{i=1}^{N}\alpha_i y^{(i)} \\ \dfrac{\partial L}{\partial \rho} = -\nu + \displaystyle\sum_{i=1}^{N}\alpha_i - \beta \\ \dfrac{\partial L}{\partial \xi_i} = \dfrac{\theta_i}{N} - \alpha_i - \gamma_i, \quad \forall i = 1, \dots, N. \end{cases} \tag{13}$$

Let the partial derivatives be zeros and plug back into (ν-FRQSSVM′), the dual problem is obtained as the following:

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{s}^{(i)^T}\hat{\mathbf{G}}^{-1}\boldsymbol{s}^{(i)} \\ s.t. \quad & \sum_{i=1}^{N}\alpha_i y^{(i)} = 0 \\ & \sum_{i=1}^{N}\alpha_i \geqslant \nu \\ & 0 \leqslant \alpha_i \leqslant \frac{\theta_i}{N}, \quad \forall i = 1, \dots, N. \end{aligned} \tag{ν-DFRQSSVM′}$$

The KKT condition is listed below.

$$\begin{cases} \boldsymbol{z} = \hat{\mathbf{G}}^{-1}\displaystyle\sum_{i=1}^{N}\alpha_i y^{(i)}\boldsymbol{r}^{(i)}, \quad \sum_{i=1}^{N}\alpha_i y^{(i)} = 0, \\ \displaystyle\sum_{i=1}^{N}\alpha_i = \beta + \nu, \quad \beta\rho = 0, \quad \rho \geqslant 0, \quad \beta \geqslant 0, \\ y^{(i)}\left(\boldsymbol{z}^T\boldsymbol{r}^{(i)} + c\right) \geqslant \rho - \xi_i, \quad \forall i = 1, \dots, N, \\ \alpha_i\left(\rho - \xi_i - y^{(i)}\left(\boldsymbol{z}^T\boldsymbol{r}^{(i)} + c\right)\right) = 0, \quad \forall i = 1, \dots, N, \\ 0 \leqslant \alpha_i \leqslant \dfrac{\theta_i}{N}, \quad \forall i = 1, \dots, N. \end{cases} \tag{14}$$

where $(\boldsymbol{z}, \boldsymbol{b}, c, \rho, \boldsymbol{\xi})$ is the primal optimal solution, $\boldsymbol{\alpha}$ is the dual optimal solution, and β is the corresponding Lagrangian multiplier of ρ.

The following theorem provides a sufficient and necessary condition for the feasibility of the dual problem (ν-DFRQSSVM′).

Theorem 3.2 (*Feasibility of* (ν-DFRQSSVM′)). *(ν-DFRQSSVM′) is feasible if and only if $\nu \in [0, \nu_{\max}]$, where*

$$\nu_{\max} \triangleq \frac{2\sum_{i\in I_{\min}} \theta_i}{N}. \tag{15}$$

Here θ_i is the fuzzy membership associated with data point $\mathbf{x}^{(i)}$.

Proof. First, we assume that the (ν-DFRQSSVM′) model is feasible. Without loss of generality, let $N_+ \leqslant N_-$ so that $I_{\min} = I_+$. For any dual feasible solution $\boldsymbol{\alpha} \in \mathbb{R}^N$, we have

$$\nu \leqslant \sum_{i=1}^{N} \alpha_i = \sum_{i\in I_+} \alpha_i + \sum_{j\in I_-} \alpha_j.$$

Since $\sum_{i=1}^N y^{(i)}\alpha_i = 0$ yields $\sum_{i\in I_+} \alpha_i = \sum_{j\in I_-} \alpha_j$ and $\alpha_i \leqslant \theta_i/N$, the inequality chain above becomes

$$\nu \leqslant 2\sum_{i\in I_+} \alpha_i \leqslant 2\frac{\sum_{i\in I_+} \theta_i}{N} = \nu_{\max}.$$

Thus, the sufficiency is proved. Next, we show the necessarity. Assume that $\nu \in [0, \nu_{\max}]$, let $\boldsymbol{\alpha} \in \mathbb{R}^N$ such that

$$\alpha_i = \begin{cases} \dfrac{\nu\theta_i}{2\sum_{j\in I_+} \theta_j} & i \in I_+, \\ \dfrac{\nu\theta_i}{2\sum_{j\in I_-} \theta_j} & i \in I_-. \end{cases} \tag{16}$$

Notice that α_i is non-negative. Besides, $\boldsymbol{\alpha}$ satisfies the following constraints:

$$\alpha_i \leqslant \frac{\nu_{\max}\theta_i}{2\sum_{j\in I_{\min}} \theta_j} = \frac{2\theta_i \sum_{j\in I_{\min}} \theta_j}{2N\sum_{j\in I_{\min}} \theta_j} = \frac{\theta_i}{N}, \quad i = 1, \dots, N. \tag{17}$$

$$\sum_{i=1}^{N} \alpha_i = \sum_{i\in I_+} \alpha_i + \sum_{i\in I_-} \alpha_i = \frac{\nu\sum_{i\in I_+} \theta_i}{2\sum_{j\in I_+} \theta_j} + \frac{\nu\sum_{i\in I_-} \theta_i}{2\sum_{j\in I_-} \theta_j} = \frac{\nu}{2} + \frac{\nu}{2} = \nu. \tag{18}$$

Similarly, $\sum_{i=1}^N y^{(i)}\alpha_i = \sum_{i\in I_+} \alpha_i - \sum_{j\in I_-} \alpha_j = \frac{\nu}{2} - \frac{\nu}{2} = 0$ is also satisfied. Hence, $\boldsymbol{\alpha}$ is a feasible solution to the dual problem, which yields the necessarity.

In conclusion, the theorem is proved. □

Theorem 3.2 provides a relatively tight upper bound for the parameter ν in the proposed model. Since (ν-FRQSSVM′) is a convex QP problem with only affine constraints, the strong duality holds [24]. Moreover, the dual problem is bounded for any $\nu \in [0, \nu_{\max}]$. Therefore, $\nu_{\max}$ guarantees an achievable primal optimal solution, which yields the decision function as define by (11).

The relationship between the parameter ν and the dual optimal solution is shown in the following:

Corollary 3.2.1. *Given $\nu \in [0, \nu_{\max}]$, the following properties hold:*

1. *There exists an optimal solution to (ν-DFRQSSVM′) such that $\sum_{i=1}^N \alpha_i = \nu$.*
2. *For any optimal solution to (ν-DFRQSSVM′) such that $\sum_{i=1}^N \alpha_i > \nu$, its corresponding primal optimal solution yields $\rho = 0$.*
3. *If (ν-FRQSSVM′) has an optimal solution such that $\rho > 0$, then $N\nu$ is the lower bound of the number of support vectors.*

Proof. We show the first property by contradiction. Assume that for any optimal solution $\boldsymbol{\alpha}^* = [\alpha_1^*, \dots, \alpha_N^*]^T$, $\sum_{i=1}^N \alpha_i^* > \nu$. Then define $\bar{\boldsymbol{\alpha}}$ such that

$$\bar{\boldsymbol{\alpha}} = \gamma\boldsymbol{\alpha}^*.$$

where $\gamma = \frac{\nu}{\sum_{i=1}^N \alpha_i^*} \in (0, 1)$ and denote the optimal value as d^*. Hence, the following inequality chain is obtained

$$\begin{aligned} d^* &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i^*\alpha_j^* y^{(i)}y^{(j)}\mathbf{s}^{(i)^T}\hat{\mathbf{G}}^{-1}\mathbf{s}^{(i)} \\ &\leqslant \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \bar{\alpha}_i\bar{\alpha}_j y^{(i)}y^{(j)}\mathbf{s}^{(i)^T}\hat{\mathbf{G}}^{-1}\mathbf{s}^{(i)} \\ &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \gamma^2\alpha_i^*\alpha_j^* y^{(i)}y^{(j)}\mathbf{s}^{(i)^T}\hat{\mathbf{G}}^{-1}\mathbf{s}^{(i)} = \gamma^2 d^* < d^*, \end{aligned} \tag{19}$$

which yields a contradiction. Hence, there exists at least one optimal solution such that $\sum_{i=1}^N \alpha_i = \nu$.

Next, we show the second property. Recall the KKT condition discussed in (14), and we have $\beta > 0$ since $\sum_{i=1}^N \alpha_i > \nu$. Thus, $\rho = 0$.

For the third one, recall the KKT condition in (14). A positive ρ yields a zero multiplier β so that the equality $\sum_{i=1}^N \alpha_i = \nu$ holds. Let $\mathcal{A} \triangleq \{\alpha_i > 0 | i = 1, \dots, N\}$ be the set of support vectors and notice that $\alpha_i \leqslant \theta_i/N \leqslant 1/N$ for all $i = 1, \dots, N$. Hence,

$$\frac{|\mathcal{A}|}{N} \geqslant \sum_{i=1}^{N} \frac{\alpha_i}{\theta_i} \geqslant \sum_{i=1}^{N} \alpha_i = \nu.$$

which yields $|\mathcal{A}| \geqslant N\nu$.

In conclusion, the corollary is proved. □

4. Computational experiments

In this section, some computational experiments are conducted to investigate the effectiveness and efficiency of the proposed (ν-FRQSSVM) model. We first introduce the settings of the experiments and then test the proposed model along with some well-known binary classification models on some public benchmark data sets.

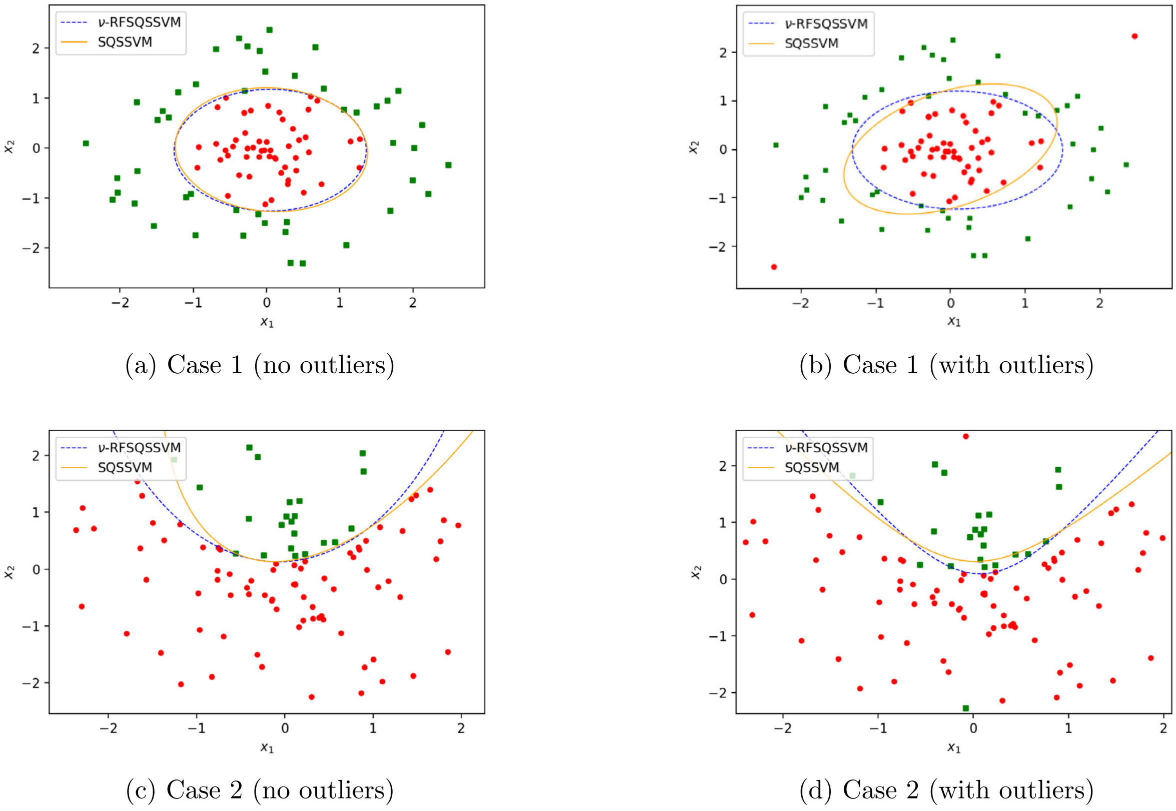
4.1. Experiment settings

The proposed model is implemented along with some benchmark models for comparison, including some well-developed SVM models such as the SVMs with the RBF kernel and the quadratic kernel, the ν-SVM models with the RBF kernel and quadratic kernel. Two kernel-free SVM models are also implemented, including the SQSSVM model [9], the quadratic least squares SVM model [10] and the fuzzy QSSVM model [11]. Moreover, two state-of-the-art network-based models are also implemented for comparison, including the intuitionistic fuzzy random vector functional link network [20] and the stochastic configuration network [25]. In addition, other typical methods for binary classification are tested for comparison, including the logistic regression model, the decision tree model, the Gaussian naive Bayes model and the artificial neural network model. The abbreviation of each tested model is listed in Table 1 below. We also listed the commercial solvers or packages that are utilized to implement some of the models. Notice that, the (ν-FRQSSVM′) is implemented for (ν-FRQSSVM) by utilizing the Cplex solver.

All the computational experiments are conducted on a computer with twelve Intel(R) Core(TM) i7-9750H CPU @ 2.60 GHz CPUs and 16 GB RAM. A ten-fold cross validation procedure is applied for each experiment. And to make the results statistically meaningful, each experiment is repeated ten times for each tested model. The mean and the standard deviation of accuracy scores are recorded to qualify the effectiveness of each model. In addition, the CPU time consumed by each model is also

Table 1
Abbreviations and solvers of tested models.

Model	Abbreviation	Solver/Package	Parameters
Logistic regression	LR	Scikit-learn	–
Decision tree	DT	Scikit-learn	–
Artificial neural network	ANN	Scikit-learn	–
Gaussian naive Bayes	GNB	Scikit-learn	–
Stochastic configuration network	SCN	–	$(L_{\max}, T_{\max})$
Intuitionistic fuzzy random vector functional link network	IFRVFL	–	(C, μ, N)
SVM with RBF kernel	SVM-rbf	LIBSVM	(C, γ)
SVM with quadratic kernel	SVM-quad	LIBSVM	(C, r)
ν-SVM with RBF kernel	ν-SVM-rbf	LIBSVM	(ν, γ)
ν-SVM with quadratic kernel	ν-SVM-quad	LIBSVM	(ν, r)
Soft margin quadratic surface SVM	SQSSVM	Cplex	C
Quadratic least squares SVM	QLSSVM	–	C
Fuzzy quadratic surface SVM	FQSSVM	Cplex	C
The proposed model	ν-FRQSSVM	Cplex	ν



(a) Case 1 (no outliers)

(b) Case 1 (with outliers)

(c) Case 2 (no outliers)

(d) Case 2 (with outliers)

Fig. 1. (ν-FRQSSVM) vs. SQSSVM on artificial data sets with/without outliers.

recorded. It is the time when implementing a model once with fixed parameters.

The parameters for each tested model are tuned by using grid method, which is a common approach in literature [9,14]. The range of each parameter is listed in Appendix D.

We first use two artificial data sets to show the effects that the fuzzy membership brings to the proposed model. Both the (ν-FRQSSVM) model and the SQSSVM model are conducted on the 2-D artificial data sets as plotted in Figs. 1(a)–1(d). Fig. 1(b) shares the same data pattern as that in Figs. 1(a), and 1(d) shares the same data pattern as that in Fig. 1(c). But, both Figs. 1(b) and 1(d) have outliers.

From the classification results plotted in Figs. 1(a) and 1(c), we see that both the SQSSVM model and the proposed (ν-FRQSSVM) model perfectly separate the artificial data sets, respectively. However, for the data sets with outliers as shown in 1(b) and 1(d), the (ν-FRQSSVM) model is still able to characterize a proper separation surface, while the SQSSVM is not able to. It shows the robustness of the proposed (ν-FRQSSVM) model despite the existence of outliers.

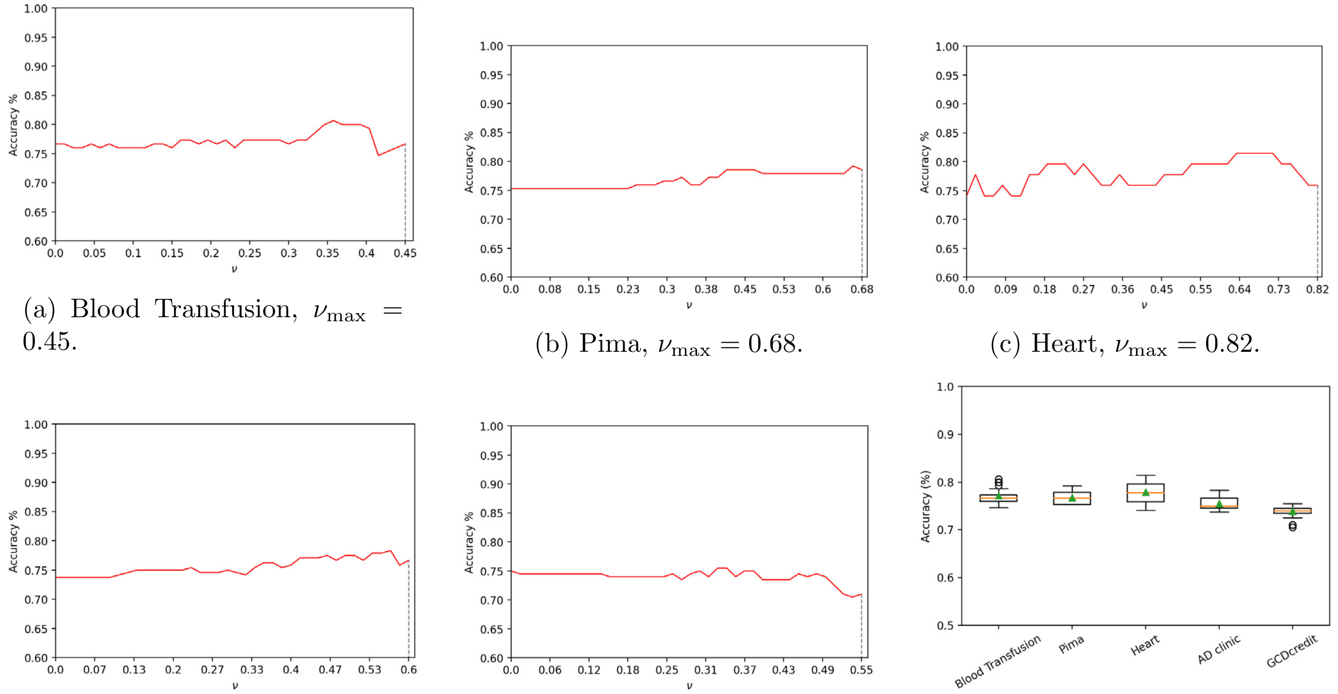
Next, we plot the following figures to show the sensibility of parameter ν on a few commonly used benchmark data sets.

Fig. 2 verifies that the parameter ν is bounded by 0 and $\nu_{\max}$, which is related to the weights θ_i's. In Fig. 2(f), we have also plotted the boxplot for each result shown in Figs. 2(a) to 2(e), respectively. The boxplot of each data set is flat, i.e., the accuracy scores obtained with different values of ν are tightly distributed. Hence, we may conclude that the classification accuracy of the proposed model is not very sensitive to the value of its parameter $\nu \in [0, \nu_{\max}]$.

4.2. Tested on public benchmark data

The proposed (ν-FRQSSVM) model is tested with some public benchmark data sets, whose basic information is listed in Table 2.

We have the following observations from the results listed in Tables 3 and 4:



(a) Blood Transfusion, $\nu_{\max} = 0.45$.

(b) Pima, $\nu_{\max} = 0.68$.

(c) Heart, $\nu_{\max} = 0.82$.

(d) AD clinic, $\nu_{\max} = 0.60$.

(e) GCDcredit, $\nu_{\max} = 0.55$.

(f) Boxplots of the results.

Fig. 2. Accuracy scores vs. ν.

Table 2
Public benchmark data sets.

Data set	n	Sample size (N_1 vs. N_2)	Data set	n	Sample size (N_1 vs. N_2)
Blood transfusion	4	178 vs. 570	Glass	9	51 vs. 163
AD clinic	5	383 vs. 846	Wine	13	59 vs. 71
Cryotherapy	6	48 vs. 42	Heart	13	120 vs. 150
Liver disorders	6	185 vs. 134	AUScredit	14	383 vs. 307
Ecoli	7	143 vs. 193	JAPcredit	15	296 vs. 357
Wholesale	7	142 vs. 298	GCDcredit	24	300 vs. 700
Pima	8	268 vs. 500	Loan	44	1000 vs. 1000
Disease basic	9	37 vs. 127	Sonar	60	97 vs. 110

- Compared with other tested models, the proposed (ν-FRQSSVM) model provides the highest mean accuracy scores on all the tested data sets. For the health-care related data sets such as the *AD clinic*, *Cryotherapy*, *Liver-disorders*, *Disease basic* and *Heart*, the proposed model dominates in terms of accuracy. In other words, it may have the potential in solving binary classification problems in health-care.
- The CPU time consumed by the proposed (ν-FRQSSVM) model is shorter than those by the SQSSVM and the FQSSVM models. Indeed, the proposed model has a smaller number of variables than SQSSVM or FQSSVM does, which leads to a much less worst case computational complexity order. Moreover, even though (ν-FRQSSVM) consumes a bit longer CPU time than the least squares based QLSSVM model does on those small-scaled data sets, it becomes more efficient than QLSSVM when the number of features and data points increases.
- Notice that, except the kernel-free QSSVM models, the SCN model and the IFRVFL model, all the other tested models are conducted by utilizing Scikit-learn [26] and LIBSVM [27] packages, which are professionally supported. Even though the proposed model consumes a longer CPU time, it is still acceptable.

Moreover, some statistical tests [20,28] are conducted to evaluate the all the tested models on the public benchmark data sets. Denote the number of models as p and the number of data sets as m.

First, the average ranks of all the models are calculated in Table 5. Denote the average rank of ith model as r_j. The results show that the classification accuracy of the proposed model is the highest with respect to the average rank among all the tested models.

In addition, we conduct the Friedman tests [28] to assess all the tested models. Recall that the Friedman statistic follows the τ_{χ^2} distribution (when p and m are not small value) with a $p-1$ degrees of freedom, where $\tau_{\chi^2} = \frac{12m}{p(p+1)} \sum_{i=1}^{p} \left(r_i - \frac{p+1}{2}\right)^2$. Also, define $\tau_F = \frac{(m-1)\tau_{\chi^2}}{m(p-1)-\tau_{\chi^2}}$, where τ_F follows an F-distribution with degrees of freedom $p-1$ and $(p-1)(m-1)$. With the results in Table 5, we have $\tau_{\chi^2} = 83.957$ and $\tau_F = 10.153$ for $m = 16$ and $p = 14$. The F-distribution table at the 95% level of significance is 1.771, which is smaller than $\tau_F = 10.153$. Hence, it is confident to reject the null hypothesis. In other words, there is a significant difference among all the tested models.

In order to check the significant difference, the Nemenyi post-hoc test is also conducted. Recall that the critical difference is defined by $CD = q_\alpha \sqrt{\frac{p(p+1)}{6m}}$, where q_α is the critical value of the Tukey distribution [28]. Two models are significantly different if the difference of their average ranks is greater than CD. In

Table 3
The results of tested models on Benchmark data.

Data set	Model																											
	Accuracy score %																											
	LR		DT		ANN		GNB		SCN		IFRVFL		SVM-rbf		SVM-quad		ν-SVM-rbf		ν-SVM-quad		SQSSVM		QLSSVM		FQSSVM		ν-FRQSSVM	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Blood transfusion	77.53	1.73	72.30	2.61	79.79	2.66	75.36	3.17	79.54	2.44	76.30	2.66	78.63	1.77	77.05	0.99	78.90	2.02	76.94	0.75	77.91	1.22	79.03	1.86	77.88	1.10	79.84	2.32
AD clinic	76.57	2.98	75.10	2.59	76.08	2.30	71.63	2.98	75.78	3.01	73.92	2.41	75.72	2.31	75.13	2.04	76.05	2.38	71.93	2.51	76.67	1.93	76.15	2.10	76.69	1.94	76.92	1.81
Cryotherapy	86.59	5.76	89.41	6.58	85.88	9.45	84.94	5.54	87.76	6.43	84.24	7.00	88.00	6.46	82.12	8.05	88.00	5.75	88.94	7.07	91.29	5.91	92.24	5.19	90.35	6.33	95.06	4.71
Liver disorders	68.25	4.91	64.13	5.69	65.81	5.99	56.57	7.68	71.94	4.70	70.79	1.62	70.29	5.23	69.97	5.70	70.48	4.94	69.40	6.44	73.59	3.88	73.78	4.42	73.14	4.12	75.68	4.39
Ecoli	96.72	2.30	92.42	2.59	94.36	2.89	60.68	10.51	96.55	2.08	92.00	6.59	96.67	1.96	95.76	2.40	96.85	1.95	96.30	1.86	97.27	1.69	97.50	1.54	97.09	1.69	97.58	1.42
Wholesale	91.03	2.78	88.18	3.78	87.22	3.90	89.79	3.17	89.89	3.05	90.11	2.76	90.62	2.54	90.80	2.72	90.57	2.32	87.45	2.79	92.37	2.37	90.94	3.71	92.14	2.24	92.74	2.25
Pima	77.36	3.63	70.69	3.15	73.52	4.07	75.53	3.28	76.26	3.56	74.51	3.09	77.05	3.50	75.90	3.51	76.63	3.45	76.58	3.19	77.36	3.09	77.52	3.14	77.39	3.21	78.22	3.53
Disease basic	92.50	4.86	85.88	4.95	90.50	5.14	87.75	6.31	88.25	4.53	84.38	5.23	89.88	4.26	89.25	3.73	90.13	4.75	88.00	3.57	92.00	3.73	91.88	4.13	91.88	4.03	95.75	2.84
Glass	93.43	3.66	94.19	3.83	93.62	3.42	90.76	2.96	91.33	3.49	87.90	3.86	94.00	2.84	92.48	3.48	93.71	3.50	92.76	3.98	94.29	3.07	93.33	3.63	94.67	2.41	95.62	2.63
Wine	97.92	3.08	96.80	3.83	98.24	2.60	98.08	3.02	98.40	2.26	94.56	5.52	99.04	1.74	97.12	2.95	98.88	2.95	96.96	3.32	99.36	1.50	99.20	1.60	99.20	1.63	100.00	0.00
Heart	82.96	5.01	74.96	4.99	75.85	5.94	84.00	5.23	83.56	5.03	82.22	5.12	83.19	5.26	78.89	5.45	82.30	5.60	81.33	4.90	82.00	3.80	81.41	3.37	81.48	3.59	86.37	4.80
AUScredit	86.13	2.99	81.75	2.78	80.53	3.03	80.32	3.40	85.64	2.95	85.66	3.57	85.99	2.75	82.98	2.65	86.25	2.88	84.26	3.12	86.86	2.84	87.12	2.78	86.66	2.46	87.94	3.04
JAPcredit	86.34	3.29	81.81	3.52	81.17	3.55	80.80	3.03	86.34	3.21	85.66	3.92	85.88	3.55	84.68	3.26	85.94	3.28	85.82	3.50	87.11	3.43	87.32	3.25	87.14	3.22	87.75	2.71
GCDcredit	76.70	2.48	68.16	2.70	69.20	3.54	72.42	3.54	72.76	2.13	70.76	2.59	75.84	2.11	75.08	2.71	76.34	2.27	74.66	2.94	75.62	2.18	76.42	2.17	75.68	2.21	77.60	2.25
Loan	65.74	1.52	55.71	2.08	58.94	2.69	60.93	5.14	60.81	2.61	64.81	1.97	65.48	1.36	62.05	2.53	65.58	1.75	61.64	3.75	64.18	1.69	64.43	1.98	64.19	1.67	66.65	1.52
Sonar	76.29	4.20	74.05	6.86	84.98	5.25	67.90	4.91	72.78	6.85	78.54	6.01	86.63	4.88	82.54	7.17	85.95	4.94	83.90	7.55	77.46	4.74	79.22	5.35	79.02	5.09	86.83	4.88

Table 4
The CPU time of tested models on Benchmark data.

Data set	Model													
	CPU time (s)													
	LR	DT	ANN	GNB	SCN	IFRVFL	SVM-rbf	SVM-quad	ν-SVM-rbf	ν-SVM-quad	SQSSVM	QLSSVM	FQSSVM	ν-FRQSSVM
Blood Transfusion	0.006	0.003	1.380	0.004	0.244	0.156	0.014	0.003	0.007	0.003	0.223	0.029	0.467	0.129
AD clinic	0.013	0.005	2.435	0.004	0.220	0.872	0.015	0.017	0.017	0.009	0.582	0.046	0.505	0.398
Cryotherapy	0.005	0.002	0.154	0.005	0.128	0.031	<0.001	<0.001	<0.001	<0.001	0.037	0.007	0.074	0.023
Liver Disorders	0.005	0.003	1.397	0.004	0.169	0.101	0.002	0.004	0.002	0.001	0.093	0.016	0.188	0.057
Ecoli	0.008	0.002	0.500	0.004	0.145	0.048	0.001	0.001	0.001	0.001	0.108	0.056	0.230	0.064
Wholesale	0.006	0.003	1.134	0.005	0.187	0.081	0.001	0.001	0.001	0.001	0.140	0.038	0.299	0.084
Pima	0.006	0.006	2.698	0.003	0.187	0.174	0.007	0.019	0.007	0.006	0.323	0.053	0.668	0.166
Disease basic	0.011	0.003	0.121	0.003	0.121	0.044	<0.001	<0.001	<0.001	<0.001	0.095	0.022	0.103	0.050
Glass	0.010	0.003	0.103	0.003	0.129	0.061	<0.001	<0.001	<0.001	0.001	0.107	0.023	0.223	0.048
Wine	0.006	0.003	0.082	0.004	0.125	0.036	<0.001	<0.001	<0.001	0.001	0.220	0.025	0.431	0.037
Heart	0.005	0.003	0.159	0.003	0.136	0.086	0.001	0.001	0.001	0.001	0.286	0.047	0.587	0.064
AUScredit	0.007	0.005	0.806	0.004	0.174	0.124	0.007	0.026	0.008	0.008	0.679	0.151	1.345	0.188
JAPcredit	0.011	0.005	0.714	0.004	0.142	0.209	0.006	0.015	0.007	0.005	0.741	0.147	1.507	0.177
GCDcredit	0.009	0.009	0.628	0.004	0.143	0.601	0.021	0.019	0.024	0.020	3.447	0.676	3.557	0.451
Loan	0.034	0.041	2.161	0.006	0.211	2.686	0.132	0.175	0.153	0.206	39.381	17.822	43.742	2.078
Sonar	0.009	0.007	0.296	0.005	0.120	0.191	0.002	0.002	0.002	0.002	40.423	9.994	80.740	0.174

Table 5
Average rank of all the tested models on public benchmark data sets.

Model	Average rank	Model	Average rank
LR	11.063	SVM-quad	6.500
DT	10.688	ν-SVM-rbf	8.313
ANN	10.438	ν-SVM-quad	7.875
GNB	9.500	SQSSVM	6.188
SCN	7.625	QLSSVM	4.438
IFRVFL	7.625	FSQSSVM	6.438
SVM-rbf	7.313	ν-FRQSSVM	1.000

our case, $q_{.05} = 3.354$ and $CD = 4.961$. Hence, the Nemenyi test detects the significant differences between the proposed ν-FRQSSVM model and all the other tested models except the QLSSVM model. Nevertheless, it is not hard to notice that the average rank of the proposed model is much better than that of the QLSSVM model.

Remark. With the results in Tables 3 and 4, and the results from the statistical tests, we are confident to conclude that the proposed ν-FRQSSVM model is strongly competitive among all the tested models. It is meaningful to apply the proposed model to real-life applications.

4.3. Application to the prodromal detection of Alzheimer's disease

In this section, the proposed (ν-FRQSSVM) model is applied to the prodromal detection of AD. First, we briefly introduce the background of AD and the data set.

4.3.1. Background and data

AD is a type of brain disorder which progressively destroys patients' memories, thinking skills and, eventually, the ability to carry out some simplest tasks in daily life. Named after Dr. Alois Alzheimer, who identified this unusual disease of the cerebral cortex in 1906, AD have been attracting much attention from governments and medical researchers [29]. However, the causes of this common dementia is still not fully understand by human beings. In fact, a definite diagnosis, most of times, can only be made after confirming the hallmarks of AD, such as the neurofibrillary tangles or amyloid plaques.

Mild cognitive impairment (MCI) is a prodromal stage of AD. It has been studied that MCI tends to progress to AD at a rate of approximately 10%–15% each year [30]. Hence, an accurate determination of specific markers of early AD progression is crucial in preventing AD from worsening [31]. It not only aids the doctors to develop proper treatments for potential patients, but also save the cost and time of clinical trials.

Recently, machine learning methods have been proposed to automatically classify patients with AD, MCI and the control subjects (CN) based on volumetric measurements of regions of interest (ROI) [15,20,21]. In this subsection, the proposed (ν-FRQSSVM) model is extended and applied to AD forecasting, as well as all other tested models. The AD data set utilized in this study is obtained from the ADNI database, which is a project launched at the beginning of this millennium for investigating the progression of AD. The data set contains a total of 816 data points, including 228 for the CN class, 189 for the AD class, and 399 for the MCI class. The data set has 331 features, including the age, marriage status, gender, education level and 327 ROI-based features. To avoid the dominance of some input features with larger numerical values over those with smaller values, all the data points are normalized into [0,1].

Table 6
The accuracy score results of tested models on AD data.

Model	Classes					
	Accuracy score %					
	CN vs. AD		CN vs. MCI		AD vs. MCI	
	Mean	Std	Mean	Std	Mean	Std
LR	87.46	3.73	73.36	2.52	65.41	3.17
DT	77.95	4.64	73.78	3.04	64.65	4.10
ANN	88.10	3.24	74.05	2.35	66.03	3.47
GNB	81.41	4.73	68.72	2.85	67.14	3.90
SCN	77.22	4.84	70.42	3.00	63.79	3.31
IFRVFL	85.90	3.59	73.41	3.92	64.62	11.48
SVM-rbf	89.51	2.86	77.28	3.01	68.93	3.70
SVM-quad	88.63	2.63	76.96	2.69	70.34	3.75
ν-SVM-rbf	89.61	3.01	77.16	2.61	70.86	3.41
ν-SVM-quad	89.07	2.74	77.16	2.43	70.69	3.24
ν-FRQSSVM	91.17	2.51	77.63	2.21	70.90	2.96

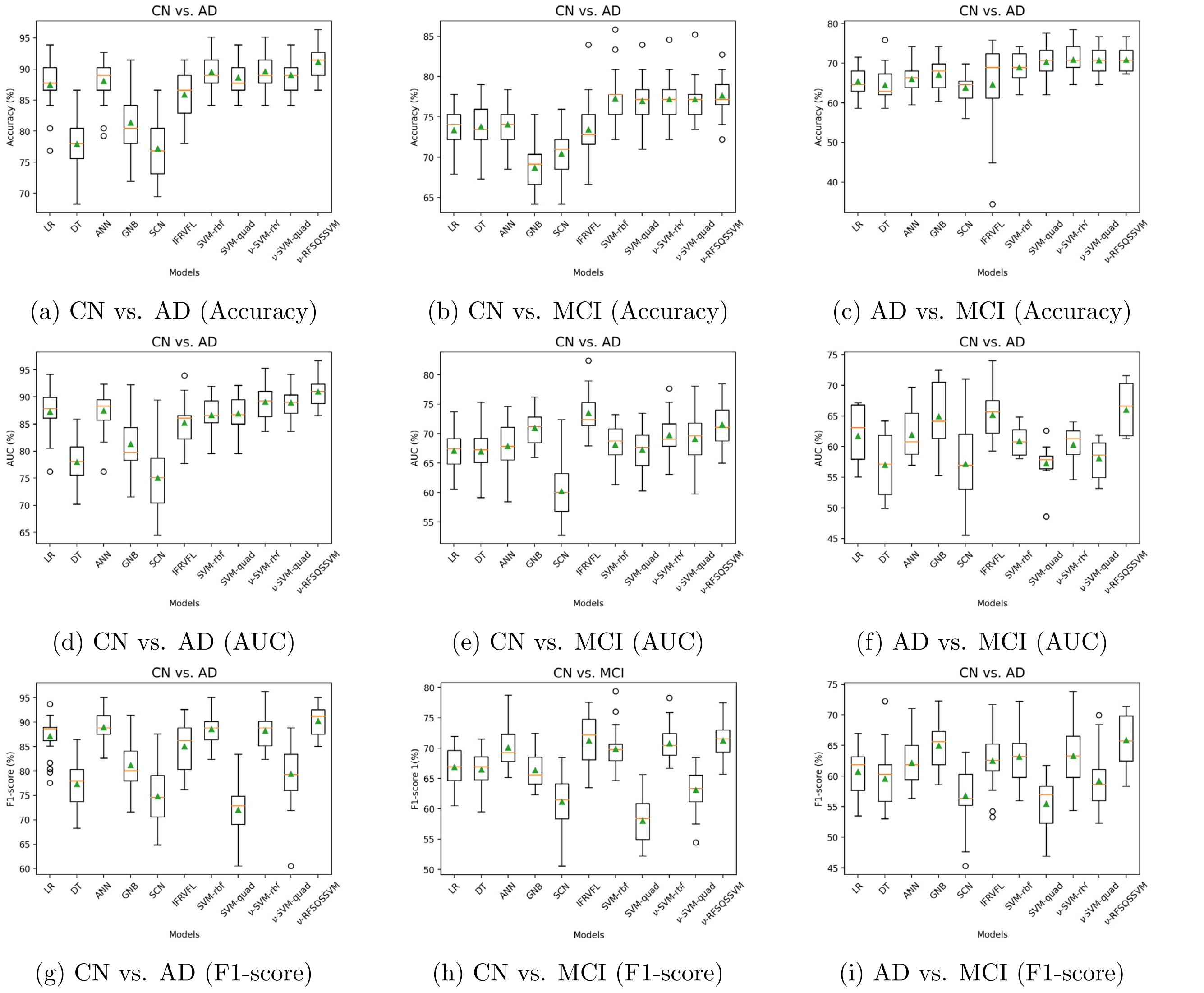
4.3.2. Numerical experiments on the AD data set

The proposed (ν-FRQSSVM) model has been applied to the AD data set as well as all the other tested models except SQSSVM due to the computer memory issue. Indeed, the number of features is $n = 331$ which yields $n(n+1)/2 = 54\,946$ variables in the SQSSVM model, which not amenable for the computer we used throughout the computational experiments. However, the proposed (ν-FRQSSVM) model does not encounter the computational issue. All the results are listed in the following Tables. In Appendix E, we also plotted the ROC curves for all the tested models on the AD data set to show the performance with different training rates (see Tables 8 and 9).

The boxplots of the results are shown below (see Fig. 3):

We have the following observations from the results listed above:

- From the mean accuracy scores listed in Table 6, it might be easier to detect an AD patient from the control subjects



(a) CN vs. AD (Accuracy)

(b) CN vs. MCI (Accuracy)

(c) AD vs. MCI (Accuracy)

(d) CN vs. AD (AUC)

(e) CN vs. MCI (AUC)

(f) AD vs. MCI (AUC)

(g) CN vs. AD (F1-score)

(h) CN vs. MCI (F1-score)

(i) AD vs. MCI (F1-score)

Fig. 3. Boxplots of AD data results.

Table 7
The AUC results of tested models on AD data.

Model	Classes					
	AUC %					
	CN vs. AD		CN vs. MCI		AD vs. MCI	
	Mean	Std	Mean	Std	Mean	Std
LR	87.31	3.81	67.14	3.70	61.47	4.72
DT	77.97	4.66	66.98	3.72	56.90	5.16
ANN	87.61	3.55	67.87	3.82	62.27	4.34
GNB	81.31	4.87	70.98	3.04	64.92	5.56
SCN	75.07	5.91	60.24	4.55	57.17	6.20
IFRVFL	85.26	3.53	73.54	3.53	65.16	3.77
SVM-rbf	86.63	3.30	68.16	3.44	60.79	2.53
SVM-quad	86.91	3.37	67.31	3.47	57.20	3.58
ν-SVM-rbf	89.07	2.97	69.74	3.42	60.50	2.76
ν-SVM-quad	88.98	2.80	69.14	4.88	57.93	2.89
ν-FRQSSVM	90.95	2.69	71.54	3.24	65.94	4.28

Table 8
The F1-score results of tested models on AD data.

Model	Classes					
	F1-score %					
	CN vs. AD		CN vs. MCI		AD vs. MCI	
	Mean	Std	Mean	Std	Mean	Std
LR	87.15	4.18	66.90	3.01	60.70	3.80
DT	77.30	4.50	66.46	2.99	59.58	4.58
ANN	88.97	3.22	70.09	3.49	62.19	3.62
GNB	81.23	4.79	66.40	2.77	64.96	3.99
SCN	74.78	5.40	60.91	4.42	56.77	4.81
IFRVFL	85.05	4.54	70.84	4.14	62.48	3.99
SVM-rbf	88.61	3.57	69.87	2.90	63.17	4.23
SVM-quad	72.08	5.54	58.01	3.49	55.50	4.34
ν-SVM-rbf	88.26	3.56	70.80	2.75	63.32	4.71
ν-SVM-quad	79.38	6.12	63.13	3.49	59.20	4.40
ν-FRQSSVM	90.26	3.17	71.26	2.89	65.91	4.04

than to detect a MCI patient from the control subjects. But it is relatively difficult for any of the tested models to separate the AD and the MCI patients.

- The proposed (ν-FRQSSVM) outperforms all the other tested models in terms of classification accuracy. Notice that, for the binary classification of between CN and AD, only the mean accuracy score provided by (ν-FRQSSVM) exceeds 90%. Moreover, the standard deviation of accuracy scores provided by the proposed model is smallest among those of tested models. All in all, the (ν-FRQSSVM) model provides

Table 9
The CPU time of tested models on AD data.

Model	Classes		
	CPU time (s)		
	CN vs. AD	CN vs. MCI	AD vs. MCI
LR	0.021	0.022	0.022
DT	0.162	0.168	0.166
ANN	4.993	7.412	6.337
GNB	0.007	0.009	0.008
SCN	0.126	0.170	0.172
IFRVFL	2.708	2.582	2.435
SVM-rbf	0.139	0.138	0.141
SVM-quad	0.122	0.141	0.125
ν-SVM-rbf	0.144	0.149	0.148
ν-SVM-quad	0.135	0.140	0.138
ν-FRQSSVM	3.417	3.471	3.517

the most accurate and the most stable classifier among all tested models, for the prodromal detection of AD.

- As shown in Table 7, the mean AUC of the proposed model is the highest among those of all the other tested models. In addition, the standard deviations yielded by (ν-FRQSSVM) are the smallest when classifying the AD patients and the MCI patients against the control subjects. It favors the good classification ability of the proposed model in the prodromal detection of AD.
- Notice that, the typical kernel-free QSSVM models (e.g., SQSSVM, QLSSVM or FQSSVM) are not able to handle the AD data set under the computational environment in this paper, due to the large number of features. However, the proposed model is able to handle this problem, and outperforms all other tested models in terms of accuracy. The CPU time consumed by the proposed model is acceptable, even though it is longer than some of the tested models which are conducted by using commonly-used solvers. Indeed, taking the advantage of utilizing the reduced quadratic surface for separation, the proposed (ν-FRQSSVM) may handle the real-life problems more efficiently than other kernel-free QSSVM models.

We notice that for the binary classification of CN vs. AD, the mean accuracy score of (ν-FRQSSVM) even exceeds 90%, and the standard deviation is also the smallest. For the prodromal detection of AD, i.e., CN vs. MCI, the proposed (ν-FRQSSVM) model is still the most accurate one among all tested models. In addition, the CPU time consumed by the proposed model is acceptable, even though it is longer than those models which are conducted by using packages. Notice that the typical kernel-free SQSSVM model is even not able to handle such a large number of features in the AD data set. In general, the classification accuracy listed in Table 6 has verified that the proposed (ν-FRQSSVM) model may be preferable in machine learning based AD diagnosis.

5. Conclusion

In this paper, we have proposed a state-of-the-art kernel-free (ν-FRQSSVM) model for nonlinear binary classification. Certain theoretical properties of (ν-FRQSSVM) have been rigorously studied. Moreover, the classification effectiveness and efficiency of the proposed model have been investigated by conducting numerical experiments on some public benchmark data sets. In addition, the proposed model has been applied to AD disease diagnosis with a ROI-based MRI data set from ADNI database. The major findings of this paper are summarized below:

- The proposed (ν-FRQSSVM) performs better than all other tested well-known models in terms of the classification accuracy. Different from the nonlinear SVM models equipped with kernels, the proposed (ν-FRQSSVM) model neither requires any kernels nor tuning the kernel parameters, which saves considerable efforts in practice.
- By adopting the idea of ν-SVM, the value of parameter ν in the proposed model is bounded. Therefore, for a given data set, the exact range of parameter ν not only saves efforts, but also yield a better parameter ν while using the grid tuning method. By assigning the fuzzy membership to each data point, the proposed (ν-FRQSSVM) model is able to reduce the influence from outliers in the given data set. The fuzzy membership is easy to calculate and it does not affect the convexity of the model.
- The results from the numerical experiments have shown the dominant performance of the proposed (ν-FRQSSVM) model on the tested benchmark data sets. By generating the quadratic surface without considering the cross terms in the quadratic form, the proposed (ν-FRQSSVM) model consumed much less CPU time than the SQSSVM model did. All these results has indicated the potential of the proposed (ν-FRQSSVM) model in handling other real-life applications of binary classification.
- In particular, the proposed model has shown its promising effectiveness and its acceptable efficiency after being applied to the AD data set. The satisfying performance of the proposed model in the prodromal detection of AD has provided another reliable machine learning tool in the research field of AD forecasting.

Future study

Our investigation of the (ν-FRQSSVM) model leads to some potential research works. First, the reduced quadratic surface adopts only diagonal elements of the matrix in the quadratic term. Similar idea can be employed by other quadratic surface SVM models [32]. Besides, the fuzzy membership function can be customized based on different real-life applications, such as the transportation forecasting [33], credit scoring [34]. Another interesting work is to extend the proposed model to image classification [35,36], which involves matrix data inputs. Moreover, we plan to design an efficient algorithm for the proposed model.

CRediT authorship contribution statement

Zheming Gao: Conceptualization, Methodology, Software, Writing – original draft. **Yiwen Wang:** Validation, Software, Writing – review & editing. **Min Huang:** Resources, Writing – review & editing. **Jian Luo:** Conceptualization, Writing – review & editing. **Shanshan Tang:** Data curation.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (NSFC) Major International (Regional) Joint Research Project Grant No. 71620107003; the National Key R&D Program of China Grant # 2021YFB3300900; the Liaoning Revitalizing Talent Program # XLYC1802115; the Fundamental Research

Funds for State Key Laboratory of Synthetical Automation for Process Industries Grant # 2013ZCX11; the 111 Incubating Program of Overseas Expert Introduction Grant # BC2018010 and the "High-level Overseas Expert" Introduction Program Grant # G20190006026. This research has also been supported by the NSFC, China Grant # 71701035. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI Author ship list.pdf).

Appendix A. Motivation of the proposed (ν-FRQSSVM) model

In this section, we derive the idea of keeping the diagonal elements of the quadratic term of the separation function. Assume the given data set defined by (6) is quadratically separable, i.e., there exists a quadratic surface $\mathcal{Q} \triangleq \{\boldsymbol{x} \in \mathbb{R}^n | \frac{1}{2}\boldsymbol{x}^T\mathbf{W}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c = 0\}$ such that

$$\begin{cases} \frac{1}{2}{\boldsymbol{x}^{(i)}}^T\mathbf{W}\boldsymbol{x}^{(i)} + \boldsymbol{b}^T\boldsymbol{x}^{(i)} + c > 0, & \text{if} \quad y^{(i)} = 1, \\ \frac{1}{2}{\boldsymbol{x}^{(i)}}^T\mathbf{W}\boldsymbol{x}^{(i)} + \boldsymbol{b}^T\boldsymbol{x}^{(i)} + c < 0, & \text{if} \quad y^{(i)} = -1. \end{cases} \tag{20}$$

where matrix $\mathbf{W}$ is symmetric. Recall that for any symmetric matrix $\mathbf{W}$, the singular value decomposition (SVD) of $\mathbf{W}$ tells that there exists an orthonormal matrix $\mathbf{U}$ such that

$$\mathbf{W} = \mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U}, \tag{21}$$

where $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ is a diagonal matrix with $\sigma_1 \geqslant \cdots \geqslant \sigma_n \geqslant 0$ [37]. Hence, $\forall i = 1, \dots, N$,

$$\begin{aligned} &\frac{1}{2}{\boldsymbol{x}^{(i)}}^T\mathbf{W}\boldsymbol{x}^{(i)} + \boldsymbol{b}^T\boldsymbol{x}^{(i)} + c \\ =&\frac{1}{2}{\boldsymbol{x}^{(i)}}^T\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U}\boldsymbol{x}^{(i)} + \boldsymbol{b}^T\mathbf{U}^T\mathbf{U}\boldsymbol{x}^{(i)} + c \\ =&\frac{1}{2}\left(\mathbf{U}\boldsymbol{x}^{(i)}\right)^T\boldsymbol{\Sigma}\left(\mathbf{U}\boldsymbol{x}^{(i)}\right) + (\mathbf{U}\boldsymbol{b})^T\left(\mathbf{U}\boldsymbol{x}^{(i)}\right) + c. \end{aligned} \tag{22}$$

Since $\mathbf{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_n]$ is orthonormal, the columns of $\mathbf{U}$ forms an orthonormal basis of $\mathbb{R}^n$, i.e., $\mathbb{R}^n = \operatorname{span}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_n\}$. More importantly, the linear transformation induced by $\mathbf{U}$ preserves the inner product. In other words, it preserves the length of the vectors and the angle between the vectors. Let the data set $\mathcal{D}$ is linearly transformed by matrix $\mathbf{U}$ and denote the data set as $\mathcal{D}_U$ after transformation:

$$\mathcal{D}_U = \left\{\left(\mathbf{U}\boldsymbol{x}^{(i)}, y^{(i)}\right) \,\middle|\, (\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}, \; i = 1, \dots, N\right\}, \tag{23}$$

Moreover, define quadratic surface $\mathcal{Q}_U \triangleq \{\boldsymbol{z} \in \mathbb{R}^n | \frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}\boldsymbol{z} + (\mathbf{U}\boldsymbol{b})^T\boldsymbol{z} + c = 0\}$. Hence, when $\mathcal{D}$ is quadratically separated by $\mathcal{Q}$, it is equivalent to state that $\mathcal{D}_U$ is quadratically separated by $\mathcal{Q}_U$. In other words, for any quadratically separable data set $\mathcal{D}$, there exists an orthonormal matrix $\mathbf{U}$ such that the data set can be separated by $\mathcal{Q}_U$ after the orthogonal transformation induced by $\mathbf{U}$.

Since the orthogonal transformation preserves length and the angle between the vectors, the shape of a quadratic surface is not changed by a orthogonal transformation. Hence, an orthogonal transformation does not change the separability of a given data set, which yields the motivation of recording only the diagonal elements of the matrix in the quadratic term of the separation surface.

Appendix B. Reformulation from (ν-FRQSSVM) to (ν-FRQSSVM′)

Recall the (SQSSVM) model introduced in (2). For any $i = 1, \dots, N$, let $\boldsymbol{s}^{(i)} = \begin{bmatrix} \operatorname{lvec}(\boldsymbol{x}^{(i)}) \\ \boldsymbol{x}^{(i)} \end{bmatrix}$, $\boldsymbol{z} = \begin{bmatrix} \operatorname{hvec}(\mathbf{W}) \\ \boldsymbol{b} \end{bmatrix}$, where hvec and lvec are defined by (1) and (4), respectively. Hence, the quadratic constraint becomes a linear one as the following:

$$y^{(i)}\left({\boldsymbol{s}^{(i)}}^T\boldsymbol{z} + c\right) \geqslant 1 - \xi_i. \tag{24}$$

The first term in the objective function of the (SQSSVM) model is minimized for maximizing the relative geometrical margin of the separation quadratic surface [9]. For any $i = 1, \dots, N$, a matrix $\boldsymbol{M}^{(i)} \in \mathbb{R}^{n \times \frac{n(n+1)}{2}}$ is constructed. For the kth row of $\mathbf{M}^{(i)}$ ($k = 1, \dots n$), i.e. $\mathbf{M}^{(i)}_{k\bullet}$, assign

$$M^{(i)}_{kj} = \begin{cases} x^{(i)}_p & \text{if} \quad \operatorname{hvec}(\mathbf{W})_j = W_{kp}, \\ 0 & \text{otherwise}. \end{cases} \tag{25}$$

where $\operatorname{hvec}(\mathbf{W})_j$ denotes the jth element of $\operatorname{hvec}(\mathbf{W})$. Then define matrix $\mathbf{H}^{(i)} = [\mathbf{M}^{(i)}, \mathbf{I}_n]$ and let matrix $\mathbf{G} = 2\sum_{i=1}^{N} {\mathbf{H}^{(i)}}^T\mathbf{H}^{(i)}$. It has been shown in [9] that $\sum_{i=1}^{N} \|\mathbf{W}\boldsymbol{x}^{(i)} + \boldsymbol{b}\|_2^2 = \frac{1}{2}\boldsymbol{z}^T\mathbf{G}\boldsymbol{z}$. Thus, (SQSSVM) can be reformulated as (SQSSVM′).

For the (ν-FRQSSVM) model formulated as the following, the idea is similar.

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} \theta_i\|\boldsymbol{\Sigma}\boldsymbol{x}^{(i)} + \boldsymbol{b}\|_2^2 - \nu\rho + \frac{1}{N}\sum_{i=1}^{N} \theta_i\xi_i \\ s.t. \quad & y^{(i)}\left(\frac{1}{2}{\boldsymbol{x}^{(i)}}^T\boldsymbol{\Sigma}\boldsymbol{x}^{(i)} + {\boldsymbol{x}^{(i)}}^T\boldsymbol{b} + c\right) \geqslant \rho - \xi_i, \quad i = 1, \dots, N, \\ & \boldsymbol{\Sigma} \in \mathbb{D}^n, \; \boldsymbol{b} \in \mathbb{R}^n, \; c \in \mathbb{R}, \; \rho \in \mathbb{R}_+, \; \boldsymbol{\xi} \in \mathbb{R}^N_+. \end{aligned} \tag{ν-FRQSSVM}$$

where ν is a given parameter.

For any $i = 1, \dots, N$, let $\boldsymbol{r}^{(i)} = \begin{bmatrix} \operatorname{qvec}(\boldsymbol{x}^{(i)}) \\ \boldsymbol{x}^{(i)} \end{bmatrix}$, $\boldsymbol{v} = \begin{bmatrix} \operatorname{dvec}(\boldsymbol{\Sigma}) \\ \boldsymbol{b} \end{bmatrix}$, where dvec and qvec are defined by (2) and (3), respectively. Hence, the quadratic constraint becomes a linear one as the following:

$$y^{(i)}\left({\boldsymbol{r}^{(i)}}^T\boldsymbol{v} + c\right) \geqslant \rho - \xi_i. \tag{26}$$

Define matrix $\mathbf{T}^{(i)} \triangleq \left[\operatorname{diag}(\boldsymbol{x}^{(i)}) \quad \mathbf{I}_n\right]$, and define matrix $\hat{\mathbf{G}} \triangleq 2\sum_{i=1}^{N} \theta_i{\mathbf{T}^{(i)}}^T\mathbf{T}^{(i)}$. One can verify that $\sum_{i=1}^{N} \|\boldsymbol{\Sigma}\boldsymbol{x}^{(i)} + \boldsymbol{b}\|_2^2 = \frac{1}{2}\boldsymbol{v}^T\hat{\mathbf{G}}\boldsymbol{v}$. Thus, the ($\nu$-FRQSSVM) can be equivalently reformulated as model (ν-FRQSSVM′).

Recall that in the inverse of matrix $\hat{\mathbf{G}}$ is required in the dual formulation (ν-DFRQSSVM′). The positiveness of matrix $\hat{\mathbf{G}}$ is shown in next subsection.

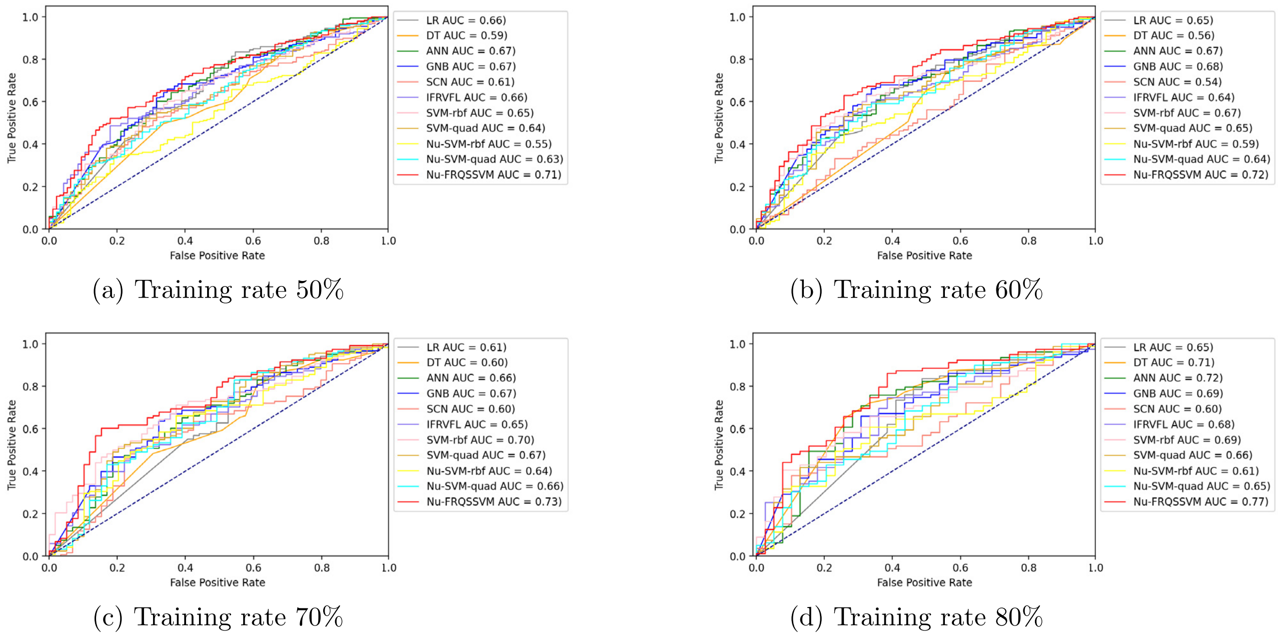
Appendix C. Proof of Lemma 3.1

Proof. The definition of $\hat{\mathbf{G}}$ yields the following representation after conducting the matrix product and the summation.

$$\hat{\mathbf{G}} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{C} \end{bmatrix} \tag{27}$$

where blocks $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are all diagonal matrices and

$$\mathbf{A} = \operatorname{diag}\left(\sum_{i=1}^{N} \theta_i {x^{(i)}_k}^2\right), \quad \mathbf{B} = \operatorname{diag}\left(\sum_{i=1}^{N} \theta_i x^{(i)}_k\right), \quad \mathbf{C} = \sum_{i=1}^{N} \theta_i\mathbf{I}_n. \tag{28}$$



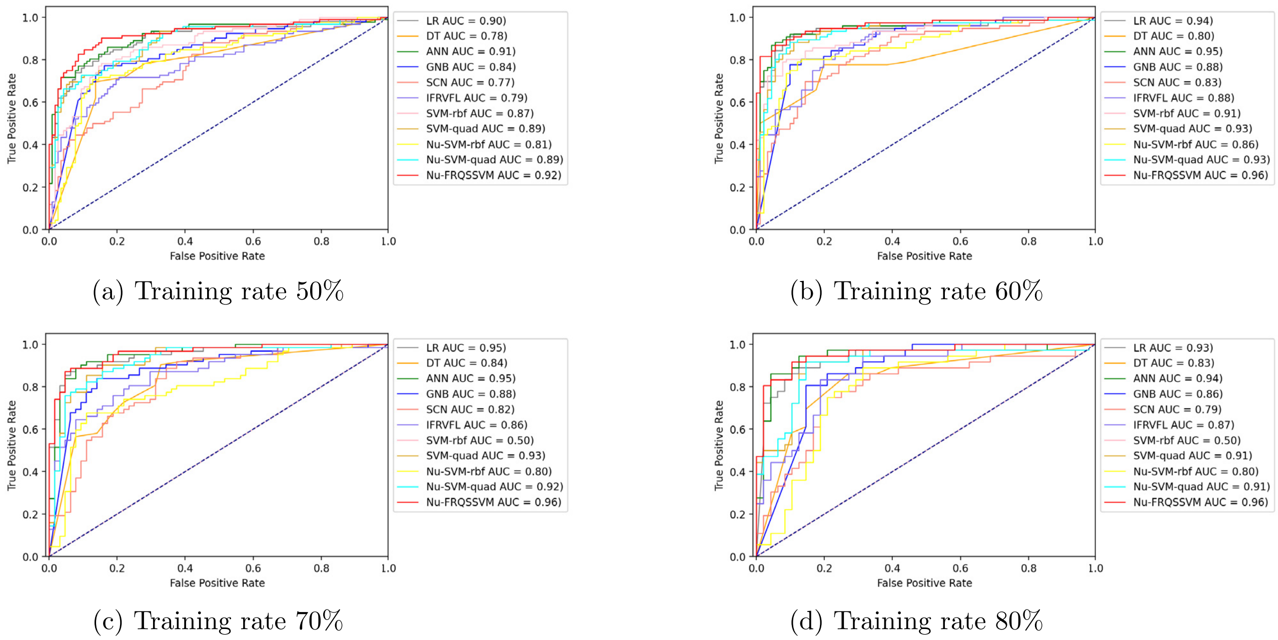
(a) Training rate 50%

(b) Training rate 60%

(c) Training rate 70%

(d) Training rate 80%

Fig. 4. ROC curves on the AD data (AD vs. MCI) with different training rates.



(a) Training rate 50%

(b) Training rate 60%

(c) Training rate 70%

(d) Training rate 80%

Fig. 5. ROC curves on the AD data (CN vs. AD) with different training rates.

Since the fuzzy membership $\theta_i > 0$, matrix $\mathbf{C}$ is non-singular. By Shur's complement [37], $\hat{\mathbf{G}}$ is invertible if and only if $\mathbf{A}-\mathbf{B}\mathbf{C}^{-1}\mathbf{B}$ is invertible. Notice that $\mathbf{A}-\mathbf{B}\mathbf{C}^{-1}\mathbf{B}$ is also a diagonal matrix, which is denoted as $\operatorname{diag}(d_k)$. Hence,

$$d_k = \sum_{i=1}^{N} \theta_i x_k^{(i)2} - \frac{\left(\sum_{i=1}^{N} \theta_i x_k^{(i)}\right)^2}{\sum_{i=1}^{N} \theta_i} \tag{29}$$

Let $\boldsymbol{a}_k = [\sqrt{\theta_1}, \ldots, \sqrt{\theta_N}]^T$ and $\boldsymbol{b}_k = [\sqrt{\theta_1} x_k^{(1)}, \ldots, \sqrt{\theta_N} x_k^{(N)}]^T$. By Cauchy–Schwartz's inequality,

$$\boldsymbol{a}_k^T \boldsymbol{b}_k \leqslant \|\boldsymbol{a}_k\| \|\boldsymbol{b}_k\|. \tag{30}$$

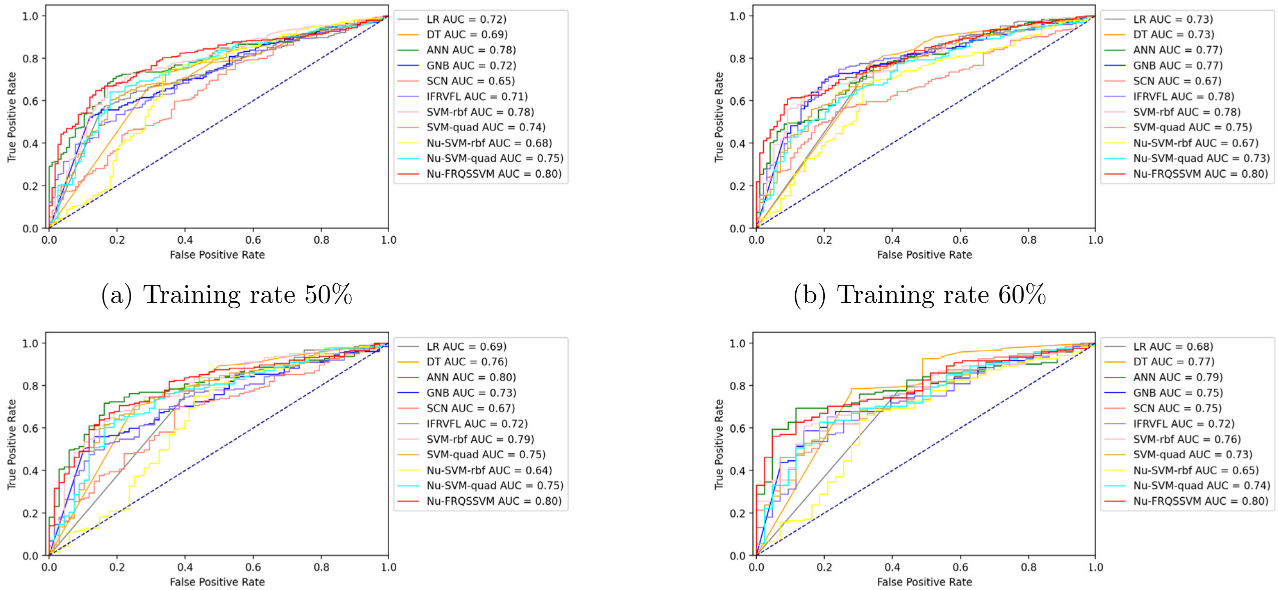
where the equation holds if and only if $\frac{(a_k)_j}{(b_k)_j} = \tau$ is a constant for all$j = 1, \ldots, N$.

Hence, $\sum_{i=1}^{N} \theta_i x_k^{(i)} \leqslant \sqrt{\sum_{i=1}^{N} \theta_i}\sqrt{\sum_{i=1}^{N} \theta_i x_k^{(i)2}}$. In addition, $\frac{(a_k)_j}{(b_k)_j} = \frac{1}{x_k^{(j)}}$ cannot be a constant, otherwise, the kth feature in data set $\mathcal{D}$ will be a trivial and yields a contradiction to Assumption 1.

Therefore, $d_k > 0$ for all $k = 1, \ldots, n$, and the lemma is proved. □

Appendix D. Ranges of parameters

The hyper-parameters in each tested model are listed in Table 1. Each parameter is tuned as the following: $\log_2 C \in \{-8, -3, \ldots, 21, 20\}$, $\log_2 r \in \{-3, -2, \ldots, 2, 3\}$, $\log_2 \gamma \in \{-3, -2, \ldots,$



(a) Training rate 50%

(b) Training rate 60%

(c) Training rate 70%

(d) Training rate 80%

Fig. 6. ROC curves on the AD data (CN vs. MCI) with different training rates.

2, 3}. The maximum number of hidden nodes $L_{\max}$ of the SCN model is set as 10, and the maximum times of random configuration $T_{\max}$ is set as 100.

The ANN model and the IFRVFL is implemented with the ReLU activation function. The ANN model implemented in our experiments has one hidden layer. The size of the hidden layer k is tuned within range $\log_2 \frac{k}{n} \in \{0, \ldots, 3\}$ [14]. The parameters of the IFRVFL model, including the number of hidden neurons N, the kernel parameter μ, and the penalty parameter C are tuned within ranges $N = 3 : 20 : 203$, $\log_2 \mu \in \{-3, -2, \ldots, 2, 3\}$ and $\log_{10} C \in \{-2, \ldots, 4\}$ [20].

For the ν-SVM models, including the proposed model, the range of ν is set as defined in Lemma 2.1 and Theorem 3.2.

Appendix E. Additional experiment results

The ROC curves for all the tested models on the AD data set with different training rates are plotted as follows.

Figs. 4–6 show that the ROC curves of the proposed model stays above all the curves of other tested models with different training rates. It verifies the dominant performance of the proposed model over other tested models, and its potential in the application to AD detection.

References

[1] Corinna Cortes, Vladimir Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[2] Hikmet Esen, Mustafa Inalli, Abdulkadir Sengur, Mehmet Esen, Modeling a ground-coupled heat pump system by a support vector machine, Renew. Energy 33 (8) (2008) 1814–1823.
[3] Hikmet Esen, Filiz Ozgen, Mehmet Esen, Abdulkadir Sengur, Modelling of a new solar air heater through least-squares support vector machines, Expert Syst. Appl. 36 (7) (2009) 10673–10682.
[4] Bernhard Schölkopf, Alex J Smola, Robert C Williamson, Peter L Bartlett, New support vector algorithms, Neural Comput. 12 (5) (2000) 1207–1245.
[5] Chih-Chung Chang, Chih-Jen Lin, Training ν-support vector classifiers: theory and algorithms, Neural Comput. 13 (9) (2001) 2119–2147.
[6] Pai-Hsuen Chen, Chih-Jen Lin, Bernhard Schölkopf, A tutorial on ν-support vector machines, Appl. Stoch. Models Bus. Ind. 21 (2) (2005) 111–136.
[7] Hong-Sen Yan, Duo Xu, An approach to estimating product design time based on fuzzy ν-support vector machine, IEEE Trans. Neural Netw. 18 (3) (2007) 721–731.
[8] Issam Dagher, Quadratic kernel-free non-linear support vector machine, J. Global Optim. 41 (1) (2008) 15–30.
[9] Jian Luo, Shu-Cherng Fang, Zhibin Deng, Xiaoling Guo, Soft quadratic surface support vector machine for binary classification, Asia-Pac. J. Oper. Res. 33 (06) (2016) 1650046.
[10] Yanqin Bai, Xiao Han, Tong Chen, Hua Yu, Quadratic kernel-free least squares support vector machine for target diseases classification, J. Combinator. Optim. 30 (4) (2015) 850–870.
[11] Ye Tian, Ziyang Yong, Jian Luo, A new approach for reject inference in credit scoring using kernel-free fuzzy quadratic surface support vector machines, Appl. Soft Comput. 73 (2018) 96–105.
[12] Ahmad Mousavi, Zheming Gao, Lanshan Han, Alvin Lim, Quadratic surface support vector machine with L1 norm regularization, J. Ind. Manage. Optim. 18 (3) (2022) 1835–1861.
[13] Xin Yan, Yanqin Bai, Shu-Cherng Fang, Jian Luo, A proximal quadratic surface support vector machine for semi-supervised binary classification, Soft Comput. 22 (20) (2018) 6905–6919.
[14] Zheming Gao, Shu-Cherng Fang, Jian Luo, Negash Medhin, A kernel-free double well potential support vector machine with applications, European J. Oper. Res. 290 (1) (2021) 248–262.
[15] Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, Olivier Colliot, Alzheimer's Disease Neuroimaging Initiative, et al., Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database, NeuroImage 56 (2) (2011) 766–781.
[16] Yong Fan, Dinggang Shen, Ruben C Gur, Raquel E Gur, Christos Davatzikos, COMPARE: classification of morphological patterns using adaptive regional elements, IEEE Trans. Med. Imaging 26 (1) (2006) 93–105.
[17] Roman Filipovych, Christos Davatzikos, Alzheimer's Disease Neuroimaging Initiative, et al., Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI), NeuroImage 55 (3) (2011) 1109–1119.
[18] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, Enchi Liu, et al., The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception, Alzheimer's Dementia 9 (5) (2013) e111–e194.
[19] Choong Ho Lee, Hyung-Jin Yoon, Medical big data: promise and challenges, Kidney Res. Clin. Pract. 36 (1) (2017) 3.
[20] Ashwani Kumar Malik, MA Ganaie, M Tanveer, PN Suganthan, Alzheimer's Disease Neuroimaging Initiative Initiative, et al., Alzheimer's disease diagnosis via intuitionistic fuzzy random vector functional link network, IEEE Trans. Comput. Soc. Syst. (2022).
[21] Shanshan Tang, Peng Cao, Min Huang, Xiaoli Liu, Osmar Zaiane, Dual feature correlation guided multi-task learning for Alzheimer's disease prediction, Comput. Biol. Med. 140 (2022) 105090.
[22] Chun-Fu Lin, Sheng-De Wang, Fuzzy support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 464–471.

[23] Jian Luo, Ye Tian, Xin Yan, Clustering via fuzzy one-class quadratic surface support vector machine, Soft Comput. 21 (19) (2017) 5859–5865.

[24] Stephen Boyd, Lieven Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[25] Dianhui Wang, Ming Li, Stochastic configuration networks: Fundamentals and algorithms, IEEE Trans. Cybern. 47 (10) (2017) 3466–3479.

[26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[27] Chih-Chung Chang, Chih-Jen Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 1–27.

[28] Zhi-Hua Zhou, Machine Learning, Springer Singapore, 2021.

[29] Jianping Jia, Cuibai Wei, Shuoqi Chen, Fangyu Li, YI Tang, Wei Qin, Lina Zhao, Hongmei Jin, Hui Xu, Fen Wang, et al., The cost of Alzheimer's disease in China and re-estimation of costs worldwide, Alzheimer's Dementia 14 (4) (2018) 483–491.

[30] Biao Jie, Mingxia Liu, Jun Liu, Daoqiang Zhang, Dinggang Shen, Temporally constrained group sparse learning for longitudinal data analysis in Alzheimer's disease, IEEE Trans. Biomed. Eng. 64 (1) (2016) 238–249.

[31] Yun Yang, Xinfa Li, Pei Wang, Yuelong Xia, Qiongwei Ye, Multi-source transfer learning via ensemble approach for initial diagnosis of alzheimer's disease, IEEE J. Transl. Eng. Health Med. 8 (2020) 1–10.

[32] Zheming Gao, Shu-Cherng Fang, Xuerui Gao, Jian Luo, Negash Medhin, A novel kernel-free least squares twin support vector machine for fast and accurate multi-class classification, Knowl.-Based Syst. 226 (2021) 107123.

[33] Wencheng Huang, Hongyi Liu, Yue Zhang, Mirong Wei, Chuangui Tong, Wei Xiao, Bin Shuai, Railway dangerous goods transportation system risk identification: Comparisons among SVM, PSO-SVM, GA-SVM and GS-SVM, Appl. Soft Comput. (2021) 107541.

[34] Paweł Pławiak, Moloud Abdar, U. Rajendra Acharya, Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring, Appl. Soft Comput. 84 (2019) 105740.

[35] Qing Song, Wenjie Hu, Wenfang Xie, Robust support vector machine with bullet hole image classification, IEEE Trans. Syst. Man Cybern. Part C 32 (4) (2002) 440–448.

[36] Luo Luo, Yubo Xie, Zhihua Zhang, Wu-Jun Li, Support matrix machines, in: International Conference on Machine Learning, PMLR, 2015, pp. 938–947.

[37] Carl D. Meyer, Matrix Analysis and Applied Linear Algebra, Vol. 71, SIAM, 2000.