Twin support vector machine with local structural information for pattern classification

MAOXIANG CHU*, LIMING LIU, YONGHUI YANG*, AND RONGFEN GONG
School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan 114051, China
Corresponding authors: Moxiang Chu (chu52_2004@163.com) and Yonghui Yang (yangyh2636688@163.com)

This work was supported in part by Liaoning Province PhD Start-up Fund (201601291), Natural Science Foundation of Liaoning Province of China (20180550067), Liaoning Province Ministry of Education Scientific Study Project (2017LNQN11 and 2016TSPY13).

ABSTRACT Many versions of support vector machine with structural information exploit the useful prior knowledge to directly improve the algorithm’s generalization. The prior knowledge embodies the structure of data, but it cannot fully reflect the local nonlinear structure of data. In this paper, a twin support vector machine with local structural information (LSI-TSVM) is proposed. LSI-TSVM embeds the local within-class and between-class distribution information of data, which makes it contain not only the original global within-class clustering and between-class margin, but also the local within-class and between-class scatters. Furthermore, our LSI-TSVM is extended to a nonlinear version with a kernel trick. All experiments show that our LSI-TSVM is superior to the state-of-the-art algorithms in generalization performance.

INDEX TERMS pattern classification, structural twin support vector machine, local structural information, generalization performance.

I. INTRODUCTION

Standard support vector machine (SVM) proposed by Vapnik et al. [1] is an excellent machine learning method. SVM adopts the principle of structural risk minimization and obeys the idea of margin maximization. It builds an optimal classification hyperplane by solving a quadratic programming problem (QPP) [2] to distinguish the types of samples. In order to solve nonlinear classification problem, the kernel function is adopted. Because of the advantage of global optimality and powerful generalization, SVM has been used in many fields, such as pattern classification and regression analysis [3-5]. As time goes on, many developed versions of SVM are proposed, such as least squares support vector machine (LSSVM) [6], ν-SVM [7], support vector machine with pinball loss (Pin-SVM) [8], support vector data description (SVDD) [9], large margin distribution machine (LDM) [10] and fuzzy support vector machine (FSVM) [11].

Though SVM has many merits, it costs much time to solve an entire QPP. In 2007, twin support vector machine (TSVM) [12] settles that problem, and it becomes popular. The idea of TSVM is originated from the generalized eigenvalue proximal SVM (GEP SVM) [13], which is accord with SVM. It aims at building two nonparallel hyperplanes by solving two smaller QPPs, which improves four times in efficiency. Moreover, TSVM is very fit for crossed dataset and has been widely used in many fields [14-16]. In addition to application, its model has been extended and improved. Shao et al. [17] proposed twin bounded support vector machine (TBSVM) based on TSVM, which realizes structural risk minimization by introducing the regularization item into objective function. On the other hand, many scholars proposed many developed TSVM, such as least squares TSVM (LST SVM) [18], projection TSVM (PTSVM) [19], two-parametric-margin SVM (TPMSVM) [20], nonparallel support vector machine (NPSVM) [21] and best fitting hyperplane classifier (BFSC) [22].

In fact, traditional SVM and TSVM do not sufficiently exploit the prior samples distribution information. They only concern on the types of samples in separability, but they ignore the structure distribution of samples. In order to fully use the structural information of samples, Xue et al. [23] and Qi et al. [24] proposed the structural regularized SVM (SRSVM) and structural TSVM (S-TSVM) respectively. As applied in the structured large margin machine (SLMM), the Ward’s linkage clustering method is also used to investigate the underlying structure of data in SRSVM and S-TSVM. The experimental results show that these two algorithms have perfect generalization ability and improve classification accuracy compared with traditional SVM and TSVM. Inspired by the idea of SRSVM and S-TSVM, the structural information of samples is widely used to improve generalization ability for different version of SVM. For examples, Chen et al. [25] proposed structural nonparallel support vector machine (SNPSVM) by adding minimization
within-class structural information into NPSVM model. Zhang et al. [26] proposed fisher regularized support vector machine (Fisher-SVM) by introducing minimization within-class scatter. Peng et al. [27] proposed structural regularized PTVM (SRPTSVM) by inserting two classes of structural information into PTVM. Pan et al. [28] proposed K-nearest neighbor based structural twin support vector machine (KNN-STSVM), which strengthens the structural information by applying the inter-class KNN method. An et al. [29] proposed a novel FSVM based on within-class scatter (WCS-FSVM), which incorporates minimum within-class scatter in fisher discriminant analysis (FDA) into FSVM. Luo et al. [30] proposed a fuzzy quadratic surface SVM model (FQSSVM) using the concept of FDA.

At present, the existing structural SVMs have two types:
(1) One uses the concept of FDA [31], such as WCS-SVM, Fisher-SVM and FQSSVM.
(2) The other one uses the Ward’s linkage clustering method, such as SRSVM, S-TSVM, KNN-STSVM, SNPSVM and SRPTSVM.

Nevertheless, we find two drawbacks of these structural SVMs:
(1) Some structural SVMs using the concept of FDA only can excavate the global linear structure of data but the local nonlinear structure.
(2) The other structural SVMs using the Ward’s linkage clustering method can exploit the local nonlinear structure, but they only consider the local within-class structure not including the local between-class margin.

Exactly, the neighborhood margin FDA (NMFDA) proposed by Wei et al. [32] tries to minimize the local within-class scatter and maximize the local between-class scatter. It aims at enlarging the boundary margin between the farthest within-class neighbor and the nearest between-class neighbor of sample. It has been proved theoretically and experimentally that the boundary is the real boundary of two classes of samples. It reflects the local between-class margin. Moreover, NMFDA uses KNN to find the local nonlinear structural information of data.

Inspired by the studies above, a twin support vector machine with local structural information (LSI-TSVM) is proposed in this paper. Similar to the existing structural SVMs, LSI-TSVM is built by introducing the structural information of data into the objective function of TSVM. As WCS-FSVM, Fisher-SVM and FQSSVM, LSI-TSVM contains the original global structural information. Differently, LSI-TSVM also contains the local structural information. As SRSVM, S-TSVM, KNN-STSVM, SNPSVM and SRPTSVM, LSI-TSVM captures the local information of each class and global margin. In addition, they also can deal with the nonlinear data with kernel trick. Differently, LSI-TSVM also contains the local between-class scatter. The comparative experiments on all datasets indicate our LSI-TSVM has better generalization performance.

The remaining parts of the paper are organized as follows. Section 2 mainly introduces the related classification models including TSVM and S-TSVM. Section 3 describes the extraction method of local structural information. And LSI-TSVM models are presented for linear and nonlinear conditions. Section 4 analyzes the characteristics of LSI-TSVM. Section 5 displays the experiment results and analysis. Conclusions are given in the last section.

II. RELATED WORK

Assume that $X = [(X^+)^T (X^-)^T]^T$ is a training dataset. $X^+ = [x^+_1, x^+_2, ..., x^+_m]^T$ and $X^- = [x^-_1, x^-_2, ..., x^-_m]^T$ are datasets for class $+1$ and class $-1$ respectively, where $x^+_i \in \mathbb{R}^{n+}$, $x^-_i \in \mathbb{R}^{n+}$, $m^+$ and $m^-$ are the number of samples, where $m^++m^- = m$.

A. TWIN SUPPORT VECTOR MACHINE

TSVM searches for two optimal nonparallel hyperplanes:

$$x^+_i w^+ + b^+ = 0,$$
$$x^-_i w^- + b^- = 0,$$

and requires each hyperplane to be close to one class of samples and to be far from the other class of samples as much as possible. TSVM can be described as the following two QPPs:

$$\min_{w^+, b^+} \frac{1}{2} \|X^+ w^+ + e^+ b^+\|^2 + c^+(e^-)^T \xi^-,$$
$$\text{s.t. } X^+ w^+ + e^+ b^+ \leq -e^- + \xi^-,$$
$$\xi^- \geq 0,$$

(2)

$$\min_{w^-, b^-} \frac{1}{2} \|X^- w^- + e^- b^-\|^2 + c^-(e^+)^T \xi^+,$$
$$\text{s.t. } X^- w^- + e^- b^- \geq e^+ - \xi^+,$$
$$\xi^+ \geq 0,$$

(3)

where $c^+$ and $c^-$ are penalty factors. $e^+$ and $e^-$ are column vectors of ones. According to the Lagrange function and KKT conditions, the dual problems of QPPs (2) and (3) can be obtained as follows:

$$\min_{\alpha^+} \frac{1}{2} (\alpha^+)^T H(G^T G)^{-1} H^T \alpha^- -(e^+)^T \alpha^+,$$
$$0 \leq \alpha^+ \leq c^+ e^+,$$

(4)

$$\min_{\alpha^-} \frac{1}{2} (\alpha^-)^T G(H^T H)^{-1} G^T \alpha^+ -(e^-)^T \alpha^-,$$
$$0 \leq \alpha^- \leq c^- e^-,$$

(5)

where $G = [X^+, e^+]$, $H = [X^-, e^-]$, $\alpha^+$ and $\alpha^-$ are Lagrangian multipliers. By solving (4) and (5), the optimal solutions of $(\alpha^+)^T$ and $(\alpha^-)^T$ can be obtained. Then, according to the following augmented vectors:
the optimal classification hyperplane can be determined. The class of a testing sample $x_i$ can be determined with the following function:

$$f(x_i) = \arg\min_{+,-} \{ |x_i^T w^* + b^*|, |x_i^T w^- + b^-| \}. \quad (7)$$

**B. STRUCTURAL TWIN SUPPORT VECTOR MACHINE**

S-TSVM digs the structural information of data with clustering method [24]. Then, the structural information for samples in target class embeds into the objective function in TBSVM. The model of S-TSVM can be described with the following optimal problems:

$$\min_{w^+,b^+} \frac{1}{2} \|X^T w^* + e^* b^*\|^2 + c_1 (e^*)^T \xi^*$$

$$+ \frac{1}{2} c_6 (\|w^+\|^2 + (b^*)^2) + \frac{1}{2} c_5 (w^*)^T \Sigma w^+ \quad \text{s.t.} - (X^T w^* + e^* b^*) + \xi^* \succeq e^*, \xi^* \succeq 0, \quad (8)$$

$$\min_{w^-,b^-} \frac{1}{2} \|X^T w^- + e^* b^-\|^2 + c_4 (e^*)^T \xi^*$$

$$+ \frac{1}{2} c_6 (\|w^-\|^2 + (b^-)^2) + \frac{1}{2} c_5 (w^-)^T \Sigma w^- \quad \text{s.t.} (X^T w^- + e^* b^-) + \xi^* \succeq - e^*, \xi^* \succeq 0, \quad (9)$$

where $c_i (i = 1, \ldots, 6)$ are penalty parameters. $\Sigma_+$ and $\Sigma_-$ are covariances of samples. Suppose there are two groups in class P and class N. If $P = P_1 \cup P_2 \cup \cdots \cup P_p$ and $N = N_1 \cup N_2 \cup \cdots \cup N_N$, then $\Sigma_+ = \sum_{i=N_1 + \cdots + N_p}$ and $\Sigma_- = \sum_{i=N_1 + \cdots + N_N}$. The dual QPPs of (8) and (9) can be represented as:

$$\min_{a} \left( \alpha^T H (G^T G + c_2 I + c_6 J)^{-1} H^T \alpha - (e^*)^T \alpha \right)$$

$$\text{s.t.} \ 0 \leq \alpha^* \leq c_6 e^*, \quad (10)$$

$$\min_{a} \left( \alpha^T G (H^T H + c_1 I + c_4 F)^{-1} G^T \alpha - (e^*)^T \alpha \right)$$

$$\text{s.t.} \ 0 \leq \alpha^* \leq c_4 e^*, \quad (11)$$

where $J = \begin{bmatrix} \Sigma_+ & 0 \\ 0 & 0 \end{bmatrix}$ and $F = \begin{bmatrix} \Sigma_- & 0 \\ 0 & 0 \end{bmatrix}$.

By solving (10) and (11), the optimal solutions of $(\alpha^*)^+$ and $(\alpha^*)^-$ can be obtained. Then, according to the following augmented vectors:

$$\begin{bmatrix} w^+ \\ b^+ \end{bmatrix} = -(G^T G + c_2 I + c_6 J)^{-1} H^T \alpha^+,$$

$$\begin{bmatrix} w^- \\ b^- \end{bmatrix} = (H^T H + c_1 I + c_4 F)^{-1} G^T \alpha^-,$$  

the optimal hyperplane (1) of S-TSVM can be determined. A testing sample $x$ belonging to class +1 or -1 can be determined with the decision function (7).

**III. TWIN SUPPORT VECTOR MACHINE WITH LOCAL STRUCTURAL INFORMATION**

In this section, we propose a novel nonparallel classifier termed as LSI-TSVM. LSI-TSVM first adopts some techniques to extract the local structural information. Then it introduces some terms to limit the local within-class structural information and maximize the local between-class structural information into the objective function. Finally, the linear and nonlinear models of LSI-TSVM are formulated.

**A. LOCAL STRUCTURAL INFORMATION**

It is well known that the separability principle of FDA [31] is to minimize the within-class structural information and maximize the between-class structural information. More precisely, the structural information refers to the scatter of data. As mentioned before, the scatter has the assumption of global structure. To address the limitation of the global structure, some FDA algorithms using local structure of data [33-35] are proposed. The NMFDA in [32] is a simple and efficient algorithm. Following the NMFDA algorithm, we consider the following local scatter of data.

Here the case of two classes is only shown. For sample $x^*_i$ in class +1, its within-class KNNs set $N^+(x^*_i)$ and between-class KNNs set $N^-(x^*_i)$ are defined as:

$$N^+(x^*_i) = \{ x^*_i | t_i \in \{1,2,\ldots,m^+\}, k=1,2,\ldots,K \},$$

$$N^-(x^*_i) = \{ x^*_i | t_i \in \{1,2,\ldots,m^+\}, k=1,2,\ldots,K \}, \quad (13)$$

where $K$ is determined by user. By calculating the euclidean distances between $x^*_i$ and samples in two sets, the farthest within-class neighbor sample and the nearest between-class neighbor sample are determined:

$$x^*_m = \arg \max_{x^*_i} \{ \| x^*_i - x^*_j \|^2 \},$$

$$x^*_o = \arg \max_{x^*_i} \{ \| x^*_i - x^*_j \|^2 \}. \quad (14)$$

Based on $x^*_w$ and $x^*_b$, the local within-class scatter and between-class scatter can be obtained:

$$\bar{S}^w = w^T S^w w^T,$$

$$\bar{S}^b = w^T S^b w^T,$$  

where $S^w$ and $S^b$ are the local within-class and between-class covariance matrices. They are computed as:
\[ S_u^\alpha = \sum_{i=1}^{m} (x_i^\alpha - x_{m}^\alpha)(x_i^\alpha - x_{m}^\alpha)^T, \]
\[ S_b^\alpha = \sum_{i=1}^{m} (x_i^\alpha - x_{m}^\alpha)(x_i^\alpha - x_{m}^\alpha)^T. \]  

(16)

\[ \bar{S}_u^\alpha \text{ selects the farthest sample } x_i^\alpha \text{ from } N^\alpha(x_j^\alpha) \text{ to be the local within-class neighbor boundary. And it describes the local within-class neighbor distribution with the covariance } S_u^\alpha. \text{ So, } \bar{S}_u^\alpha \text{ reflects the local within-class structural information. Similarly, } \bar{S}_b^\alpha \text{ selects the nearest sample } x_j^\alpha \text{ from } N^\alpha(x_j^\alpha) \text{ to be the local between-class neighbor boundary. And the covariance } S_b^\alpha \text{ is used to describe the local between-class neighbor margin. So, } \bar{S}_b^\alpha \text{ reflects the local between-class structural information.} \]

Similarly, for the sample \( x_i^\alpha \) in class -1, the local within-class scatter \( \bar{S}_u^\alpha \) and local between-class scatter \( \bar{S}_b^\alpha \) can be determined as:

\[ \bar{S}_u^\alpha = w^T S_u^\alpha w, \]
\[ \bar{S}_b^\alpha = w^T S_b^\alpha w, \]  

(17)

where

\[ N^\alpha(x_j^\alpha) = \{x_i^\alpha | x_i^\alpha \in [1,2,\ldots,m^\alpha], k = 1,2,\ldots,K\}, \]

\[ x_{m}^\alpha = \arg \max_{x_i^\alpha} \|x_i^\alpha - x_j^\alpha\|^2 | k = 1,2,\ldots,K, \]  

(18)

\[ S_u^\alpha = \sum_{j=1}^{m} (x_j^\alpha - x_{m}^\alpha)(x_j^\alpha - x_{m}^\alpha)^T, \]

\[ N^\alpha(x_j^\alpha) = \{x_i^\alpha t_k \in [1,2,\ldots,m^\alpha], k = 1,2,\ldots,K\}, \]

\[ x_{m}^\alpha = \arg \max_{x_i^\alpha} \|x_i^\alpha - x_j^\alpha\|^2 | k = 1,2,\ldots,K, \]  

(19)

\[ S_b^\alpha = \sum_{j=1}^{m} (x_j^\alpha - x_{m}^\alpha)(x_j^\alpha - x_{m}^\alpha)^T. \]

B. LINEAR LSI-TSVM

Following the S-TSVM classifier, our LSI-TSVM introduces the local structural information for two classes by using the concept of NMFD. The linear LSI-TSVM optimizes the following two QPPs:

\[ \min_{w,b} \frac{1}{2} \|X^+ w^+ + e^+ b^+\|^2 + \frac{1}{2} v^+ \gamma^+ S_u^\alpha w^+ \]
\[ + \frac{1}{2} v^+ \gamma^+ (w^+)^T (I^- \gamma^+ S_b^\alpha) w^+ + (b^+)^T c^+ (e^+)^T \xi^- \]  

s.t. \( X^+ w^+ + e^+ b^+ \leq -e^- + \xi^- \), \( \xi^- \geq 0, \) \( \gamma^+ \) and \( \gamma^- \) are trade-off parameters, \( e^+ \) and \( e^- \) are vectors of ones, \( I^+ \) and \( I^- \) are two identity matrices with appropriate dimensions. Consider the illustration of the QPP (20). The second term minimizes the scatter of samples in class +1, which makes LSI-TSVM pay more attention to the local within-class structural information. It is more reasonable for real-world problems. The third term contains three sub-parts: \( 1/2(v^+_m (w^+)^T I^+ w^+), \)
\[ -1/2(v^-_m \gamma^- (w^-)^T S_b^\alpha w^-) \) and \( 1/2(v^-_m (b^-)^2) \). Minimizing the first sub-part is equivalent to maximizing the one side margin with respect to the hyperplane \( x^+ w^+ + b^+ = 0 \) \( [17]. \)

It embodies the maximization of global margin. The second sub-part aims to maximize the local between-class scatter, which makes LSI-TSVM pay attention to the local between-class structural information. The third sub-part is a regularization term, which avoids the inverse of matrix with possible ill conditioning. The other terms and constraints are the same as TSVM. It would be specially mentioned that the first term also minimizes the global within-class clustering information. For QPP (21), the similar conclusions can be given.

To solve (20), the Lagrange function is given by

\[ L = \frac{1}{2} \|X^+ w^+ + e^+ b^+\|^2 + \frac{1}{2} v^+ \gamma^+ S_u^\alpha w^+ \]
\[ + \frac{1}{2} v^+ \gamma^+ (w^+)^T (I^- \gamma^+ S_b^\alpha) w^+ + (b^+)^T c^+ (e^+)^T \xi^- \]  

(22)

where \( a^+ = [a^+_1,a^+_2,\ldots,a^+_m]^T \) and \( b^+ = [b^+_1,b^+_2,\ldots,b^+_m]^T \) are Lagrange vectors. According to the Karush-Kuhn-Tucker optimality condition:

\[ \frac{\partial L}{\partial w^+} = (X^+ w^+ + e^+ b^+) + \frac{1}{2} v^+ \gamma^+ S_u^\alpha w^+ + \frac{1}{2} v^+ \gamma^+ (w^+)^T (I^- \gamma^+ S_b^\alpha) w^+ \]
\[ + \frac{1}{2} v^+ \gamma^+ (b^-)^2 + c^+ (e^+)^T \xi^- \]
\[ \text{s.t. } X^+ w^+ + e^+ b^+ \leq -e^- + \xi^- + c^+ (e^+) \]  

(23)

\[ \frac{\partial L}{\partial b^+} = (e^+)^T (X^+ w^+ + e^+ b^+) + v^+ (b^-)^2 + (c^+) \xi^- = 0, \]

(24)

the augmented vector \( \begin{bmatrix} w^+ \\ b^+ \end{bmatrix} \) is obtained as follows:

\[ \begin{bmatrix} w^+ \\ b^+ \end{bmatrix} = -(G^T G + P)^{-1} H^T a^+. \]

(26)
\[ G = \{X^+, e^+\}, \]
\[ H = \{X^-, e^-\}, \]
\[ P = \begin{bmatrix} v^+_mS^+_e + v^-_m(I^+ - \gamma^+S^+_e) & 0 \\ 0 & 0 \end{bmatrix}. \]  
(27)

According to (22-26), the dual QPP of (20) is given by
\[
\text{max} -\frac{1}{2} (\alpha^+)^T H(G^T G + P)^{-1} H^T \alpha^- + (e^-)^T \alpha^- \\
s.t. \quad 0 \leq \alpha^- \leq c^- e^-.
\]  
(28)

Once the solution of \( \alpha^- \) is found, the vector \([ (w^-)^T \ b^- ]^T \) will be determined by (26). Similarly, the dual problem of (21) is
\[
\text{max} -\frac{1}{2} (\alpha^*)^T G(H^T G + Q)^{-1} G^T \alpha^* + (e^*)^T \alpha^* \\
s.t. \quad 0 \leq \alpha^* \leq c^* e^* ,
\]  
where
\[
Q = \begin{bmatrix} v^+_mS^+_e + v^-_m(I^+ - \gamma^+S^+_e) & 0 \\ 0 & 0 \end{bmatrix},
\]  
(30)

and the augmented vector \([ (w^-)^T \ b^- ]^T \) is determined as
\[
\begin{bmatrix} w^- \\ b^- \end{bmatrix} = (H^T H + Q)^{-1} G^T \alpha^* .
\]  
(31)

According to (26) and (31), two decision hyperplanes can be given by (1).

In the prediction stage, a testing sample \( x \) is assigned to class 1+ or class -1, depending on which of the two decision hyperplanes is it closer to. So, the decision function of LSI-TSVM is also expressed as (7).

**C. NONLINEAR LSI-TSVM**

Similar to the linear case, the two decision hyperplanes in high dimensional mapping space can be written as
\[
\phi^T(x_i)w^*{\omega}^* + b^* = 0,
\]
\[
\phi^T(x_i)w^*{-\omega}^* + b^* = 0,
\]  
(32)

where \( \phi(\cdot) \) is a nonlinear mapping function. According to the reproducing kernel Hilbert space theory [36], the following equations are given by
\[
\hat{w}^*{\omega}^* = \sum_{i=1}^{m^+} \hat{w}^+_i \phi(x^+_i) + \sum_{j=1}^{n^-} \hat{w}^-_j \phi(x^-_j) = \phi(X^+) \hat{w}^*{\omega}^*,
\]
\[
\hat{w}^*-\omega^* = \sum_{i=1}^{m^+} \hat{w}^+_-i \phi(x^-_i) + \sum_{j=1}^{n^-} \hat{w}^-_-j \phi(x^+_j) = \phi(X^-) \hat{w}^*-\omega^*,
\]  
(33)

where \( \hat{w}^*{\omega}^*, \hat{w}^*-\omega^* \in \mathbb{R}^{(m^-+m^+)\times 1} \). So, the two decision hyperplanes in (32) can be rewritten as
\[
\psi(x^T, X^T) \hat{w}^* + \hat{b}^* = 0,
\]
\[
\psi(x^-T, X^-T) \hat{w}^- + \hat{b}^- = 0,
\]  
where \( \psi(\cdot) \) is a chosen kernel function and \( \psi(x^T, X^T) = \phi(x^T) \cdot \phi(X^T) \). Meanwhile, the local within-class scatterers \( (\hat{w}^*)^T \psi^T(X^T) S_{\omega}^* \psi(X^T) \hat{w}^* \) and the local between-class scatterers \( (\hat{w}^-)^T \psi^T(X^-) S_{\omega^-} \psi(X^-) \hat{w}^- \) for the nonlinear LSI-TSVM are determined, where \( S_{\omega}^* \) and \( S_{\omega^-} \) are the scatter matrices in the high dimensional mapping space. The farthest within-class neighbor sample and the nearest between-class neighbor sample are defined in the nonlinear case. Take \( S_{\omega}^* \) for example, the local within-class scatter can be expressed as
\[
(\hat{w}^*)^T \psi^T(X^T) S_{\omega}^* \psi(X^T) \hat{w}^*
\]
\[
= (\hat{w}^*)^T \psi^T(X^T) \sum_{i=1}^{n^-} (\phi(x^-_i) - \phi(x^-_i)) (\phi(x^-_i) - \phi(x^-_i))^T \psi(X^-) \hat{w}^- + (\hat{b}^-)^T
\]
\[
+ c^+ (e^-)^T \xi^-.
\]
\[
\text{s.t.} \quad \psi(x^+, X^+) \hat{w}^* + e^+ \hat{b}^* \leq -e^- + \xi^+,
\]
\[
\xi^+ \geq 0.
\]  
(34)

Following the linear case, the two primal QPPs of nonlinear LSI-TSVM are respectively described as follows:
\[
\min_{\hat{w}^*, \hat{b}^-} \frac{1}{2} \| \psi(X^+, X^T) \hat{w}^* + e^+ \hat{b}^* \|^2
\]
\[
+ \frac{1}{2} v^-_m((\hat{w}^-)^T \psi^T(X^-) (I^- - \gamma^- S^- e^-) \psi(X^-) \hat{w}^- + (\hat{b}^-)^T)
\]
\[
+ c^+ (e^-)^T \xi^-.
\]
\[
\text{s.t.} \quad \psi(x^+, X^+) \hat{w}^* + e^+ \hat{b}^* \geq e^- + \xi^-,
\]
\[
\xi^- \geq 0.
\]  
(35)

Similarly, the dual QPPs of (36) and (37) are as follows:
\[
\max_{\alpha^-} -\frac{1}{2} (\alpha^+)^T H^T((G^T)^T G^T + P^T)^{-1} (H^T)^T \alpha^- + (e^-)^T \alpha^- \\
s.t. \quad 0 \leq \alpha^- \leq c^- e^-.
\]  
(36)

\[
\max_{\alpha^-} -\frac{1}{2} (\alpha^*)^T G^T((H^T)^T H^T + Q^T)^{-1} (G^T)^T \alpha^- + (e^*)^T \alpha^* \\
s.t. \quad 0 \leq \alpha^- \leq c^- e^-,
\]  
(37)

Whereas, the dual QPPs of (36) and (37) are as follows:
\[
\max_{\alpha^-} -\frac{1}{2} (\alpha^+)^T H^T((G^T)^T G^T + P^T)^{-1} (H^T)^T \alpha^- + (e^-)^T \alpha^- \\
s.t. \quad 0 \leq \alpha^- \leq c^- e^-.
\]  
(38)

\[
\max_{\alpha^-} -\frac{1}{2} (\alpha^*)^T G^T((H^T)^T H^T + Q^T)^{-1} (G^T)^T \alpha^- + (e^*)^T \alpha^* \\
s.t. \quad 0 \leq \alpha^- \leq c^- e^-,
\]  
(39)
where
\[
G^\circ = [\psi(X^\circ, X^T) \cdot e^\circ],
\]
\[
H^\circ = [\psi(X^\circ, X^T) \cdot e^\circ],
\]
\[
\mathbf{p}^\circ = \begin{bmatrix}
  v_1 \psi(X^T) S_{\mu, e} \psi(X^T) + v_1 \psi(X^T) (I - \gamma S_{\mu, e} \psi(X^T)) & 0 \\
  0 & v_2
\end{bmatrix},
\]
\[
Q^\circ = \begin{bmatrix}
  v_1 \psi(X^T) S_{\mu, e} \psi(X^T) + v_1 \psi(X^T) (I - \gamma S_{\mu, e} \psi(X^T)) & 0 \\
  0 & v_2
\end{bmatrix}.
\]
\[\text{and the augmented vectors } [(\mathbf{w}_1^\circ)^\top \mathbf{b}_1^\circ]^\top \text{ and } [(\mathbf{w}_2^\circ)^\top \mathbf{b}_2^\circ]^\top \text{ are given by}
\]
\[
\mathbf{w}_1^\circ = -(G^\circ)^\top G^\circ + P^\circ)^{-1} (H^\circ)^\top a^-, 
\]
\[
\mathbf{w}_2^\circ = (H^\circ)^\top H^\circ + Q^\circ)^{-1} (G^\circ)^\top a^+.
\]
\[
\text{The decision function of LSI-TSVM is determined as}
\]
\[
f(x_i) = \arg \min_{x_i \in \mathbb{R}^d} \left\{ |\psi(x_i, X^\circ) \mathbf{w}^\circ + \mathbf{b}^\circ|, |\psi(x_i, X^\circ) \mathbf{w}^\circ + \mathbf{b}^\circ| \right\}.
\]

IV. ANALYSIS OF ALGORITHM

A. THE THEORETICAL FRAMEWORK OF LSI-TSVM

Recently, Yeung [37] described a system that applying the operable measures, the homogeneous space and the structured degree to classify the learning models. The theory emphasizes the importance of sample structure and obtains many new algorithms, such as SRSVM, WCS-FSVM, S-TSVM, and KNN-STSVM. Just like these algorithms, our LSI-TSVM also captures the structural information of training samples.

Besides, LSI-TSVM captures the local structure distribution by using the concept of NMFDA. The local structure distribution is described through the farthest homogeneous sample and the nearest heterogeneous sample in neighbor set of every sample. The margin of two boundaries between them should be the real margin. It is more reasonable that our LSI-TSVM tries to enlarge the margin. So, our LSI-TSVM can more fully exploit these prior knowledges to improve the model’s generalization performance.

B. THE COMPUTATIONAL COMPLEXITY OF LSI-TSVM

The computational complexity of traditional SVM is $O(m^2)$. As for the TSVM, it costs around $O(m^3/4)$ under the assumption that the patterns in two classes are approximately equal. Based on SVM, the structure information and the fuzzy technique are introduced into WCS-FSVM. So, WCS-FSVM costs more time than SVM. As for S-TSVM, the time complexity for the clustering step is $O((|m|^3 + |m|^2)d)$. Hence, the optimization of S-TSVM costs around $O((|m|^3 + |m|^2)d + m^3/4)$. Similarly, the computational complexity of KNN-STSVM is around $O(N^2\log(N) + (|m|^3 + |m|^2)d + m^3/4)$. However, in reality, the authentic computational complexity of KNN-STSVM is less than $O((|m|^3 + |m|^2)d + m^3/4)$. As for the new proposed method LSI-TSVM, compared with S-TSVM, it needs to solve and optimize more structural information terms, so it is slower.

C. RELATIONSHIP WITH TSVM AND TBSVM

It can be easily seen that LSI-TSVM has a similar structure of TSVM by comparing (20) and (2). And LSI-TSVM degenerates to TSVM when parameters are chosen appropriately. This means TSVM is a special case of LSI-TSVM. Suppose $v_1^\circ = 0$ and $v_2^\circ = 0$, the primal QPP (20) of linear LSI-TSVM becomes
\[
\min_{w, b} \frac{1}{2} \|X^\circ w^\circ + e^\circ b^\circ\|^2 + \frac{1}{2} v_1^\circ \|w^\circ\|^2 + (b^\circ)^2
\]
\[
+ c(e^\circ)^\top \xi^\circ
\]
\[
\text{s.t. } X^\circ w^\circ + e^\circ b^\circ \leq -e^\circ + \xi^\circ,
\]
\[
\xi^\circ \geq 0.
\]

The degenerated (43) is one of the primal QPPs of linear TBSVM in [17]. This also means TBSVM is a special case of LSI-TSVM. So, LSI-TSVM inherits all the advantages of TSVM and TBSVM. Meanwhile LSI-TSVM pays attention to the structural information of training samples. This implies LSI-TSVM is better than traditional SVM, TSVM and TBSVM in generalized capability.

D. RELATIONSHIP WITH KNN-STSVM AND S-TSVM

S-TSVM exploits the within-class structural information by using the Ward’s linkage clustering method and directly embeds the structural information into the objective function of TBSVM. Different from S-TSVM, KNN-STSVM not only incorporates the distribution information by using the clustering method, but also gives different weight to each sample of different classes by using KNN method. With regard to LSI-TSVM, it not only contains the original global within-class clustering information and between-class margin, but also embeds the local within-class scatter and between-class scatter by using the NMFDA method.

E. RELATIONSHIP WITH SRSVM AND WCS-FSVM

The same as SRSVM and WCS-FSVM, LSI-TSVM also captures the structural information of data. Similar to SRSVM and WCS-FSVM, LSI-TSVM directly embeds the local structural information into the TSVM objective function. Differently, our LSI-TSVM captures the within-class structural information by using the concept of NMFDA rather than using some clustering technology into SRSVM. LSI-TSVM seems to be more reasonable. Compared with WCS-FSVM, LSI-TSVM contains not only the global structural information, but also the local within-class scatter and the local between-class scatter. And LSI-TSVM can solve the nonlinear problem. In addition, LSI-TSVM is
proposed inspired by the success of T SVM methods. So, 

LSI-T SVM inherits all the advantages of T SVM.

V. EXPERIMENTS

In this section, in order to demonstrate the superiority of our proposed method, LSI-T SVM, S-T SVM, KNN-STSVM, WCS-F SVM, SRSVM, T SVM and SVM are tested and compared. All algorithms are implemented by using MATLAB (7.11.0) R2010b environment on a PC with an Intel I5 processor (3.0GHz) and 8GB RAM.

These algorithms are tested on synthetic dataset, image datasets and UCI datasets respectively. And the optimal performance of all algorithms is obtained with five-fold cross validation and grid searching. In the experiments, two conditions are considered respectively: linear kernel

\[ k(x_i, x_j) = x_i \cdot x_j \]

and Gaussian kernel

\[ k(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / \sigma^2) \].

As for the problem of selecting parameters, standard 5-fold cross-validation technique is employed. The optimal value of regularization parameters and penalty parameters are chosen from \( \{2^i \mid i = -7, -6, \ldots, 7\} \), and the kernel parameter \( \sigma \) is chosen from \( \{2^i \mid i = -4, 0, \ldots, 6\} \) by grid searching. In LSI-T SVM and KNN-STSVM, the optimal \( K \) is chosen from \( \{7, 8, 9, 10, 20\} \). Once these parameters are determined, the final decision function can be obtained. For brevity’s sake, in LSI-T SVM, \( \gamma^+ = \gamma^- \), \( v^+ = v^- \), \( \gamma^+ = \gamma^- \) and \( c^+ = c^- \). These simplified treatments will greatly reduce the training time on large-scale dataset.

A. TOY DATASET

In order to geometrically illustrate the effectiveness of the novel algorithm, our LSI-T SVM is compared with S-T SVM and T SVM on a two-dimensional dataset. In the toy dataset, the samples in positive class or negative class are randomly generated under two Gaussian distribution. For the positive (negative) class, there are 50 samples satisfying Gaussian distribution 1 (3) and 50 samples satisfying Gaussian distribution 2 (4). The corresponding attributes of toy dataset are shown in Table 1.

All the samples are regarded as training dataset and testing dataset. And the corresponding decision hyperplanes are given in Fig. 1, where “*” and “.” represent the samples in positive class and negative class respectively. It can be seen that the samples in positive class appear vertical distribution and the samples in negative class appear horizontal distribution. Under the condition, the structural information is very important for classifier. The following conclusions can be drawn from Fig. 1:

1. T SVM ignores the structural information of samples. It only concerns the separability of samples, which makes classifier dissatisfy the tendency of data distribution.

2. The Fisher-T SVM is obtained by incorporating Fisher regularization into T SVM. Fisher-T SVM tries to minimize the global within-class scatter and maximize the global classification margin. Fisher-T SVM nearly satisfies the rule of Fisher, which uses the structural information of samples to fulfill the classification task. So, Fisher-T SVM obtains better statistical separability than T SVM.

3. The within-class structural information is added onto the objective function of T SVM, which can make the tradeoff between within-class structural information and between-class discriminant information. And T SVM also can capture the local information of each class. So, T SVM is better than Fisher-T SVM and T SVM.

4. The local between-class scatter is embedded into the objective function of LSI-T SVM, which really reflects the aggregation of within-class samples and the separability of between-class samples. So, compared with T SVM, LSI-T SVM is more reasonable and has better generalization performance.

Table 1: Attributes of toy datasets.

<table>
<thead>
<tr>
<th>Class</th>
<th>Distribution</th>
<th>Probability</th>
<th>Mean</th>
<th>Covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive class</td>
<td>Gaussian distribution 1</td>
<td>1/2</td>
<td>[0.1]</td>
<td>[0.002, 0.0, 0.06]</td>
</tr>
<tr>
<td></td>
<td>Gaussian distribution 2</td>
<td>1/2</td>
<td>[0.5, 0.7]</td>
<td>[0.002, 0.0, 0.2]</td>
</tr>
<tr>
<td>Negative class</td>
<td>Gaussian distribution 3</td>
<td>1/2</td>
<td>[-0.5, -0.4]</td>
<td>[0.1, 0.0, 0.01]</td>
</tr>
<tr>
<td></td>
<td>Gaussian distribution 4</td>
<td>1/2</td>
<td>[1.0]</td>
<td>[0.1, 0.0, 0.01]</td>
</tr>
</tbody>
</table>

(a) TSVM

(b) Fisher-T SVM

M. Chu: TSVM with Local Structural Information for Pattern Classification

FIGURE 1. The classification results on toy dataset.

B. IMAGE DATASETS
In this subsection, our LSI-TSVM is compared with the other algorithms on two datasets. Because only the performance of algorithms is compared, the experiment is executed on the original pixel features.

(1) MNIST dataset [38]
There are 60000 training images and 10000 testing images in MNIST dataset. And ten types of handwriting digits represented as 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9 are contained in the dataset. All images are adjusted to gray level images with the size of 28×28. For brevity, four binary classification models are built with one-against-one method. And ‘0’ and ‘6’, ‘1’ and ‘2’, ‘3’ and ‘5’, ‘7’ and ‘9’ are taken as their corresponding labels. Each classification model is trained by using five datasets with the different number of samples: 20, 50, 80, 110 and 140. All testing images are tested and the testing results are shown in Fig. 2.

(2) PASCAL 2012 dataset [39]
This dataset is composed of 17125 images including twenty different categories. And ten types of them are chosen for experiments, such as cat, motorbike, cow, dog, car, bus, bicycle, sheep, people and horse. All these images are firstly converted into greyscale images which have 4000 pixels and the same length-width ratio with the original images. 800, 1000, 1200, 1400 and 1600 greyscale images are chosen as training samples respectively, and the others are testing samples. Each sample is represented as a 4000×1 vector. For convenience, ten binary classification models are built with one-against-rest method. The final classification performance is determined with the mean of ten accuracies. Because the classification accuracy in nonlinear space is superior to that in linear space, all algorithms are testified in nonlinear conditions and the results are shown in Fig. 3.
The reason is that the dataset has a regular structure distribution. Moreover, it can be easily seen that the accuracies of five algorithms grow as the training samples growing, which shows the number of samples will affect the digging of structural information.

**C. UCI DATASETS**

In order to further testify the effectiveness of LSI-TSVM, sixteen regular scale UCI datasets [40] and eight large scale datasets are chosen. They are Australian, balance, blood, diabetes, dsatset, german, heart, iris, titanic, liverdisorder, msplice, glass, segment, spambase, splice, vehicle, vote, waveform, wdbc, wine, wpbc, X8D5K, DNA and landsat. These datasets include binary datasets and multi-class datasets. And their attributes are shown in Table 2. For multi-class datasets, the binary tree algorithm is adopted. Our LSI-TSVM, KNN-STSVM, S-TSVM, TSVM, SR SVM, WCSS-F SVM and SVM are tested and compared on these datasets.

**TABLE 2. The attributes of UCI datasets.**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Feature</th>
<th>Data</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>australian</td>
<td>14</td>
<td>6901</td>
<td>2</td>
</tr>
<tr>
<td>balance</td>
<td>4</td>
<td>625</td>
<td>3</td>
</tr>
<tr>
<td>blood</td>
<td>4</td>
<td>748</td>
<td>2</td>
</tr>
<tr>
<td>diabetes</td>
<td>8</td>
<td>768</td>
<td>2</td>
</tr>
<tr>
<td>dsatset</td>
<td>180</td>
<td>1186</td>
<td>2</td>
</tr>
<tr>
<td>german</td>
<td>24</td>
<td>1000</td>
<td>2</td>
</tr>
<tr>
<td>heart</td>
<td>12</td>
<td>270</td>
<td>2</td>
</tr>
<tr>
<td>iris</td>
<td>4</td>
<td>150</td>
<td>3</td>
</tr>
<tr>
<td>titanic</td>
<td>12</td>
<td>1309</td>
<td>2</td>
</tr>
<tr>
<td>liverdisorder</td>
<td>6</td>
<td>345</td>
<td>2</td>
</tr>
<tr>
<td>msplice</td>
<td>23</td>
<td>3175</td>
<td>3</td>
</tr>
<tr>
<td>glass</td>
<td>9</td>
<td>214</td>
<td>6</td>
</tr>
<tr>
<td>segment</td>
<td>8</td>
<td>2310</td>
<td>7</td>
</tr>
<tr>
<td>spambase</td>
<td>57</td>
<td>4601</td>
<td>2</td>
</tr>
<tr>
<td>splice</td>
<td>60</td>
<td>1000</td>
<td>2</td>
</tr>
<tr>
<td>vehicle</td>
<td>18</td>
<td>846</td>
<td>4</td>
</tr>
<tr>
<td>vote</td>
<td>16</td>
<td>435</td>
<td>2</td>
</tr>
<tr>
<td>waveform</td>
<td>22</td>
<td>5000</td>
<td>3</td>
</tr>
<tr>
<td>wdbc</td>
<td>30</td>
<td>569</td>
<td>2</td>
</tr>
<tr>
<td>wine</td>
<td>12</td>
<td>178</td>
<td>3</td>
</tr>
<tr>
<td>wpbc</td>
<td>33</td>
<td>198</td>
<td>2</td>
</tr>
<tr>
<td>X8D5K</td>
<td>8</td>
<td>1000</td>
<td>5</td>
</tr>
<tr>
<td>DNA</td>
<td>180</td>
<td>3186</td>
<td>3</td>
</tr>
<tr>
<td>landsat</td>
<td>36</td>
<td>2000</td>
<td>6</td>
</tr>
</tbody>
</table>

For each dataset, the samples are divided into two nonoverlapping parts. 50% samples are concerned as training.
dataset, and the others are testing dataset. Moreover, all samples are normalized into [0, 1]. All parameters of algorithms are obtained by grid searching. The final testing results are shown in Table 3 and Table 4. Table 3 lists the training time and the mean and standard deviation of testing accuracies for seven classifiers with linear kernel. Table 4 lists the same items of classifiers with Gaussian kernel. Moreover, the optimal performance of all algorithms is determined with five-fold cross validation.

Table 3 shows that our LSI-TSVM obtains the best classification accuracy for sixteen UCI datasets in linear space. For liverdisorder, splice, X8D5K and landsat datasets, the classification accuracy of LSI-TSVM is nearly the best. Because the local within-class scatter and between-class scatter are not fit for blood, heart, glass and vehicle datasets, the accuracy of our LSI-TSVM is low but not the worst. Testing results in Table 4 and Table 3 are similar. And the following conclusions can be drawn from them:

1. For most cases, LSI-TSVM, KNN-STSVM, S-TSVM, WSC-FSV and SRSVM are superior to TSV and SVM, which means the prior structural information can improve classification performance.

2. For most datasets, the testing accuracy of LSI-TSVM is superior to that of S-TSVM, which means the local within-class scatter and between-class scatter obtained by using the concept of NMFDA have better generalization compared with the structural information which is captured by using clustering method. For few cases, the accuracy of LSI-TSVM is lower than that of S-TSVM because the NMFDA is not fit for some datasets.

3. For most cases, SRSVM and WCS-FSV are better than TSV. However, LSI-TSVM, KNN-STSVM and S-TSVM are better than SRSVM and WCS-FSV for most cases, which shows that the nonparallel classification hyperplanes can avoid the conflict of structural information between two classes of samples. And the results further prove that the prior structural information can improve algorithm’s generalization performance.

4. In terms of computational time, the LSI-TSVM, KNN-STSVM, S-TSVM and TSV are faster than SVM, SRSVM and WCS-FSV, because they solve a pair of QPPs rather than a single one. For large scale datasets, we can find that KNN-STSVM becomes much faster. In addition, for most cases, we can find that our LSI-TSVM is slower than S-TSVM and KNN-STSVM.

Table 3: The testing accuracy of linear classifiers on UCI datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>SVM</th>
<th>TSVM</th>
<th>SRSVM</th>
<th>WSC-FSV</th>
<th>S-TSVM</th>
<th>KNN-STSVM</th>
<th>LSI-TSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accurancy</td>
<td>Time(s)</td>
<td>Accurancy</td>
<td>Time(s)</td>
<td>Accurancy</td>
<td>Time(s)</td>
<td>Accurancy</td>
<td>Time(s)</td>
</tr>
<tr>
<td>Australian</td>
<td>86.09±2.30</td>
<td>87.24±2.79</td>
<td>86.38±1.94</td>
<td>86.95±2.57</td>
<td>87.54±2.04</td>
<td>86.67±2.10</td>
<td>87.54±2.66</td>
</tr>
<tr>
<td>Balance</td>
<td>86.67±2.06</td>
<td>87.21±2.62</td>
<td>87.53±2.25</td>
<td>87.21±3.62</td>
<td>87.37±3.45</td>
<td>86.73±4.04</td>
<td>87.85±3.24</td>
</tr>
<tr>
<td>Blood</td>
<td>76.34±0.47</td>
<td>77.81±1.17</td>
<td>76.34±0.47</td>
<td>76.20±0.25</td>
<td>77.04±1.07</td>
<td>75.32±1.03</td>
<td>78.35±1.03</td>
</tr>
<tr>
<td>Diabetes</td>
<td>76.16±3.90</td>
<td>77.60±2.13</td>
<td>77.60±2.79</td>
<td>75.13±2.40</td>
<td>77.86±2.27</td>
<td>77.86±2.82</td>
<td>78.76±1.98</td>
</tr>
<tr>
<td>Dna test</td>
<td>85.92±3.77</td>
<td>93.42±1.73</td>
<td>88.19±2.98</td>
<td>92.74±1.04</td>
<td>93.59±1.55</td>
<td>92.41±1.52</td>
<td>93.93±1.64</td>
</tr>
<tr>
<td>German</td>
<td>75.70±4.06</td>
<td>76.40±2.33</td>
<td>76.90±2.30</td>
<td>75.68±2.41</td>
<td>76.20±1.18</td>
<td>76.60±0.31</td>
<td>83.00±2.48</td>
</tr>
<tr>
<td>Heart</td>
<td>81.48±2.16</td>
<td>82.98±3.19</td>
<td>85.19±2.87</td>
<td>84.81±2.72</td>
<td>82.96±3.19</td>
<td>82.96±3.19</td>
<td>83.00±2.48</td>
</tr>
<tr>
<td>Iris</td>
<td>98.00±1.63</td>
<td>98.00±1.63</td>
<td>98.00±1.63</td>
<td>98.00±1.63</td>
<td>98.00±1.63</td>
<td>98.00±1.63</td>
<td>98.00±1.63</td>
</tr>
<tr>
<td>Titanic</td>
<td>75.69±1.48</td>
<td>77.57±0.41</td>
<td>77.60±0.37</td>
<td>77.60±1.12</td>
<td>77.64±0.42</td>
<td>78.42±0.37</td>
<td>78.59±0.65</td>
</tr>
<tr>
<td>Liver</td>
<td>66.77±6.26</td>
<td>70.43±6.12</td>
<td>68.70±5.84</td>
<td>69.56±0.21</td>
<td>71.30±5.38</td>
<td>69.57±0.61</td>
<td>71.01±6.01</td>
</tr>
<tr>
<td>Mtsplice</td>
<td>92.38±0.79</td>
<td>92.85±0.74</td>
<td>92.88±0.72</td>
<td>93.05±0.76</td>
<td>93.10±0.82</td>
<td>92.91±1.01</td>
<td>93.26±0.82</td>
</tr>
<tr>
<td>Glass</td>
<td>60.30±9.83</td>
<td>63.48±7.07</td>
<td>66.32±7.79</td>
<td>60.35±8.52</td>
<td>65.79±6.81</td>
<td>62.95±6.19</td>
<td>64.39±8.12</td>
</tr>
<tr>
<td>Segment</td>
<td>93.27±0.24</td>
<td>95.11±0.22</td>
<td>97.60±0.08</td>
<td>97.56±0.55</td>
<td>97.66±0.14</td>
<td>97.68±0.15</td>
<td>97.99±0.91</td>
</tr>
<tr>
<td>Sparabase</td>
<td>84.16±4.15</td>
<td>86.12±5.53</td>
<td>82.25±2.96</td>
<td>87.42</td>
<td>90.09±8.45</td>
<td>92.99±2.54</td>
<td>93.78±0.79</td>
</tr>
<tr>
<td>Splice</td>
<td>79.90±1.88</td>
<td>80.81±2.17</td>
<td>80.50±1.95</td>
<td><strong>81.50±1.96</strong></td>
<td>80.80±1.17</td>
<td>81.00±1.17</td>
<td>81.00±1.17</td>
</tr>
<tr>
<td>Vehicle</td>
<td>72.22±5.52</td>
<td>74.35±2.93</td>
<td>75.29±2.70</td>
<td>74.56±0.29</td>
<td>74.70±0.17</td>
<td>75.52±2.50</td>
<td>74.94±2.33</td>
</tr>
<tr>
<td>Vote</td>
<td>94.48±1.36</td>
<td>94.74±1.56</td>
<td>94.25±0.72</td>
<td>95.62±2.87</td>
<td>94.80±1.55</td>
<td>95.00±1.76</td>
<td>95.10±1.01</td>
</tr>
<tr>
<td>Waveform</td>
<td>79.46±1.39</td>
<td>82.72±0.81</td>
<td>88.05±0.89</td>
<td>88.19±0.64</td>
<td>88.26±0.51</td>
<td>88.28±0.60</td>
<td>88.42±0.25</td>
</tr>
<tr>
<td>Wdbc</td>
<td>97.19±0.96</td>
<td>96.84±3.51</td>
<td>96.14±1.04</td>
<td>97.39±0.34</td>
<td>96.84±1.02</td>
<td>96.49±1.45</td>
<td>97.54±1.02</td>
</tr>
<tr>
<td>Wine</td>
<td>96.65±3.29</td>
<td>98.32±1.38</td>
<td>97.29±1.36</td>
<td>97.30±1.36</td>
<td><strong>98.32±1.38</strong></td>
<td>97.78±2.04</td>
<td>98.32±1.38</td>
</tr>
<tr>
<td>Wpbc</td>
<td>80.56±9.59</td>
<td>81.85±9.17</td>
<td>80.25±4.45</td>
<td>80.85±7.09</td>
<td>83.86±5.57</td>
<td>81.85±8.89</td>
<td>85.88±3.34</td>
</tr>
<tr>
<td>X8D5K</td>
<td>99.60±0.37</td>
<td>99.90±0.20</td>
<td>99.80±0.24</td>
<td><strong>100±0.00</strong></td>
<td>99.90±0.20</td>
<td>99.90±0.20</td>
<td>99.90±0.20</td>
</tr>
</tbody>
</table>
VI. CONCLUSION
Many structural SVMs exploit the structure of data to improve the algorithm’s generalization. However, the structural information in WCS-SVM, Fisher-SVM and FQSSVM cannot discover the local structure of data. And the structural information in SRSVM, T-SVM, KNN-STSVM, SNPSVM and SRPTSVSVM does not include local between-class margin. So, we introduced the local structural information into TSV model, resulting in the LSI-TSVSVM classifier. The LSI-TSVSVM classifier first adopts some techniques to extract the local scatter of data. Then it introduces not only the original global within-class clustering and between-class margin, but also the local within-class scatter and between-class scatter. The structural information really reflects the aggregation of within-class samples and the separability of between-class samples. In addition to the linear model of LSI-TSVSVM, we have extended it to the nonlinear case with the kernel trick. The new classifier has been tested on synthetic dataset, image dataset and UCI dataset respectively. The testing results have confirmed that our LSI-TSVSVM successfully digs the nonlinear structure of data and obtains the good performance. But, LSI-TSVSVM is slower than S-TSVSVM and KNN-STSVM. So, further work may focus on how to solve and optimize QPP in LSI-TSVSVM more effectively.

ACKNOWLEDGMENT
The authors gratefully acknowledge the helpful comments and suggestions of anonymous reviewers.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>SVM</th>
<th>TSVM</th>
<th>SRSVM</th>
<th>WCSSVM</th>
<th>ST-SVM</th>
<th>KNN-STSVM</th>
<th>LSI-TSVSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Time(s)</td>
<td>Acc</td>
<td>Time(s)</td>
<td>Acc</td>
<td>Time(s)</td>
<td>Acc</td>
</tr>
<tr>
<td>australian</td>
<td>0.86</td>
<td>0.50</td>
<td>0.87</td>
<td>0.88</td>
<td>0.87</td>
<td>1.16</td>
<td>0.86</td>
</tr>
<tr>
<td>balance</td>
<td>0.28</td>
<td>1.20</td>
<td>0.45</td>
<td>1.20</td>
<td>0.45</td>
<td>1.20</td>
<td>0.45</td>
</tr>
<tr>
<td>blood</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>diabetes</td>
<td>0.77</td>
<td>2.26</td>
<td>0.78</td>
<td>2.26</td>
<td>0.78</td>
<td>2.26</td>
<td>0.78</td>
</tr>
<tr>
<td>dnatest</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>german</td>
<td>0.74</td>
<td>3.72</td>
<td>0.75</td>
<td>3.72</td>
<td>0.75</td>
<td>3.72</td>
<td>0.75</td>
</tr>
<tr>
<td>heart</td>
<td>0.19</td>
<td>1.13</td>
<td>0.20</td>
<td>1.13</td>
<td>0.20</td>
<td>1.13</td>
<td>0.20</td>
</tr>
<tr>
<td>iris</td>
<td>0.97</td>
<td>3.49</td>
<td>0.98</td>
<td>3.49</td>
<td>0.98</td>
<td>3.49</td>
<td>0.98</td>
</tr>
<tr>
<td>titanic</td>
<td>0.76</td>
<td>2.84</td>
<td>0.77</td>
<td>2.84</td>
<td>0.77</td>
<td>2.84</td>
<td>0.77</td>
</tr>
<tr>
<td>liverdisorder</td>
<td>0.72</td>
<td>7.67</td>
<td>0.73</td>
<td>7.67</td>
<td>0.73</td>
<td>7.67</td>
<td>0.73</td>
</tr>
<tr>
<td>missplice</td>
<td>0.94</td>
<td>2.61</td>
<td>0.95</td>
<td>2.61</td>
<td>0.95</td>
<td>2.61</td>
<td>0.95</td>
</tr>
<tr>
<td>glass</td>
<td>0.73</td>
<td>8.66</td>
<td>0.74</td>
<td>8.66</td>
<td>0.74</td>
<td>8.66</td>
<td>0.74</td>
</tr>
<tr>
<td>segment</td>
<td>0.97</td>
<td>4.21</td>
<td>0.98</td>
<td>4.21</td>
<td>0.98</td>
<td>4.21</td>
<td>0.98</td>
</tr>
<tr>
<td>spambase</td>
<td>0.85</td>
<td>4.66</td>
<td>0.86</td>
<td>4.66</td>
<td>0.86</td>
<td>4.66</td>
<td>0.86</td>
</tr>
<tr>
<td>splice</td>
<td>0.83</td>
<td>5.86</td>
<td>0.84</td>
<td>5.86</td>
<td>0.84</td>
<td>5.86</td>
<td>0.84</td>
</tr>
<tr>
<td>vehicle</td>
<td>0.97</td>
<td>5.81</td>
<td>0.98</td>
<td>5.81</td>
<td>0.98</td>
<td>5.81</td>
<td>0.98</td>
</tr>
<tr>
<td>vote</td>
<td>0.95</td>
<td>5.13</td>
<td>0.95</td>
<td>5.13</td>
<td>0.95</td>
<td>5.13</td>
<td>0.95</td>
</tr>
<tr>
<td>waveform</td>
<td>0.89</td>
<td>0.79</td>
<td>0.90</td>
<td>0.79</td>
<td>0.90</td>
<td>0.79</td>
<td>0.90</td>
</tr>
<tr>
<td>wdbc</td>
<td>0.97</td>
<td>4.19</td>
<td>0.98</td>
<td>4.19</td>
<td>0.98</td>
<td>4.19</td>
<td>0.98</td>
</tr>
<tr>
<td>wine</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>wpbc</td>
<td>0.82</td>
<td>8.24</td>
<td>0.82</td>
<td>8.24</td>
<td>0.82</td>
<td>8.24</td>
<td>0.82</td>
</tr>
<tr>
<td>X8D5K</td>
<td>100.00</td>
<td>0.00</td>
<td>100.00</td>
<td>0.00</td>
<td>100.00</td>
<td>0.00</td>
<td>100.00</td>
</tr>
<tr>
<td>DNA</td>
<td>0.25</td>
<td>0.42</td>
<td>0.38</td>
<td>0.42</td>
<td>0.38</td>
<td>0.42</td>
<td>0.38</td>
</tr>
<tr>
<td>landsat</td>
<td>0.31</td>
<td>0.52</td>
<td>0.31</td>
<td>0.52</td>
<td>0.31</td>
<td>0.52</td>
<td>0.31</td>
</tr>
</tbody>
</table>

| TABLE 4. The testing accuracy of nonlinear classifiers on UCI datasets. |
REFERENCES


MAOXIANG CHU was born in September 7, 1978. He received Ph.D degree from Northeastern University in 2015. Now he is an associate professor and doctoral supervisor at School of Electronic and Information Engineering in University of Science and Technology Liaoning. His research interests include pattern recognition, machine learning, image processing and intelligent control.

LIMING LIU was born in June 10, 1994. She is a graduate student at School of Electronic and Information Engineering in University of Science and Technology Liaoning. Her main research direction is pattern recognition and machine learning.
YONGHUI YANG was born in March 23, 1971. He is a professor and dean of School of Electronic and Information Engineering in University of Science and Technology Liaoning. His research interests include machine learning, information processing, intelligent control and information engineering.

RONGFEN GONG was born in April 12, 1979. She received Master's degree from University of Science and Technology Liaoning in 2005. Now she is a lecturer at School of Electronic and Information Engineering in University of Science and Technology Liaoning. Her main research direction is pattern recognition and machine learning.