Expectation-based and Quantile-based Probabilistic Support Vector Machine Classification for Histogram-Valued Data

Fathimah Al-Ma'shumah, Mostafa Razmkhah, Sohrab Effati

Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran
razmkhah_m@um.ac.ir

Abstract: A histogram-valued random variable represents its value by a list of pairs of bins and their corresponding probabilities or relative frequencies. This type of data is a part of the symbolic data. There are many cases such as colors in image learning where histogram-valued data are naturally found. This study focuses on classification of the histogram-valued data by extending two approaches of support vector machine (SVM), namely, the expected-based and quantile-based probabilistic SVM on histogram-valued data. In both approaches, the cases of linear and nonlinear problems as well as the least-square classification are discussed. In addition, the extension to multi-class classification is also discussed. To compare the performance of the proposed procedures a simulation study has been done based on some generated data sets. The data are generated from various distributions with various parameters to represent different cases of classification, including binary and multi-class classification. Further, the methods are applied on two different real data sets. From the results, it can be concluded that our proposed methods perform well on wide range of classification problems.

Keywords: Expectation-based probabilistic SVM (EPSVM), Kernel function, Least-square SVM (LS-SVM), Quantile-based probabilistic SVM (QPSVM), Symbolic data analysis

1. Introduction

In recent years, there has been rapid development on symbolic data analysis (SDA). It is remarkably different from classical data analysis which usually represents a unit of data with a single value, be it numerical, ordinal, or nominal, the SDA offers many other alternative representations of data. Usually, the representations contain more detailed descriptions than only a single-valued representation. Such descriptions are termed as symbolic objects (SOs). The interested readers may refer to the key books. [1], [2]

Some data sets and observation can naturally consist of some SOs. For example, in some scientific measurements, the data are commonly in the forms of interval-valued data. However, the SOs can come from some pre-defined aggregations from massive classical data sets. Because of the vast progress on computerization of data, data sets are growing considerably more massive. The objective of these aggregations can be to obtain more manageable smaller data sets. *There are also some conditions in which the researchers or data analysts might not have the access to obtain the raw data and only have summary data.* Another possible objective of these pre-defined aggregations is to explore as much information as possible. Of course, specific scientific inquiries are crucial to decide the rules of these pre-defined aggregations of some data sets.

This work focuses on histogram-valued variable as a particular type of SOs, which represents its value by a list of pairs of bins and their corresponding probabilities or relative frequencies. In practice, the histogram-valued data can naturally be found, because for various economic, technical, or political reasons, the researchers or data analysts might not have the access to obtain the raw data and only have summary data in the form of histograms. For example, in various financial industries, usually some individual data such as incomes, expenditures, etc. are not single real-valued but histogram-valued data because of privacy and regulations. Such data are often needed to solve use case customer relationship management, underwriting, or fraud early warning system.

Received: March 8th, 2022. Accepted: March 31st, 2022
DOI: 10.15676/ijeei.2022.14.1.15

Summarizing data to be histogram-valued data can be considered as a data reduction technique. Besides, there might be some other analytical reasons to summarize data by constructing histogram-valued data. For example, one may be interested to analyze data in the form of groups of observations. The groups can be in the form of countries, institutions, schools, etc. An interesting example by Kejzar et al. [3] consists of modal-valued variables, and one of them is in the form of histogram-valued variable. Their example shows that representing the observations with histograms provides more information than some real values such as means or other instances. Besides, presenting all the raw data will make the data size considerably larger. Therefore, building histogram-valued data can be considered as data reduction and dimensionality reduction, along with some advantages related to some specific scientific questions.

Studies in supervised learning and unsupervised learning have been done on histogram-valued data. For example, in supervised learning, [4], [5] studied classification of histogram-valued data using learning random forest and decision trees. Linear regression of histogram-valued random variables also had been studied [6], [7]. *In the field of supervised learning method, the research on principal component analysis [8] and clustering [6] have been explored extensively in previous works*. Meanwhile, the SVM method is one of the most well-known method in the field of classification. It was originally introduced by Cortes and Vapnik [9]. Some successful applications of this method are: face detection in images [10], text categorization [11], nonlinear equalization in communication systems [12], optical character recognition [13], and producing of fuzzy rules using SVM [14]–[18].

The SVM method has some advantages. For instance, the successful applications mentioned above concluded that SVM works well with unstructured and semi-structured data such as text, images, and trees. Moreover, by converting the SVM optimization problem to its dual form, it will be much easier to solve and it enables SVM to scale relatively well to high-dimensional data. *In practice, however, the SVM model is relatively efficient in memory storing with the disadvantage of performing less optimally in huge data sets.* Above all, the main strength of SVM is its flexibility to use the kernel trick so that many complex problems can be solved. Later, there are some improvements of SVM such as LS-SVM. This method outperforms the standard SVM in the computation speed because it is based on the equality constraints [19]. However, this method is more sensitive to outliers and noises.

The classical standard SVM assumes that the data sets are in classical form, i.e., every unit of data is described with a single value. However, in practice there are many other forms of data with uncertainty such as sensor database, biometric information systems, measurements with margins of errors, and also the symbolic data mentioned before. Li et. al. [20] introduced noise elimination and probabilistic framework to solve such problems. There are many other researchers that have worked on some probabilistic set-up of SVM method [21]–[28].

Recently, Abaszade and Effati [29] proposed SVM and support vector regression (SVR) with probabilistic constraints, where support vectors are random variables. This work focuses on the classification method for histogram-valued data. More specifically, we propose an extension of PSVM that has been developed first by Abaszade and Effati [29].

The first-proposed classical SVM for histogram-valued data in [34] considers the limits of the bins and their corresponding relative frequencies as the data's features. Then, they directly used classical SVM with its inner product. Recently, instead of using the SVM's inner product, Support Histogram Machine (SHM) of [35] was recently proposed by adopting the inner product induced by the Wasserstein-Kontorovich distance which is defined between probability distributions. Specifically, they considered the bins and relative frequencies as the data's features, then they mapped all of those features through the kernel function induced by the metric. Even so, they did not include the nonlinear kernel model to deal with the case of nonlinear classification, i.e., the case when the data are linearly non-separable. Meanwhile, the methods proposed in this work are derived from PSVM model using its sufficient conditions for optimality, resulting to a mathematical justification to use means or quantiles as the representatives of the histogram-valued variables along with some adjustments in the resulting

optimization problem's formulations. Our proposed methods also include the derivations of the models for nonlinear cases using non-linear kernel methods and the Least-Squares PSVM model.

The rest of the paper is organized as follows. In Section 2, some preliminaries about the backgrounds of this study are presented. The probabilistic SVM (PSVM) is investigated for histogram-valued data in Section 3. The EPSVM is studied in this section as an extension of PSVM which can be applied to histogram-valued data as well as the other forms of data sets in the probabilistic frameworks. The QPSVM method for classification is proposed in Section 4. Some simulations on generated data and real examples are also conducted in Section 5. Finally, our conclusions are presented.

2. Preliminaries

In this section, some brief descriptions about SVM for classification as well as some details about histogram-valued data are presented.

A. Support Vector Machine for Classification

Classification aims to predict the class labels of some observations, given a set of n training samples. Suppose that χ and γ are input (feature) space and output (label) space, respectively. Let $S = \{(\mathbf{x}_i, y_i); i = 1, \dots, n\} \in \chi \times \gamma$, where $\mathbf{x}_i = (x_i^1, \dots, x_i^m)^T$ is an m-dimensional sample in the feature space. In binary classification, $y_i \in \{-1,1\}$ is the two-class labels of $\mathbf{x}_i$; in other words, S is a set of n training samples. A classifier is a functional relation f between the input and output spaces, that is $f: \chi \to \gamma$. More generally, in multi-class classification, $y_i \in \{C_1, \dots, C_{n_c}\}$ is the n_c-class labels of $\mathbf{x}_i$. The objective in every classification method is to find a classifier, should it be a function or a rule, that gives a relation between input and output spaces. The SVM method looks for the optimal separating hyperplane $f(\mathbf{x}_i) = 0$ having the maximum margin or, equivalently, minimum structural risk. In a binary classification, the classifier will be found such that $f(\mathbf{x}_i) \geq 0$ for $y_i = 1, i = 1, \dots, n$ and $f(\mathbf{x}_i) < 0$ for $y_i = -1, i = 1, \dots, n$, or equivalently, $y_i f(\mathbf{x}_i) \geq 0$ for any correct classification. *Refer to [9]* for more detailed assumptions and derivations.

The hyperplane $f(\mathbf{x})$ can be represented as $f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i + b$ where $\mathbf{w} = (w_1, \dots, w_m)^T$ is the hyperplane's normal vector and b is the scalar bias. Since most cases are not perfectly separable, some positive slack variables ξ_i are used. Therefore, the optimal separating hyperplane is obtained by solving the following minimization problem:

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$s.t. \begin{cases} y_i(w^T x_i + b) \geq 1 - \xi_i, & i = 1, \dots, n \\ \xi_i \geq 0, & i = 1, \dots, n \end{cases} \quad (1)$$

From the saddle point conditions, it follows that the partial derivatives of the Lagrange function with respect to the primal variables w, b, ξ_i , $i = 1, \dots, n$ have to vanish for optimality. Later, by proceeding this step, its corresponding dual optimization problem can be formulated. There are some reasons why the dual form is preferred in the computation. In some cases, it is simpler to solve the dual problem, i.e., the computational complexity will be lower. Moreover, using the duality concept allows to introduce the kernel trick which is one of the most important advantages of SVM. The kernel trick is a very flexible tool to deal with non-linear data as well as unstructured data.

Definition 1. *A function* $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ *is called a kernel function if there exists a Hilbert space* $\mathcal{H}$ *and a map* $\phi: \mathbb{R}^d \to \mathcal{H}$ *such that for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,
$K(\mathbf{x}, \mathbf{y}) = <\phi(\mathbf{x}), \phi(\mathbf{y})>_{\mathcal{H}}$,
where $<\cdot,\cdot>_{\mathcal{H}}$ *is an inner product on the Hilbert space* $\mathcal{H}$.

There are some well-known kernels, such as

$K(x,y) = x^T y, \quad (Linear),$
$K(x,y) = (\gamma x^T y + C)^P, p = 2,3,\ldots \quad (Polynomial),$
$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right), \ p = 2,3,\ldots \ (Gaussian),$
$K(x,y) = \exp(-\gamma\|x-y\|^2), \quad (Radial\ Basis),$
$K(x,y) = \tanh(x^T y + C) \quad (Sigmoid).$

B. Histogram-valued variables

Most of the analytic methods for interval-valued data assume a uniform distribution across the intervals. Histogram-valued data contain more information, which will produce more accurate results than those obtained by interval data. The common assumption on the bins of the histogram-valued data is that they have uniform width. For more details, see for example [1], [2], [30].

Definition 2. *Let* $\mathbf{X} = (X_1, \ldots, X_m)^T$ *be an* m*-dimensional random variable, and let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be* n *copies of* $\mathbf{X}$*. For* $i = 1, \ldots, n$*, a histogram-valued observation* $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ *is represented as*

$$\mathbf{x_i} = \{x_{ij}, j = 1, \ldots, m\} = \{[b_{ijk}, b_{ij(k+1)}), p_{ijk};\ j = 1, \ldots, m;\ k = 1, \ldots, t_{ij}\}, \quad (2)$$

where p_{ijk} *stands for the frequency of the subinterval or bin* $[b_{ijk}, b_{ij(k+1)})$*, and* t_{ij} *is the number of bins in* x_{ij}*, such that* $\sum_{k=1}^{t_{ij}} p_{ijk} = 1$.

Remark: Note that an interval-valued data is a special case of histogram-valued data with $t_{ij} = 1$ and therefore $p_{ij1} = 1$ for all i, j.

From (2), we know that each histogram-valued observation is assumed to have different lengths and different numbers of subintervals across i and j. Histogram-valued observations can be aggregated from classical raw data if they are given. By first pre-specifying the same subintervals and then assigning the corresponding relative frequencies, one can obtain the aggregated histogram data having common subintervals with common lengths and numbers of subintervals for each variable. However, there are some situations that raw data are not available. For example, suppose that we want to compare some districts by the distribution of companies considering the features such as asset, total wage, and the number of employees. These data might originate from statistical tables published by each district, where these tables might be of the histogram form. Since the data might be from different sources, the lengths and numbers of subintervals might be different across districts. It is not easy to computationally handle histogram data obtained in this case. To solve this, we can consider a transformation of such histogram data to obtain common subintervals across observations. More details can be found in the Appendix of [32]. Based on their idea, histogram-valued observations in (2) can be transformed as

$$\boldsymbol{x_i} = \{x_{ij}, j = 1, \ldots, m\} = \{[b_{jk}, b_{j(k+1)}), p_{ijk}; j = 1, \ldots, m, k = 1, \ldots, t_j\}, \quad (3)$$

where $\sum_{k=1}^{t_j} p_{ijk} = 1$. Indeed, we assume that all histogram-valued observations in this study have common subinterval lengths and the same number of subintervals for each observation.

Remark: A set of non-intersecting interval-valued objects is a special case of histogram-valued observations with common subinterval lengths and the same numbers of subintervals among observations, with p_{ijk} is either 0 or 1 for each i and j.

Billard and Diday [2] defined the empirical mean and standard deviation for a histogram-valued observation as

$$M_{ij} = \sum_{k=1}^{t_j} \left(\frac{b_{jk} + b_{j(k+1)}}{2}\right) p_{ijk} \quad (4)$$

and

$$S_{ij} = \left(\sum_{k=1}^{t_j} \frac{(b_{jk} - M_{ij})^2 + (b_{jk} - M_{ij})(b_{j(k+1)} - M_{ij}) + (b_{j(k+1)} - M_{ij})^2}{3} p_{ijk} \right)^{\frac{1}{2}},$$

respectively.

3. The PSVM classification for histogram-valued data

In this section, both cases of linear and non-linear PSVM classification problem are discussed for histogram-valued data. Then the results are extended to the LS-SVM problem. Because of the complexity of the problems, all of them are solved based on the expectations of histogram-valued data.

A. Linear Case

Here, the classification for histogram-valued data is done by extending the concept of SVM. Abaszade and Effati [29] first proposed the probabilistic SVM and SVR. The main idea is to assume that the training sets are random variables and to formulate the constraints probabilistically. More precisely, in a PSVM for a set of histogram-valued random variables, $X_i, i = 1, \dots, n$, the minimization problem (1) transformed to

$$\begin{aligned} &\min_{\boldsymbol{w},b,\xi_i} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t.} \quad \begin{cases} Pr\left(y_i(\boldsymbol{w}^T\boldsymbol{X}_i + b) \geq 1 - \xi_i\right) \geq \delta_i, & i = 1, \dots, n, \\ \xi_i \geq 0, & i = 1, \dots, n, \end{cases} \end{aligned} \tag{5}$$

where $\delta_i \in [0,1]$ represents the effect of the ith sample in determining the optimal hyperplane position.

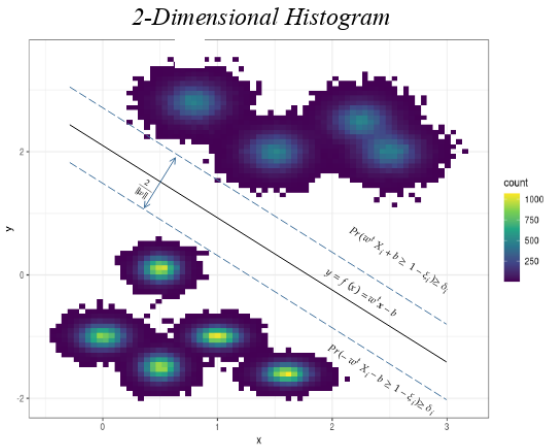

Figure 1. Illustration of two-dimensional histogram-valued data and the PSVM model

Abaszade and Effati [29] when first proposed PSVM for distributional data, didn't solve the PSVM formula by directly computing from the probabilistic expressions included in the constraints. Therefore, they formulated a sufficient condition involving sample means of the distributional data to obtain a more practical solution of the PSVM formulation. In case of histogram-valued data, it is more difficult to solve the PSVM as in (5) directly. Considering that the pdf of the histogram-valued data is a step function which is both discontinuous and nonlinear, the constraints in (5) will be nonlinear and discontinuous. An alternative solution to solve this problem is to apply Abaszade and Effati's theorem of PSVM's sufficient condition, leading to the formulation of EPSVM method.

A sufficient condition for the probabilistic constraint in the above problem to hold is

$$y_i(\mathbf{w}^T E(\mathbf{X}_i) + b) \geq 2a\delta_i + 1 - \xi_i,$$

where $a > 1$, for more details see [29]. On the other hand, using (4), the terms $E(\mathbf{X}_i) = (E(X_{i1}), \dots, E(X_{im}))^T$ can be estimated by

$$\begin{aligned} \boldsymbol{M}_i &= (M_{i1}, \dots, M_{im})^T \\ &= \left(\sum_{k=1}^{t_1} \left(\frac{b_{1k} + b_{1(k+1)}}{2} \right) p_{i1k}, \dots, \sum_{k=1}^{t_m} \left(\frac{b_{mk} + b_{m(k+1)}}{2} \right) p_{imk} \right)^T. \end{aligned} \tag{6}$$

Therefore, the PSVM problem in (5) can be rewritten as an equivalent problem

$$\begin{aligned} &\min_{\boldsymbol{w},b,\xi_i} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i \\ &\text{s.t.} \ \begin{cases} y_i(\boldsymbol{w}^T \boldsymbol{M}_i + b) \geq 2a\delta_i + 1 - \xi_i, & i = 1, \dots, n, \\ \xi_i \geq 0, & i = 1, \dots, n, \end{cases} \end{aligned} \tag{7}$$

which we call it as EPSVM problem. It follows from the saddle point condition that the partial derivatives of the corresponding Lagrange function with respect to $\mathbf{w}$, b and ξ_i should vanish for optimality. Therefore, using the following Lagrange function

$$\begin{aligned} L \ &\equiv \ L(\boldsymbol{w}, b, \xi_i, \alpha_i, \beta_i) \\ &= \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i \left(2a\delta_i + 1 - \xi_i - y_i \boldsymbol{w}^T \boldsymbol{M}_i - y_i b\right) - \sum_{i=1}^{n} \beta_i \xi_i, \end{aligned}$$

we obtain

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{M}_i = 0 \quad \Rightarrow \quad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{M}_i,$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0,$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0, \quad \Rightarrow \quad C = \alpha_i + \beta_i, \quad i = 1, \dots, n.$$

Using the above results, the problem (7) can be converted to its dual form to find the Lagrange multipliers $\alpha_1, \dots, \alpha_n$ that maximize the following objective function:

$$\begin{aligned} &\max \ -\tfrac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{M_i})^T \mathbf{M_j} + \sum_{i=1}^{n} \alpha_i (2a\delta_i + 1) \\ &\text{s.t.} \ \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0, \\ \alpha_i \in [0, C], \qquad i = 1, \dots, n. \end{cases} \end{aligned} \tag{8}$$

Now, suppose that $\alpha_{0,1}, \dots, \alpha_{0,n}$ are the optimal Lagrange multipliers. Then, the estimated optimal hyperplane coefficients equals to

$$\widehat{\mathbf{w}}_0 = \textstyle\sum_{i=1}^{n} \alpha_{0,i}\, y_i \mathbf{M}_i. \tag{9}$$

Moreover, the optimal hyperplane constant may be estimated as $\hat{b}_0 = \frac{1}{n}\sum_{i=1}^{n} \hat{b}_{0,i}$, where

$$\hat{b}_{0,i} = \begin{cases} 2a\delta_i + 1 - \widehat{\mathbf{w}}_0^T \mathbf{M}_i, & \text{if } y_i = 1, \ \alpha_{0,i} \in (0, C), \\ -2a\delta_i - 1 - \widehat{\mathbf{w}}_0^T \mathbf{M}_i, & \text{if } y_i = -1, \ \alpha_{0,i} \in (0, C). \end{cases} \tag{10}$$

For computational reasons, it is simpler to solve the dual programming (8) to find the optimal $\alpha_{0,1}, \dots, \alpha_{0,n}$; which then will be used to find the estimated optimal hyperplane coefficients (9) and constants (10). Thus, as explained in section 2.1, the hyperplane $\hat{f}(\mathbf{x}_i) = \widehat{\mathbf{w}}^T \hat{\mathbf{x}}_i + \hat{b}$ will determine the class labels of the observations. For example, in a two-class classification problem, if $\hat{f}(\mathbf{x}_i) \geq 0$ then $y_i = 1, i = 1, \dots, n$, i.e., the observation $\mathbf{x}_i$ will be assigned to the class 1; and if $\hat{f}(\mathbf{x}_i) < 0$ then $y_i = -1, i = 1, \dots, n$, then the observation will be assigned to the other class. *Moreover, the dual form allows the introduction of the kernel trick which acts as the basis of this*

classifier. This attribute enables the kernel trick to be a highly flexible tool to learn non-linear data as well as well as unstructured data.

B. Nonlinear case

When any linear hyperplane can not work well as classifier, a nonlinear problem will be considered, such that it is solved by fitting a linear classifier in a feature space with a higher dimension. Suppose that the transformation ϕ is used to transform histogram-valued data points from the m-dimensional input space into an m_1-dimensional feature space, i.e., $\phi: R^m \to R^{m_1}$, where $m < m_1$. Therefore, in the nonlinear PSVM model for histogram-valued data, the optimal separating hyperplane can be obtained by solving

$$\min_{\boldsymbol{w},b,\xi_i} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t.} \ \begin{cases} Pr(y_i(\boldsymbol{w}^T\phi(\boldsymbol{X}_i)+b) \geq 1-\xi_i) \geq \delta_i, & i=1,\dots,n, \\ \xi_i \geq 0, & i=1,\dots,n, \end{cases} \tag{11}$$

where $\phi(\mathbf{X}_i) = \left(\phi^1(\mathbf{X}_i), \dots, \phi^{m_1}(\mathbf{X}_i)\right)^T$ is a random vector. A sufficient condition for the probabilistic constraints in (11) is

$$y_i\left(\mathbf{w}^T E\left(\phi(\mathbf{X}_i)\right) + b\right) \geq 2a\delta_i + 1 - \xi_i,$$

where $a > 1$ and $E\left(\phi(\mathbf{X}_i)\right) = \left(E\left(\phi^1(\mathbf{X}_i)\right), \dots, \phi^{m_1}(\mathbf{X}_i)\right)^T$. Hence, to find the optimal values of $\mathbf{w}$, b and ξ_i, the suitable Lagrange function is given by

$$L(\mathbf{w},b,\xi_i,\alpha_i,\beta_i) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(2a\delta_i + 1 - \xi_i - y_i\mathbf{w}^T E\left(\phi(\mathbf{X}_i)\right) - y_i b\right) - \sum_{i=1}^{n}\beta_i\,\xi_i.$$

Similar to the case of linear EPSVM, it can be shown that the problem (11) can be converted to its dual form to find the Lagrange multipliers $\alpha_1, \dots, \alpha_n$ that maximize the following objective function

$$\max \ -\tfrac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j y_i y_j \left(E\left(\phi(\mathbf{X}_i)\right)\right)^T E\left(\phi(\mathbf{X}_i)\right) + \sum_{i=1}^{n}\alpha_i\,(2a\delta_i + 1), \tag{12}$$
$$\text{s.t. } \{\textstyle\sum_{i=1}^{n}\alpha_i\,y_i = 0, \alpha_i \in [0,C], \quad i = 1,\dots,n.$$

In practice, the problem (12) can be solved by some available optimization tools such as Sequential Minimal Optimization (SMO) or other methods such as Bregman methods. See [36] for more details.

From the Karush-Kuhn-Tucker (KKT) conditions for optimality, we obtain, for $i = 1, \dots, n$,

$$\begin{cases} \alpha_i\left(2a\delta_i + 1 - \xi_i - y_i\mathbf{w}^T E\left(\phi(\mathbf{X}_i)\right) - y_i b\right) = 0, \\ \beta_i\xi_i = 0, \\ (C-\alpha_i)\xi_i = 0. \end{cases}$$

Denoting the optimal Lagrange multipliers by $\alpha_{0,1}, \dots, \alpha_{0,n}$, the estimated optimal hyperplane coefficients are obtained as

$$\widehat{\mathbf{w}}_0 = \textstyle\sum_{i=1}^{n}\alpha_{0,i}\,y_i E\left(\phi(\mathbf{X}_i)\right) \tag{13}$$

Also, the estimated optimal hyperplane constant equals to $\hat{b}_0 = \frac{1}{n}\sum_{i=1}^{n}\hat{b}_{0,i}$, where

$$\hat{b}_{0,i} = \begin{cases} 2a\delta_i + 1 - \widehat{\mathbf{w}}_0^T E\left(\phi(\mathbf{X}_i)\right), & \text{if } y_i = 1, \alpha_{0,i} \in (0,C), \\ -2a\delta_i - 1 - \widehat{\mathbf{w}}_0^T E\left(\phi(\mathbf{X}_i)\right), & \text{if } y_i = -1, \alpha_{0,i} \in (0,C). \end{cases} \tag{14}$$

Note that, in practice, the term $E\left(\phi(\mathbf{X}_i)\right)$ can not be easily computed. It is easier to use kernel trick to avoid direct nonlinear mapping ϕ. Thus, by applying the properties of kernel trick (see [32]), the objective function in (12) is equivalent to

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j y_i y_j K\left(E(\mathbf{X}_i), E\left(\mathbf{X}_j\right)\right) + \sum_{i=1}^{n}\alpha_i\,(2a\delta_i + 1).$$

On the other hand, the term $E(\mathbf{X}_i)$ can be estimated by using the empirical mean of the histogram-valued random variable $\mathbf{M}_i$ presented in (6). Hence, the dual optimization problem (12) can be rewritten as an equivalent problem

$$\max \; -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j y_i y_j K(\mathbf{M}_i,\mathbf{M}_j)+\sum_{i=1}^{n}\alpha_i\,(2a\delta_i+1),$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{n}\alpha_i\,y_i=0,\\ \alpha_i\in[0,C], \qquad i=1,\dots,n.\end{cases}$$

Using (13) and (14), the optimal separating hyperplane is:

$$\hat{\mathbf{w}}_0^T\phi(\mathbf{M})+\hat{b}_0=\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{0,i}\,y_i K(\mathbf{M}_i,\mathbf{M}_j)+\hat{b}_0=0.$$

C. Least squares EPSVM (LS-EPSVM) for histogram data

The least squares version of EPSVM is a quadratic programming problem containing only equality constraints. Hence, its computational complexity and speed are much faster than EPSVM. This algorithm is suitable for classifying data sets with large sample sizes. The least squares expected-based version of (11) is defined as:

$$\begin{aligned}&\min_{\boldsymbol{w},b,e_i}\ \frac{1}{2}\|\boldsymbol{w}\|^2+\gamma\frac{1}{2}\sum_{i=1}^{n}e_i^2,\\ &\text{s.t.}\ \ y_i\left(\boldsymbol{w}^T E(\phi(\boldsymbol{X}_i))+b\right)=2a\delta_i+1-e_i,\quad i=1,\dots,n.\end{aligned}\tag{14}$$

The Lagrange function of the problem (15) is defined as

$$L(\mathbf{w},b,e_i,\alpha_i)\ =\frac{1}{2}\mathbf{w}^T\mathbf{w}+\gamma\frac{1}{2}\sum_{i=1}^{n}e_i^2-\sum_{i=1}^{n}\alpha_i\left(y_i\left(\mathbf{w}^T E\left(\phi(\mathbf{X}_i)\right)+b\right)-2a\delta_i-1+e_i\right),$$

where real values $\alpha_1,\dots,\alpha_n$ are Lagrange multipliers. Based on the optimality conditions, the partial derivatives of the Lagrange function should vanish.

$$\frac{\partial L}{\partial \mathbf{w}}\ =\mathbf{w}-\sum_{i=1}^{n}\alpha_i\,y_i E\left(\phi(\mathbf{X}_i)\right)=0\quad\Rightarrow\quad\mathbf{w}=\sum_{i=1}^{n}\alpha_i\,y_i E\left(\phi(\mathbf{X}_i)\right),$$

$$\frac{\partial L}{\partial b}\ =-\sum_{i=1}^{n}\alpha_i\,y_i=0\quad\Rightarrow\quad\sum_{i=1}^{n}\alpha_i\,y_i=0,$$

$$\frac{\partial L}{\partial e_i}\ =\gamma e_i-\alpha_i=0\quad\Rightarrow\quad\alpha_i=\gamma e_i,\quad i=1,\dots,n,$$

$$\frac{\partial L}{\partial \alpha_i}\ =y_i\left(\mathbf{w}^T E\left(\phi(\mathbf{X}_i)\right)+b\right)-2a\delta_i-1+e_i=0,\quad i=1,\dots,n.$$

By eliminating $\mathbf{w}$ and e_i, the solution is

$$\begin{bmatrix}0 & \mathbf{y}^T\\ \mathbf{y} & \Omega+\gamma^{-1}I\end{bmatrix}\begin{bmatrix}b\\ \alpha\end{bmatrix}=\begin{bmatrix}0\\ \vec{1}+2a\delta\end{bmatrix},$$

where

$$\mathbf{y}=[y_1,\dots,y_n],\ \vec{1}=[1,\dots,1],\ \alpha=[\alpha_1,\dots,\alpha_n],\ \delta=[\delta_1,\dots,\delta_n],$$

and for $i,j=1,\dots,n$,

$$\Omega_{ij}=y_i y_j K\left(E(\mathbf{X}_i),E\left(\mathbf{X}_j\right)\right).$$

and the terms $E(\mathbf{X}_i)$ can be estimated by $\mathbf{M}_i$ in (6).

In this section, the PSVM classification problem was studied on the basis of the expectation of histogram-valued data. An alternative method is to use the quantiles instead of the expectation, which is discussed in the next section.

4. The QPSVM classification for histogram-valued data

The idea used in this extension of PSVM model is to provide more practical approach to the the PSVM classification problem for histogram-valued data based on quantiles. Note that both of expectation and quantiles of different order are some measures to evaluate the concentration of a given data set or a random variable. However, the quantiles may be more accurate when the data are asymmetric or there are some outliers among a data set. The quantile function of a histogram-valued random variable basically is the inverse of it's cumulative distribution function (cdf). Suppose that $\mathbf{X}_1, \dots, \mathbf{X}_n$ are histogram-valued random variables as in (3). By assuming that the observations in each bin of the histogram-valued random variable $X_{ij} = \{[b_{jk}, b_{j(k+1)}), p_{ijk};\ k = 1, \dots, t_j\}$, for $i = 1, \dots, n$ and $j = 1, \dots, m$, are uniformly distributed, the X_{ij} has the following probability density function

$$f_{ij}(x) = \begin{cases} p_{ij1}, & b_{ij1} \le x < b_{ij2}, \\ p_{ij2}, & b_{ij2} \le x < b_{ij3}, \\ \vdots & \\ p_{ijt_j}, & b_{ijt_j} \le x < b_{ij(t_j+1)}. \end{cases}$$

Therefore, it can be shown that the cdf of X_{ij} is given by

$$F_{ij}(x) = \begin{cases} 0, & x < b_{ij1}, \\ p_{ij1}\left(\dfrac{x - b_{ij2}}{b_{ij2} - b_{ij1}}\right), & b_{ij1} \le x < b_{ij2}, \\ F(b_{ij2}) + p_{ij2}\left(\dfrac{x - b_{ij2}}{b_{ij3} - b_{ij2}}\right), & b_{ij2} \le x < b_{ij3}, \\ \vdots & \\ F\left(b_{ijt_j}\right) + p_{ijt_j}\left(\dfrac{x - b_{ijt_j}}{b_{ijt_j+1} - b_{ijt_j}}\right), & b_{ijt_j} \le x < b_{ij(t_j+1)}, \\ 1 & x \ge b_{ij(t_j+1)}. \end{cases}$$

Hence, the corresponding quantile function is

$$F_{X_{ij}}^{-1}(x) = \begin{cases} b_{ij1} + \frac{x}{z_{ijl}}\left(b_{ij2} - b_{ij1}\right), & 0 \le x \le z_{ij1}, \\ b_{ij2} + \frac{x}{z_{ij2}}\left(b_{ij3} - b_{ij2}\right), & z_{ij1} \le x \le z_{ij2}, \\ \vdots & \\ b_{ijt_j} + \frac{x}{z_{ijt_j}}\left(b_{ijt_j+1} - b_{ijt_j}\right), & z_{ijt_j} \le x \le z_{ijt_j+1}, \end{cases} \tag{15}$$

where

$$z_{ij\ell} = \begin{cases} 0, & \text{if } \ell = 0, \\ \sum_{h=1}^{\ell} p_{ijh}, & \text{if } \ell = 1, \dots, t_j. \end{cases}$$

The main idea in this section is to write the relationship between expectation and quantile for any random variable X to be

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx \ge \int_{Q(p)}^{\infty} x\, f(x)\, dx \ge Q(p) P\big(X \ge Q(p)\big) = Q(p)(1 - p), \tag{16}$$

where $Q(p) = F_X^{-1}(p)$ stands for the p-th quantile of X. In the sequel, both cases of linear and nonlinear problems are investigated.

A. Linear Case

Suppose that we want to solve the classification problem (1) for histogram-valued random variables in (3).

Theorem 3. *Let* $\mathbf{X}_i$ *be a histogram-valued random variable with the corresponding quantile function* $\mathbf{Q}_i(p) = F_{\mathbf{X}_i}^{-1}(p) = \left(F_{X_{i1}}^{-1}(p), \dots, F_{X_{im}}^{-1}(p)\right)^T$. *A sufficient condition for* $Pr(y_i(\mathbf{w}^T\mathbf{X}_i + b) \geq 1 - \xi_i) \geq \delta_i$ *to hold is* $y_i((1-p)\mathbf{w}^T\mathbf{Q}_i(p) + b) \geq 2a\delta_i + 1 - \xi_i$, *for a given* $p \in (0,1)$ *and* $a > 1$.

Using (16), we obtain

$$E(\mathbf{X}_i) = (E(X_{i1}), \dots, E(X_{im}))^T \geq (1-p)\mathbf{Q}_i(p).$$

Hence,

$$y_i(\mathbf{w}^T E(\mathbf{X}_i) + b) \geq y_i\big((1-p)\mathbf{w}^T\mathbf{Q}_i(p) + b\big).$$

Now, let $V_i = y_i(\mathbf{w}^T\mathbf{X}_i + b)$ and choose $a > 1$, such that

$$-a \leq V_i - \xi_i \leq V_i \leq V_i + \xi_i \leq a.$$

Then, we obtain

$$\begin{aligned} P(V_i + \xi_i \geq 1) &\geq \frac{E(V_i) + \xi_i - 1}{2a} = \frac{y_i(\mathbf{w}^T E(\mathbf{X}_i) + b) + \xi_i - 1}{2a} \\ &\geq \frac{y_i\big((1-p)\mathbf{w}^T\mathbf{Q}_i(p) + b\big) + \xi_i - 1}{2a}, \end{aligned}$$

where the first inequality is obtained by proceeding in lines similar to the Theorem 3.2 of. Therefore, if we take

$$\frac{y_i\big((1-p)\mathbf{w}^T\mathbf{Q}_i(p) + b\big) + \xi_i - 1}{2a} \geq \delta_i$$

or equivalently,

$$y_i\big((1-p)\mathbf{w}^T\mathbf{Q}_\mathbf{i}(p) + b\big) \geq 2a\delta_i + 1 - \xi_i,$$

then, the inequality $P(V_i + \xi_i \geq 1) \geq \delta_i$ is obtained. Hence, the proof completes. ■

Using the above theorem, to solve the problem (1) for histogram-valued random variables, it is sufficient to solve the following optimization problem

$$\begin{aligned} &\min_{\boldsymbol{w},b,\xi} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i, \\ &\text{s.t.} \ \begin{cases} y_i\left((1-p)\boldsymbol{w}^T\boldsymbol{Q}_i(p) + b\right) \geq 2a\delta_i + 1 - \xi_i, & i = 1, \dots, n, \\ \xi_i \geq 0, & i = 1, \dots, n. \end{cases} \end{aligned} \tag{17}$$

Note that the vector $\mathbf{Q}_i(p)$ is in fact the quantile of order p of a histogram-valued random variable $\mathbf{X}_i$, which may be computed by using (15). In practice, different values of the p can be used to represent the concentration of $\mathbf{X}_i$, such that if we want to use the first quartile, median, or the third quartile of $\mathbf{X}_i$, we have to put $p = 0.25$, $p = 0.5$ or $p = 0.75$, respectively.

The corresponding Lagrange function to problem (17) is

$$\begin{aligned} L(w, b, \xi_i, \alpha_i, \beta_i) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ &\quad + \sum_{i=1}^{n}\alpha_i\left[2a\delta_i + 1 - \xi_i - y_i(1-p)\mathbf{w}^T\mathbf{Q}_i(p) - y_i b\right] - \sum_{i=1}^{n}\beta_i\,\xi_i. \end{aligned}$$

By doing some algebraic calculations, it can be shown that the problem (17) can be converted to its dual form to find the Lagrange multipliers $\alpha_1, \dots, \alpha_n$ which maximizes the objective function

$$\max\ -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j y_i y_j (1-p)\big(\mathbf{Q_i}(p)\big)^T\mathbf{Q_j}(p)+\sum_{i=1}^{n}\alpha_i\,(2a\delta_i+1),$$

$$\text{s.t.}\ \begin{cases}\sum_{i=1}^{n}\alpha_i\,y_i=0,\\ \alpha_i\in[0,C], \qquad i=1,\dots,n.\end{cases}$$

Therefore, the KKT conditions for optimality lead to hold the following equations for $i=1,\dots,n$,

$$\begin{cases}\alpha_i(2a\delta_i+1-\xi_i-y_i(1-p)\mathbf{w}^T\mathbf{Q}_i(p)-y_i b)=0,\\ \beta_i\xi_i=0,\\ (C-\alpha_i)\xi_i=0.\end{cases}$$

Denoting the optimal Lagrange multipliers by $\alpha_{0,i}$, the estimated optimal hyperplane coefficients equals to

$$\widehat{\mathbf{w}}_0=\sum_{i=1}^{n}\alpha_{0,i}\,y_i(1-p)\mathbf{Q}_i(p).$$

Also, the estimated optimal hyperplane constant equals to the average $\hat{b}_0=\frac{1}{n}\sum_{i=1}^{n}\hat{b}_{0,i}$, where

$$\hat{b}_{0,i}=\begin{cases}2a\delta_i+1-(1-p)\widehat{\mathbf{w}}_0^T\mathbf{Q_i}(p), & \text{if } y_i=1,\ \alpha_{0,i}\in(0,C),\\ -2a\delta_i-1-(1-p)\widehat{\mathbf{w}}_0^T\mathbf{Q_i}(p), & \text{if } y_i=-1,\ \alpha_{0,i}\in(0,C).\end{cases}$$

B. Nonlinear case

In a nonlinear SVM classification problem for histogram-valued data, the optimal separating hyperplane can be obtained by solving (11). In this section, quantiles are used to solve the problem. Suppose that $\mathbf{X}_1,\dots,\mathbf{X}_n$ are histogram-valued random variables as presented in (3). Also, suppose that $\phi: R^m\to R^{m_1}$ is a transformation from m-dimensional input space into m_1 dimensional feature space, where $m<m_1$. The nonlinear ϕ-quantile $\mathbf{Q}_i^{\phi}(p)$, for $i=1,\dots,n$, is defined to be

$$\mathbf{Q}_i^{\phi}(p)=F^{-1}_{\phi(\mathbf{X}_i)}(p)=\left(F^{-1}_{\phi^1(\mathbf{X}_i)}(p),\dots,F^{-1}_{\phi^{m_1}(\mathbf{X}_i)}(p)\right)^T.$$

Using Theorem 3, the following statement is a sufficient condition for the constraints in the problem (11):

$$y_i\left((1-p)\mathbf{w}^T\mathbf{Q}_i^{\phi}(p)+b\right)\geq 2a\delta_i+1-\xi_i,$$

where $a>1$. Therefore, the optimal separating hyperplane is the optimal solution of

$$\min_{\boldsymbol{w},b,\xi}\ \frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\xi_i,$$

$$\text{s.t.}\ \begin{cases} y_i\left((1-p)\boldsymbol{w}^T\boldsymbol{Q}_i^{\phi}(p)+b\right)\geq 2a\delta_i+1-\xi_i, & i=1,\dots,n,\\ \xi_i\geq 0, & i=1,\dots,n,\end{cases} \tag{18}$$

where $\delta_i\in[0,1]$, for $i=1,\dots,n$.

Similar to the previous subsection, the problem (18) can be converted to it's dual form to find the optimal multipliers $\alpha_1,\dots,\alpha_n$ which maximizes the following objective function

$$\max\ -\tfrac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,\alpha_j y_i y_j(1-p)^2\left(\mathbf{Q}_i^{\phi}(p)\right)^T\mathbf{Q}_j^{\phi}(p)+\sum_{i=1}^{n}\alpha_i\,(2a\delta_i+1),$$

$$\text{s.t.}\ \begin{cases}\sum_{i=1}^{n}\alpha_i\,y_i=0,\\ \alpha_i\in[0,C], \qquad i=1,\dots,n.\end{cases} \tag{19}$$

In practice, the problem (19) can be solved by some available optimization tools such as Sequential Minimal Optimization (SMO) or other methods such as Bregman methods. See [36] for more details.

From the KKT conditions for optimality we obtain, for $i = 1, \dots, n$,

$$\begin{cases} \alpha_i\left(2a\delta_i + 1 - \xi_i - y_i(1-p)\mathbf{w}^T\mathbf{Q}_i^{\phi}(p) - y_i b\right) = 0, \\ \beta_i\xi_i = 0, \\ (C - \alpha_i)\xi_i = 0. \end{cases}$$

Denoting the optimal value of α_i by $\alpha_{0,i}$, the estimated optimal hyperplane coefficient equals to

$$\widehat{\mathbf{w}}_0 = \sum_{i=1}^{n} \alpha_{0,i}\, y_i(1-p)\mathbf{Q}_i^{\phi}(p). \tag{20}$$

Moreover, the estimated optimal hyperplane constant equals to $\hat{b}_0 = \frac{1}{n}\sum_{i=1}^{n} \hat{b}_{0,i}$, where

$$\hat{b}_{0,i} = \begin{cases} 2a\delta_i + 1 - (1-p)\widehat{\mathbf{w}}_0^T\mathbf{Q}_i^{\phi}(p), & \text{if } y_i = 1, \alpha_{0,i} \in (0, C), \\ -2a\delta_i - 1 - (1-p)\widehat{\mathbf{w}}_0^T\mathbf{Q}_i^{\phi}(p), & \text{if } y_i = -1, \alpha_{0,i} \in (0, C). \end{cases} \tag{21}$$

Let K be a kernel matrix such that $K(\mathbf{x}_i, \mathbf{x}_j) = \left(\phi(\mathbf{x}_i)\right)^T \phi(\mathbf{x}_j)$. By using the properties of kernel function and applying the quantile of order p as a representation of histogram-valued random variables, the dual optimization problem (19) can be written as

$$\max\ -\frac{1}{2}(1-p)^2 \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j y_i y_j K\left(\mathbf{Q}_i(p), \mathbf{Q}_j(p)\right) + \sum_{i=1}^{n} \alpha_i\,(2a\delta_i + 1),$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{n} \alpha_i\, y_i = 0, \\ \alpha_i \in [0, C], \qquad i = 1, \dots, n. \end{cases}$$

Based on the optimality conditions, the estimated coefficients and constants of the separating hyperplane are (20) and (21). Hence, the optimal separating hyperplane is:

$$\widehat{\mathbf{w}}_0^T\mathbf{Q}_{x_i}(p) + \hat{b}_0 = (1-p)^2 \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_{0,i}\, y_i K\left(Q_{ik}(p), Q_{ij}(p)\right) + \hat{b}_0 = 0.$$

C. Least squares QPSVM (LS-QPSVM) for histogram data

Here, the least squares approach is extended to the QPSVM classification problem. The least squares version of (17) is defined as follows

$$\begin{aligned} &\min_{\boldsymbol{w}, b, e_i} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + \gamma\frac{1}{2}\sum_{i=1}^{n} e_i^2 \\ &\text{s.t.} \quad y_i\left((1-p)\boldsymbol{w}^T\boldsymbol{Q}_i^{\phi}(p) + b\right) = 2a\delta_i + 1 - e_i, \quad i = 1, \dots, n. \end{aligned} \tag{22}$$

The Lagrange function of the problem (22) is defined as

$$L(\mathbf{w}, b, e_i, \alpha_i) \quad = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \gamma\frac{1}{2}\sum_{i=1}^{n} e_i^2 - (1-p)\sum_{i=1}^{n} \alpha_i\left(y_i\left(\mathbf{w}^T\mathbf{Q}_i^{\phi}(p) + b\right) - 2a\delta_i - 1 + e_i\right),$$

where $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ are Lagrange multipliers. Based on the optimality conditions, the partial derivatives of the Lagrange function should vanish. Therefore, we obtain

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - (1-p)\sum_{i=1}^{n} \alpha_i y_i \mathbf{Q}_i^{\phi}(p) = 0 \quad \Rightarrow \quad \mathbf{w} = (1-p)\sum_{i=1}^{n} \alpha_i y_i \mathbf{Q}_i^{\phi}(p),$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0,$$

$$\frac{\partial L}{\partial e_i} = \gamma e_i - \alpha_i = 0 \quad \Rightarrow \quad \alpha_i = \gamma e_i, \quad i = 1, \dots, n,$$

$$\frac{\partial L}{\partial \alpha_i} = y_i \left((1-p)\mathbf{w}^T \mathbf{Q}_i^{\phi}(p) + b \right) - 2a\delta_i - 1 + e_i = 0, \quad i = 1, \dots, n.$$

Eliminating $\mathbf{w}$ and e_i, the solution is

$$\begin{bmatrix} 0 & \mathbf{y}^T \\ \mathbf{y} & \mathbf{\Omega} + \gamma^{-1}\mathbf{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \vec{1} + 2a\boldsymbol{\delta} \end{bmatrix},$$

where $\mathbf{I}$ stands for the identity matrix, $\mathbf{y}^T = (y_1, \dots, y_n)$, $\vec{1} = (1, \dots, 1)^T$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^T$, $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^T$, and $\mathbf{\Omega} = [\Omega_{ij}]$ such that $\Omega_{ij} = (1-p)^2 y_i y_j K\left(\mathbf{Q}_i(p), \mathbf{Q}_j(p)\right)$ for $i, j = 1, \dots, n$.

5. Extension to Multi-class Classification

The proposed methods are extended from binary to multi-class cases using the one-against-all approach [37]. It is frequently used for multi-class problems because it is a relatively simple setup and it has many computational advantages. Suppose that there are k > 2 classes in the data and $y_i \in \{1, \dots, k\}$ denotes the class to which the i-th observation belongs. The one-against-all approach fits k different binary classifiers $\widehat{f}_1, \widehat{f}_2, \dots, \widehat{f}_k$ separately, and each classifier $\widehat{f}_j, j \in \{1, \dots, k\}$ assigns the observations to the class k versus the rest. For each $j \in \{1, \dots, k\}$, the process is performed by replacing y_i as a positive class for the observations in class j and as a negative class for those not in class j. At the end, k different classifiers will be obtained. Suppose that the tuning parameter is λ. Therefore, for a single observation x_o in the test set, let $\left(\widehat{f}_1^{\lambda}(x\_0), \widehat{f}_2^{\lambda}(x_o), \dots, \widehat{f}_k^{\lambda}(x_o)\right)^T$ be a fitted vector evaluating x_0 at a certain value of the tuning parameter λ. Then, the predicted label y_0 for x_0 is predicted by

$$y_0^{\lambda} = argmax \ \widehat{f}_j^{\lambda}(x_0)$$

6. Experimental Results

To investigate the performance of the proposed procedures in the paper, we present some experimental results here by doing some simulation and applying the results on different real data sets.

A. Simulations

In simulating binary classification problems for histogram-valued data, we first construct histogram-valued data by generating data from two classes, then construct histogram-valued data from some groups of observations. After obtaining the histogram-valued data with binary class labels, we then assume that the raw data are not available and we only have the access to the histogram-valued data.

The simulation consists of the following steps. First, suppose that two groups of observations having certain distributions are labeled by classes C_1 and C_2. Their distributions are chosen in a way so that they become easy to distinguish. It is more preferred to have some adequate translations on the chosen distributions. For example, if the class C_1 consists of normally distributed samples with mean μ and variance σ^2, denoted by $N(\mu, \sigma^2)$, then the distribution of class C_2 will be $N(\mu + \theta, \sigma^2)$, where θ is the translation parameter. Let us denote the number of data generated for the classes C_1 and C_2 by N_1 and N_2 and corresponding observations by

$X_i (i = 1, \dots, N_1)$ and $W_j (j = 1, \dots, N_2)$, respectively. Then, the following cases are considered in this section:

Case 1: Normal distribution in which $X_i \sim N(\mu, \sigma^2)$ and $W_j \sim N(\mu + \theta, \sigma^2)$. Note that the $N(\mu, \sigma^2)$ distribution has the probability density function (pdf)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where μ is the mean of the distribution and σ is the standard deviation.

Case 2: Gamma distribution, where $X_i \sim \Gamma(\mu, \sigma)$ and $W_j \sim \Gamma(\mu + \theta, \sigma)$. The $\Gamma(\mu, \sigma)$ represents the gamma distribution with shape parameter μ and scale parameter σ, which has the pdf

$$f(x) = \frac{1}{\sigma^\mu \Gamma(\mu)} x^{\mu-1} e^{-\frac{x}{\sigma}}, \quad x > 0,\ \mu > 0, \sigma > 0,$$

where $\Gamma(\cdot)$ stands for the gamma function.

Case 3: Beta distribution, such that $X_i \sim Beta(\mu, \sigma)$ and $W_j = Z + \theta$, where $Z \sim Beta(\mu, \sigma)$. The $Beta(\mu, \sigma)$ distribution, where both of μ and σ are called the shape parameters, has the pdf

$$f(x) = \frac{\Gamma(\mu + \sigma)}{\Gamma(\mu)\Gamma(\sigma)} x^{\mu-1}(1-x)^{\sigma-1}, \quad 0 < x < 1, \mu > 0, \sigma > 0.$$

Case 4: Log-normal (LN) distribution in which $X_i \sim LN(\mu, \sigma)$ and $W_j \sim LN(\mu + \theta, \sigma)$. The pdf of $LN(\mu, \sigma)$ distribution is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\{-\frac{(\log x - \mu)^2}{2\pi\sigma^2}\}.$$

Note that, in Case 1 and Case 3, the parameter θ means shift in the distributions. Meanwhile, in Case 2 the change in the first parameter will give a change in the shape of distribution curve; and in Case 4, the parameter θ acts as a scale change in the location of the LN distribution. Figure 1 illustrates the shapes of the empirical pdfs of two observations from two different classes from Case 1-4. Later, the data are represented as histogram-valued data to simulate the classification using PSVM and QSVM methods.

Suppose that the number of generated data is N such that $N_1 = N_2 = \frac{N}{2}$. Our purpose is to build N_{hist} histograms from all N generated data. From N_1 data, we build $\frac{N_{hist}}{2}$ histograms with identical bins. Note that each histogram is constructed using $\frac{N}{N_{hist}}$ data. After that, we save the first N_1 labels, i.e., $y = -1$ and do the similar thing for the next N_2 data for the label $y = 1$. Here, it is assumed that all histograms have the same bin lengths and numbers; more precisely, we consider 10 bins for each histogram.

Randomly-chosen 80% of N_{hist} histogram-valued data will then be considered as the training set, and the remaining data are the test set. Finally, the accuracies of PSVM and QPSVM with various kernel types and with some order of quantiles, $p = 0.25, 0.5, 0.75$, are computed. The accuracy is calculated as the proportion of observations in the test set that were predicted correctly, divided by the total number of observations in the test set. The results are presented in Table 1.

Similarly, we simulate multi-class classification by assigning the same numbers of histogram-valued observations in each class:

Case 5: N(0,9), N(0.5,4), and N(-0.4,4)

Case 6: Gamma(1,4), Gamma(2,2), and Gamma(16,0.2)

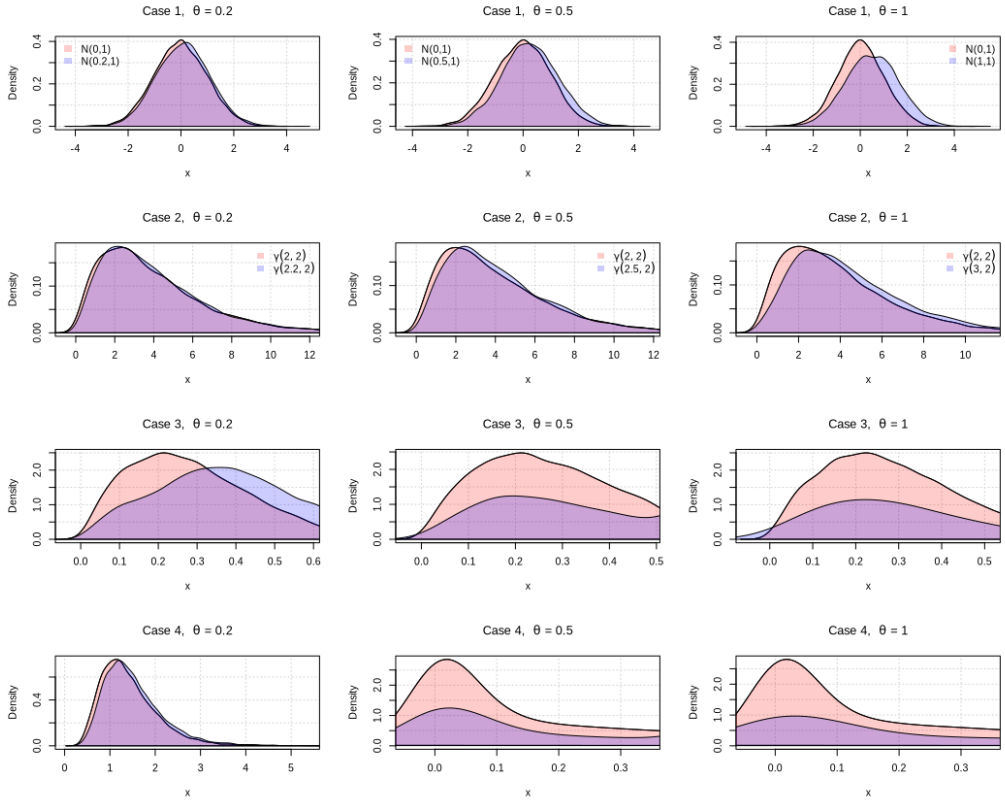

Figure 1. Empirical pdfs of two observations from two different classes from Case 1-4

Table 1. The accuracies of the classification simulations using various methods and kernels

			Accuracies Using Various Methods			
Cases	θ	Kernel Type	EPSVM	QPSVM $p = 0.25$	$p = 0.5$	$p = 0.75$
Case 1: N(0,1)	0.2000	linear	0.8156	0.7521	0.7563	0.7250
		polynomial	0.8021	0.7760	0.7438	0.7198
		radial	0.8052	0.7760	0.7583	0.7281
		sigmoid	0.8010	0.7563	0.7323	0.7167
	0.5000	linear	0.9950	0.9700	0.9750	0.9100
		polynomial	0.9850	0.9800	0.9800	0.9250
		radial	0.9950	0.9800	0.9850	0.9150
		sigmoid	1.0000	0.9700	0.9600	0.9050
	1.0000	linear	1.0000	0.9900	1.0000	0.9950
		polynomial	1.0000	0.9900	1.0000	1.0000
		radial	1.0000	1.0000	1.0000	1.0000
		sigmoid	1.0000	1.0000	1.0000	0.9900
Case 2: $\Gamma(2,2)$	0.2000	linear	0.6040	0.6245	0.5315	0.4815
		polynomial	0.5990	0.6105	0.5420	0.5395
		radial	0.5705	0.6010	0.5410	0.5080
		sigmoid	0.5965	0.5920	0.5405	0.5005
	0.5000	linear	0.7890	0.7870	0.7030	0.6020
		polynomial	0.7805	0.7985	0.7010	0.6255
		radial	0.7785	0.8005	0.7050	0.6355
		sigmoid	0.7900	0.7870	0.7040	0.5805
	1.0000	linear	0.9615	0.9410	0.8910	0.7990
		polynomial	0.9575	0.9400	0.8890	0.7905
		radial	0.9575	0.9385	0.8930	0.7915

			Accuracies Using Various Methods			
Cases	θ	Kernel Type	EPSVM	QPSVM		
				$p = 0.25$	$p = 0.5$	$p = 0.75$
		sigmoid	0.9590	0.9415	0.8825	0.7845
Case 3: Beta(2,5)	0.2000	linear	1.0000	0.9995	0.9980	0.9985
		polynomial	1.0000	0.9975	0.9975	0.9975
		radial	1.0000	0.9960	0.9970	0.9965
		sigmoid	1.0000	0.9985	0.9965	0.9945
	0.5000	linear	1.0000	1.0000	1.0000	1.0000
		polynomial	1.0000	0.9995	1.0000	0.9975
		radial	1.0000	1.0000	1.0000	1.0000
		sigmoid	1.0000	1.0000	0.9995	1.0000
	1.0000	linear	1.0000	1.0000	1.0000	1.0000
		polynomial	1.0000	1.0000	1.0000	1.0000
		radial	1.0000	1.0000	1.0000	1.0000
		sigmoid	1.0000	1.0000	1.0000	1.0000
Case 4: LN(0.25,0.44)	0.2000	Linear	0.9095	0.9280	0.8230	0.7540
		polynomial	0.9095	0.9260	0.8270	0.7555
		radial	0.9050	0.9250	0.8270	0.7580
		sigmoid	0.9110	0.9225	0.8220	0.7495
	0.5000	Linear	0.9969	0.9969	0.9875	0.9719
		polynomial	0.9938	1.0000	0.9906	0.9750
		radial	0.9969	1.0000	0.9906	0.9656
		sigmoid	1.0000	0.9969	0.9875	0.9625
	1.0000	Linear	1.0000	1.0000	1.0000	0.9985
		polynomial	0.9995	1.0000	0.9999	0.9999
		radial	1.0000	1.0000	1.0000	0.9999
		sigmoid	1.0000	1.0000	1.0000	0.9945
Case 5 (multi-class)		Linear	0.8000	0.8667	0.9667	1.000
		polynomial	0.7667	0.8788	0.9788	1.000
		radial	0.8000	0.8591	0.9500	1.000
		sigmoid	0.8000	0.8677	0.952	0.983
Case 6 (multi-class)		Linear	1.0000	1.0000	1.0000	1.0000
		polynomial	1.0000	1.0000	1.0000	1.0000
		radial	1.0000	1.0000	1.0000	1.0000
		sigmoid	1.0000	0.9667	1.0000	1.000

From Table 1, it can be seen that:
For a given distribution and a given classification method, the accuracy increases when θ increases, which is expected to happen. The higher the θ is, the more distinguishable the distributions becomes.

In the Cases 1-4, the accuracy of the QPSVM decreases when the order of quantile p increases. On the contrary, in Case 5, the higher the quantile's order is, the higher the accuracy will be.
For the case of $\Gamma(2,2)$ distribution, the accuracy of the QPSVM with $p = 0.25$ is more than that of EPSVM method. Also, for the case of $Beta(2,5)$ distribution, the accuracies of the QPSVM and the EPSVM methods are approximately the same. Comparing with the illustrations in Figure 1, it seems that the QPSVM may be better than EPSVM method when the distribution of the data is asymmetric.
In some cases, applying various kernels can improve the accuracies.
EPSVM and QPSVM perform well with the multi-class classification problems examples.

B. Aplication to Real Data Sets

Here, two data sets for binomial classification are used. These data sets are derived and summarized into histogram-valued data. Let us denote the number of single observations by N_{data}, in which every N_{member} of data are grouped as series of observations to build the number of N_{hist} histogram-valued data, each containing the number of N_{bins} bins, and each has one of

two specific class labels. The histogram-valued data are split into N_{train} training set and N_{test} test set. The summarized description of the real data sets is reported in Table 2.

Table 2. Summary of real data sets

Data Set	N_{data}	N_{hist}	N_{member}	N_{bins}	N_{train}	N_{test}	Class labels
Computers	360000	500	720	20	450	50	$Desktop \setminus Laptop$
Worms	232200	258	900	11	181	77	$Mutant \setminus Non - mutant$

- *Computers Data Set*

This problem is related to the data recorded as a part of a government-sponsored study called Powering the Nation [31]. The aim was to collect behavioral data about how consumers use electricity in their houses to help reduce the country's carbon footprint. The data contains electricity readings from 251 households, sampled in two-minute intervals over a month. Hence, each observation contains a group of 720 measurements (24 hours of readings taken every 2 minutes). The two classes are either Desktop or Laptop. This way, without surveying one-by-one, hopefully, we can know how many users of desktop or laptop only from the records of their electricity uses.

- *Worms Data Set*

Caenorhabditis elegans is a special type of worm commonly used as a model organism in the study of genetics [31]. The movement of these worms is known to be a useful indicator for understanding behavioral genetics. [32] described a system for recording the traces of the worms' motions. They captured these traces by four scalars representing the amplitudes along each dimension based on four "eigenworms". The data reports 258 traces of worms converted into four eigenworm series. The eigenworm data are of the lengths from 17984 to 100674 (sampled at 30 Hz, so from 10 minutes to 1 hour) and in four dimensions (eigenworm 1 to 4). There are five classes: N2, goa-1, unc-1, unc-38, and un63, such that N2 is wildtype (i.e., normal) and the others are mutant strains. This data set is a two-class version: mutant vs non-mutant.

The classification is done for all the data sets using the proposed EPSVM and QPSVM methods. Here, different kernels including linear and nonlinear kernels have been used. The results are presented in Table 3.

Table 3. The classification accuracies of the real data sets

Data	Kernel	Accuracies Using Various Methods			
		EPSVM	QPSVM		
			$p = 0.25$	$p = 0.5$	$p = 0.75$
Worms	Linear	0.612	0.575	0.650	0.650
	Polynomial	0.700	0.700	0.675	0.775
	Radial	0.650	0.575	0.6500	0.650
	Sigmoid	0.700	0.575	0.6500	0.650
Computers	Linear	0.552	0.556	0.664	0.612
	Polynomial	0.664	0.668	0.664	0.612
	Radial	0.552	0.556	0.664	0.612
	Sigmoid	0.552	0.560	0.664	0.598

From Table 3, it is observed that for the two data sets:

Using the Polynomial kernel leads to more accuracy than other kernel functions for both of EPSVM and QPSVM (of any order of quantile) methods.

The accuracy of the QPSVM method is usually more than the accuracy of the EPSVM method. Moreover, the choice of the quantile's order used in the QPSVM method affect the accuracies. For example, in the Worms data set, for the Linear kernel, the QPSVM with $p = 0.5$ or $p = 0.75$

is more accurate than the EPSVM method, while for the Polynomial kernel the QPSVM with $p = 0.75$ is more accurate than the EPSVM method. These facts shows that the data are better classified if they are represented by the quantiles. In other words, assymetricity of the histogram-valued observations plays an important role in classifying the Worms data set.
In both data sets, the usage of some nonlinear Kernels improve the accuracies regardless of the choice of methods (EPSVM or QPSVM) and the choice of quantile's order. This means that the Worms and Computer data sets are more likely non-linear, i.e, linearly non-separable.

7. Conclusion

The problem of PSVM classification on histogram-valued data was studied in this paper. *The original PSVM formula includes probabilistic expression which is difficult to solve. The pdfs of histogram-valued data are in the form of step functions, hence solving the PSVM formulation directly will lead to dealing with a discontinuous and nonlinear optimization problem. Because of this complexity of the PSVM method, two more practical approaches were suggested to be the alternatives to obtain the solution of the PSVM problem, which were called as EPSVM and QPSVM.* In both approaches, the linear, nonlinear and least square cases were considered in details. From the experimental results in the case of binary classification, it was deduced that the proposed classification methods may usually lead to good accuracy in either the generated data or real data sets. It was observed that the QPSVM may be better than EPSVM method when the distribution of the data is asymmetric. Moreover, depending on the linear separability of the classes and the distributions of the data, the choices of kernel and the quantile's order used in the QPSVM method are two important elements affecting the accuracy of the classification. *It is worth to note that because single real-valued data and interval-valued data are subsets of histogram-valued data, this method can easily be applied to mixed data consisting single real-valued data, interval-valued data, and histogram-valued data*. Then, further research is needed for generalizing this method to be applied to modal-valued form of data, so that this method can be used to solve linear and nonlinear classification problems of various data types or their mixtures. For the extension of the proposed model, one may build many other machine learning methods such as regression using the similar method.

8. References

[1]. Esposito F., Malerba D., and I. (Eds.). S. V. 15: pp. Tamma V. (2000) (Section 8.3), "Dissimilarity Measures for Symbolic Objects," in *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*, 2000, vol. 15, pp. 165–185.

[2]. L. Billard and E. Diday, *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. West Sussex, UK: John Wiley and Sons, 2006.

[3]. N. Kejzar, S. Korenjak-Cerne, and V. B. V, "Clustering of Modal Valued Symbolic Data," *Adv. Data Anal. Classif.*, vol. 15, pp. 513–541, 2021.

[4]. R. B. Gurung, T. Lindgren, and H. Boström, "Learning Decision Trees from Histogram Data," in *Proceedings of the 2015 International Conference on Data Mining: DMIN 2015*, 2015, pp. 139–145.

[5]. R. B. Gurung, T. Lindgren, and H. Boström, "Learning random forest from histogram data using split specific axis rotation," *Int. J. Mach. Learn. Comput.*, vol. 8, no. 1, pp. 74–79, 2018.

[6]. A. Irpino and R. Verde, "Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance," *Adv. Data Anal. Classif.*, vol. 9, pp. 81–106, 2015.

[7]. S. Dias and P. Brito, "Linear Regression Model with Histogram-Valued Variables," *Stat. Anal. Data Min.*, vol. 8, no. 2, pp. 75–113, 2015.

[8]. P. Nagabhushan and R. Pradeep Kumar, "Histogram PCA," *Adv. Neural Networks – ISNN 2007. Lect. Notes Comput. Sci.*, vol. 4492, pp. 1012–1021, 2007.

[9]. C. Cortes and V. Vapnik, "Support Vector Networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.
[10]. E. Osuna, R. Freund, and F. Girosi, "Training Support Vector Machines: An Application to Face Detection," *IEEE Conf. Comput. Vis. Pattern Recognit.*, p. 130, 1997.
[11]. T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *Proceedings European Conference on Machine Learning*, 1998, p. 137.
[12]. D. J. Sebald and J. A. Bucklew, "Support Vector Machine Techniques for Nonlinear Equalization," *IEEE Trans. Signal Process.*, vol. 48, no. 11, pp. 3217–3226, 2000.
[13]. C. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten Digit Recognition: Bench-Marking of State-of-the-Art Techniques," *Pattern Recognit.*, vol. 36, pp. 2271–2285, 2003.
[14]. Lin C.F. and S. D. Wang, "Fuzzy Support Vector Machine," *IEEE Trans. Neural Networks*, vol. 13, pp. 464–471, 2002.
[15]. Huang H.P. and Y. H. Liu, "Fuzzy Support Vector Machines for Pattern Recognition and Data Mining," *Int. J. Fuzzy Syst.*, vol. 4, pp. 826–835, 2002.
[16]. Y. Chen and J. Z. Wang, "Support Vector Learning for Fuzzy Rule-Based Classification Systems," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 6, pp. 716–728, 2003.
[17]. Y. Q. Wang, S. Y. Wang, and K. K. Lai, "A New Fuzzy Support Vector Machine to Evaluate Credit Risk," *IEEE Trans. Fuzzy Syst.*, vol. 13, pp. 820–831, 2005.
[18]. J. H. Chiang and P. Y. Hao, "Support Vector Learning Mechanism for Fuzzy Rule-Based Modeling: A New Approach," *IEEE Trans. Fuzzy Syst.*, vol. 12, no. 1, pp. 1–12, 2004.
[19]. J. A. K. Suykens and J. Vandewalle, "Least Squares Support Vector Machines Classifiers," *Neural Process. Lett.*, vol. 9, pp. 293–300, 1999.
[20]. H. Li, J. Yang, G. Zhang, and B. Fan, "Probabilistic Support Vector Machines for Classification of Noise Affected Data," *Inf. Sci. (Ny).*, vol. 221, pp. 60–71, 2013.
[21]. M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of Second-Order Cone Programming," *Linear Algebr. Its Appl.*, vol. 284, pp. 193–228, 1998.
[22]. P. Sollich, "Bayesian Methods for Support Vector Machines: Evidence and Predictive Class Probabilities," *Mach. Learn.*, vol. 46, pp. 21–52, 2002.
[23]. Y. J. Lee and S. Y. Huang, "Reduced Support Vector Machines: A Statistical Theory," *IEEE Trans. Neural Networks*, vol. 18, pp. 1–13, 2007.
[24]. J. B. Gao, S. R. Gunn, C. J. Harris, and M. Brown, "A Probabilistic Framework for SVM Regression and Error Bar Estimation," *Mach. Learn.*, vol. 46, pp. 71–89, 2002.
[25]. P. Bosch, J. Lopez, H. Ramirez, and H. Robotham, "The Evidence Framework Applied to Support Vector Machines," *IEEE Trans. Neural Networks*, vol. 11, pp. 1162–1173, 2000.
[26]. W. Y. Liu, K. Yue, and M. H. Gao, "Constructing Probabilistic Graphical Model from Predicate Formulas for Fusing Logical and Probabilistic Knowledge," *Inf. Sci. (Ny).*, vol. 181, no. 18, pp. 3828–3845, 2011.
[27]. Y. Jinglin, H. X. Li, and H. Yong, "A Probabilistic SVM Based Decision System for Pain Diagnosis," *Expert Syst. Appl.*, vol. 38, pp. 9346–9351, 2011.
[28]. P. Bosch, J. Lopez, H. Ramirez, and H. Robotham, "Support Vector Machine Under Uncertainty: An Application For Hydroacoustic Classification of Fish Schools in Chile," *Expert Syst. Appl.*, vol. 40, pp. 4029–4034, 2013.
[29]. M. Abaszade and S. Effati, "Stochastic Support Vector Machine for Classifying and Regression of Random Variables," *Neural Process. Lett.*, vol. 48, pp. 1–29, 2018.
[30]. L. Billard and E. Diday, "From the statistics of data to the statistics of knowledge: Symbolic data analysis," *J. Am. Stat. Assoc.*, vol. 98, pp. 470–487, 2003.
[31]. A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data Min. Knowl. Discov.*, vol. 31, no. 3, pp. 606–660, 2017.
[32]. J. Kim and L. Billard, "Dissimilarity measures for histogram-valued observations", Communication and Statistics — Theory and Methods 42, pp 283-303.

[33]. Brown, A.E.X., Yemini, E.I., Grundy, L.J., Jucikas, T., Schafer, W.R., "A dictionary of behavioral motifs reveals clusters of genes affecting caenorhabditis elegans locomotion", 110, 791–796, 2013.
[34]. Chapelle, Olivier & Haffner, Patrick & Vapnik, Vladimir, "Support vector machines for histogram-based image classification", IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council, 10, 1055-64, 10.1109/72.788646, 1999.
[35]. I. Kang, C. Park, Y.J. Yoon, & Park, Changyi, S. Kwon, H. Choi, Hosik, "Classification of histogram-valued data with support histogram machines", Journal of Applied Statistics. 1-16. 10.1080/02664763.2021.1947996, 2021.
[36]. J. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines" (PDF). CiteSeerX 10.1.1.43.4376, 1998.
[37]. L. Bottou, C. Cortes, J.S. Denker, H. Druncker, I. Guyon, L. Jackel, Y. LeCun, U.A. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of classifier methods: a case study in hand-written digit recognition", Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 – Conference C: Signal Processing (Cat. No.94CH3440–5), pp. 77–82 vol.2., 1994.

Fathimah Al-Ma'shumah received Master degree in mathematics from Institut Teknologi Bandung, Bandung, Indonesia, 2015. She is now a Ph.D student in statistics at Ferdowsi University Mashhad. Her interests are optimization theory and machine learning.

Mostafa Razmkhah received the Ph.D. degree in statistics from Ferdowsi University of Mashhad, Mashhad, Iran, in 2008. He is an Associate Professor at Ferdowsi University of Mashhad. His research interests are statistical inferences, reliability theory, ordered data, and censored data.

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.