Learning One-class KSVM+ for Multi-class Problems with Group Information of Data

WENSONG ZHAO
Liaocheng University
School of Mathematics Sciences
Liaocheng, 252059, P.R. CHINA
lczhaowensong@126.com

LIYA FAN[*]
Liaocheng University
School of Mathematics Sciences
Liaocheng, 252059, P.R. CHINA
fanliya63@126.com

Abstract: This paper is denoted to study the effect of the group information of data in one-class kernel support vector machines (OC-KSVMs) for classification accuracy and time consumed of multi-class classification data. Two new classification methods based on OC-KSVMs are presented. One is OC-KSVM with maximum margin from the origin and group information of data (briefly, MMOC-KSVM+) and another is OC-KSVM with hyper-sphere and group information of data (briefly, HSOC-KSVM+). We proved theoretically that MMOC-KSVM and HSOC-KSVM are equivalent for Gaussian RBF kernels. Experiments on three real-words data sets are performed in order to test and evaluate the efficacy of the proposed methods. Experimental results indicate that the group information of data can improve the classification accuracy of data and meanwhile increase the time consumed of algorithms.

Key–Words: multi-class classification problem; One-class kernel SVM; group information of data; maximum margin; hyper-sphere

1 Introduction

Over the last decades, kernel support vector machines (KSVMs) have become a novel method for many classification problems, such as object recognition [1], classification of cancer morphologies [2], handwritten characters and digit recognition [3-5] and so on [6-10], because of their flexibility, computational efficiency and capacity to handle dimensional data. KSVMs try to find an optimal decision hyperplane in the feature space by maximizing the margin or degree of separation between different class data [11-12].

SVMs have been extended to handle multi-class classification problems [13-14]. Since the standard SVMs are designed for binary classification problems, multi-class classification problems are commonly solved by a decomposition to several binary problems for which the standard SVMs can be used, for instance, one-against-all (OAA) and one-against-one (OAO) decompositions are often applied [15-18]. In OAA support vector machine (OAA-SVM), a classification problem with $c(c \geq 3)$ classes can be decomposed to c binary problems by using OAA decomposition, in which the sth class is separated from the other training patterns, and then c decision functions $f_s(x), s = 1, \cdots, c$ can be obtained by using SVMs. The classification of a pattern $\widetilde{x}$ is performed according to maximal value of functions $\{f_s(\widetilde{x})\}_{s=1}^{c}$, that is, the label of $\widetilde{x}$ can be gotten as $\arg\max\limits_{0\leq s\leq c} f_s(\widetilde{x})$.

Recently, SVMs have been extended to handle one-class classification problems, in which it is assumed that information relative to the class of interest, the target class, is only available. This means that objects from the target class are only used and that no information about the other classes is considered [19]. The task of one-class SVMs (OC-SVMs) is to determine a closed domain with maximal margin from the origin or a hyper-sphere that contains almost all the data of the target class but also allows discarding outliers. Any pattern point lying outside the enclosed region is considered as an outlier. OC-SVMs have been used for document classification [20], classification of sounds, classification of cancer morphologies and so on. Rabaoui et al. [21] introduce an advanced dissimilarity measure for OC-SVMs and illustrate the performance of these methods on an audio data. Mireille et al. [22] propose a modified maximum margin OC-SVM method as a discriminant framework to deal with multi-class problems. Yuhua Li [23] proposes a training point selection method for OC-SVMs. Heng et al. [24] demonstrate the use of principal components analysis for OC-SVMs as a dimension reduction tool. It is well-known that the learning ability of OC-SVMs originates from the kernel trick, which has been widely used to tackle complicated classification

[*]Corresponding author.

problems by a feature mapping from a original input space to a kernel feature space. Although in general the dimensionality of the feature space could be arbitrarily large or even infinite, the feature mapping can be specified implicitly by a kernel function.

Nowadays, in order to improve classification accuracy of SVMs, additional information hidden in data is considered and a KSVM with structured data (denoted by KSVM+) is presented by Vapnik [25]. A main difference between KSVM+ and KSVM is that KSVM projects inputs into one space whereas KSVM+ into two different spaces: decision space and correcting space. Liang et al. [26] describe the application of SVM+ and Multi Task Learning (MTL) to classification problems. Cai et al. [27] propose a new methodology for regression problems by means of SVM+ and MTL. Liang et al. [28] propose a new multi-task learning method (denoted by SVM+MTL) for MTL problems.

Motivated by works mentioned above, this paper is devoted to research the effect of the group information of data for multi-class classification problems. Two new kinds of classification methods are presented by means of OC-KSVMs, which are named as OC-KSVM with maximum margin from the origin (briefly, MMOC-KSVM+) and OC-KSVM with hyper-sphere (briefly, HSOC-KSVM+), respectively. In order to test and evaluate the efficacy of the proposed methods, a series of comparative experiments with KSVM, KSVM+, OC-KSVM, OAO-KSVM and OAA-KSVM are performed on Wisconsin Breast Cancer (WBC), Wine and Kennedy Space Center (KSC) three data sets. Experimental results indicate that the group information of data can improve the classification accuracy and meanwhile increase the time consumed, and that MMOC-KSVM+ is better than HSOC-KSVM+.

The rest of the paper is organized as follows. Bianry KSVM+ is briefly reviewed in Section 2 and OC-KSVM is recalled in Section 3. MMOC-KSVM+ and HSOC-KSVM+ are introduced for handling one-class classification problems in Section 4 and for handling multi-class classification problems with group information of data in Section 5. A series of comparative experiments with OAO-KSVM, OAA-KSVM and OC-KSVM are performed in Section 6 and some conclusions are given in Section 7.

2 Binary KSVM+

This section briefly recalls binary KSVM+ used in the sequel. Let $\{(x_i, y_i)\}_{i=1}^n \subset R^m \times \{\pm 1\}$ be a binary linearly nonseparable data set, where $y_i \in \{\pm 1\}$ is the class label of the ith sample $x_i, i = 1, \cdots, n$. In the last few years, kernel methods have attracted much attention [29-30] and have become one of the most popular approaches (see [31]). A kernel function $k : R^m \times R^m \to R$ satisfies

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle, \forall x, y \in R^m,$$

where $\varphi : R^m \to H$ and H are a feature mapping and a feature space corresponding to kernel k, respectively, and $\langle , \rangle$ denotes the inner product in H. By means of the feature mapping φ, binary linearly nonseparable samples $\{x_i\}_{i=1}^n$ can be mapped into H such that $\{\varphi(x_i)\}_{i=1}^n \subset H$ is approximately linearly separable. The following is several common kernel functions:

(i) linear kernel: $k(x, y) = \langle x, y \rangle$ for all $x, y \in R^m$.

(ii) polynomial kernel: $k(x, y) = (\langle x, y \rangle + c)^m$ for all $x, y \in R^m$, where $c \geq 0$ and $m > 0$ are user' parameters.

(iii) Gaussian radius base function (RBF) kernel: $k(x, y) = \exp(-\frac{1}{\sigma^2} \| x - y \|^2)$ for all $x, y \in R^m$, where $\sigma > 0$ is a user' parameter.

2.1 Binary KSVM

For approximately linearly separable binary problem, soft-margin SVM is to find an optimal separating hyperplane $\langle \omega, x \rangle + b = 0$ by considering the following optimization problem:

$$\begin{array}{l} \min\limits_{\omega,b,\xi_i} \frac{1}{2}\|\omega\|^2 + C \sum\limits_{i=1}^n \xi_i \\ s.t. \;\; y_i(\langle \omega, x_i \rangle + b) \geq 1 - \xi_i, \\ \qquad \xi_i \geq 0, i = 1, \dots, n, \end{array} \tag{1}$$

where $\omega \in R^m$ and $b \in R$ are respectively the normal vector and a bias of the separating hyperplane, $C > 0$ is a user' parameter and $\{\xi_i\}_{i=1}^n$ are slack variables. By solving the Wolfe dual form of the problem (1):

$$\begin{array}{l} \min\limits_{\alpha_i} \sum\limits_{i,j=1}^n \frac{1}{2} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - \sum\limits_{i=1}^n \alpha_i \\ s.t. \;\; \sum_{i=1}^n \alpha_i y_i = 0, \\ \qquad 0 \leq \alpha_j \leq C, j = 1, \dots, n, \end{array} \tag{2}$$

where $\{\alpha_i\}_{i=1}^n$ are Lagrange multipliers, it can get soft-margin SVM.

For linearly nonseparable binary problem, select a proper kernel function $k : R^m \times R^m \to R$ with the feature space H and feature mapping $\varphi : R^m \to H$ and map the samples $\{x_i\}_{i=1}^n$ into $\{\varphi(x_i)\}_{i=1}^n$. In this case, the problem (1) can be rewritten as the following optimization problem:

$$\begin{array}{l} \min\limits_{\omega,b,\xi_i} \frac{1}{2}\|\omega\|^2 + C \sum\limits_{i=1}^n \xi_i \\ s.t. \;\; y_i(\langle \omega, \varphi(x_i) \rangle + b) \geq 1 - \xi_i, \\ \qquad \xi_i \geq 0, i = 1, \dots, n, \end{array} \tag{3}$$

where $\omega \in H$. By solving the Wolfe dual form of the problem (3):

$$\begin{array}{ll}\min\limits_{\alpha_i} \frac{1}{2}\sum\limits_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j)-\sum\limits_{i=1}^{n}\alpha_i \\ s.t.\ \sum_{i=1}^{n}\alpha_i y_i=0, \\ \qquad 0\le\alpha_j\le C, j=1,\ldots,n,\end{array}\tag{4}$$

(ω, b) can be obtained by

$$\begin{array}{l}\omega=\sum_{i=1}^{n}\alpha_i y_i\varphi(x_i),\\ b=y_j-\sum_{i=1}^{n}\alpha_i y_i k(x_i,x_j),\end{array}$$

for some $\alpha_j \in (0, C)$. Consequently, the decision function $f(x) = \sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b$. The specific algorithm is as follows.

Algorithm 1. (KSVM)

Step 1. Give a binary data set $\{(x_i, y_i)\}_{i=1}^{n} \subset R^m \times \{\pm 1\}$.

Step 2. Select a proper kernel function $k : R^m \times R^m \to R$ and an appropriate parameter $C > 0$.

Step 3. Solve the problem (4) and obtain the optimal solution $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$.

Step 4. Select a positive component $\alpha_j^* \in (0, C)$ and calculate $b^* = y_j - \sum_{i=1}^{n} y_i \alpha_i^* k(x_j, x_i)$.

Step 5. Construct the decision function $f(x) = \sum_{i=1}^{n} \alpha_i^* y_i k(x_i, x) + b^*$.

Step 6. For a new sample $\widetilde{x}$, it can be inferred that $\widetilde{x}$ belongs to class 1 if $f(\widetilde{x}) \ge 0$ and otherwise to class -1.

2.2 Binary KSVM+

This subsection briefly recalls binary KSVM+, for details, see [25]. Let $\{(x_i, y_i)\}_{i=1}^{n} \subset R^m \times \{\pm 1\}$ be a binary linearly nonseparable data set with group information and put $X = \{x_i\}_{i=1}^{n}$, $Y = \{y_i\}_{i=1}^{n}$ and $X \times Y = \{(x_i, y_i)\}_{i=1}^{n}$. Suppose that X is a union of t groups ($t > 1$), n_r is the number of samples in group r and $n = \sum_{r=1}^{t} n_r$. Put $X_r = \{x_1^r, \cdots, x_{n_r}^r\}$, $Y_r = \{y_1^r, \cdots, y_{n_r}^r\}$ and $T_r = \{1, \cdots, n_r\}$, then $X \times Y = \cup_{r=1}^{t} X_r \times Y_r$.

The group information hidden in data is often used to impose additional constraints on the slack variables in the problem (3). By means of the group information and kernel functions, each sample can be mapped into two different spaces: a decision space Z via the feature mapping $\varphi : X \to Z(z_i = \varphi(x_i))$ corresponding to a kernel k and a correcting space Z_r via the feature mapping $\varphi_r : X_r \to Z_r(z_i^r = \varphi_r(x_i^r))$ corresponding to a kernel $k_r, r \in \{1, \cdots, t\}$. By using binary KSVM in correcting space Z_r, t correcting functions $f_r(x) = \langle w_r, z^r\rangle + d_r, r = 1, \cdots, t$ can be obtained. By imposing these correcting functions onto slack variables of the problem (3):

$$\xi_i^r=\langle w_r, z_i^r\rangle+d_r, i\in T_r, r=1,\cdots,t,$$

we can obtain the following optimization problem, which is an improvement of the problem (3):

$$\begin{array}{ll}\min\limits_{\omega,\omega_r,b,d_r} & \frac{1}{2}\|\omega\|^2+\frac{v}{2}\sum\limits_{r=1}^{t}\|\omega_r\|^2+C\sum\limits_{r=1}^{t}\sum\limits_{i\in T_r}\xi_i^r \\ s.t. & y_i(\langle\omega\cdot\varphi(x_i^r)\rangle+b)\ge 1-\xi_i^r, \\ & \xi_i^r=\langle\omega_r\cdot\varphi_r(x_i^r)\rangle+d_r\ge 0, \\ & i\in T_r,\ r=1,\ldots,t,\end{array}\tag{5}$$

where $v > 0$ and $C > 0$ are user' parameters and $b \in R$ and $d_r \in R$ are biases. v adjusts the relative weight of decision function and correcting functions and C controls the trade-off between complexity and the number of nonseparable samples. By solving the Wolfe dual from of the problem (5):

$$\begin{array}{l}\min\limits_{\alpha_i,\beta_i}\frac{1}{2}\sum\limits_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j)-\sum\limits_{i=1}^{n}\alpha_i \\ \quad +\frac{1}{2v}\sum\limits_{r=1}^{t}\sum\limits_{i,j\in T_r}(\alpha_i^r+\beta_i^r-C)(\alpha_j^r \\ \quad +\beta_j^r-C)k_r(x_i,x_j) \\ s.t.\ \sum_{i=1}^{n}\alpha_i y_i=0, \\ \quad \sum_{i\in T_r}(\alpha_i^r+\beta_i^r-C)=0, r=1,\ldots,t, \\ \quad \alpha_i,\beta_i\ge 0, i=1,\ldots,n,\end{array}\tag{6}$$

where $\{\alpha_i\}_{i=1}^{n}$ and $\{\beta_i\}_{i=1}^{n}$ are Lagrange multipliers, it gets

$$\begin{array}{l}\omega=\sum_{i=1}^{n}\alpha_i y_i\varphi(x_i),\\ \omega_r=\frac{1}{v}\sum_{i\in T_r}(\alpha_i^r+\beta_i^r-C)\varphi_r(x_i^r),\\ b=y_j-\sum_{i=1}^{n}\alpha_i y_i k(x_i,x_j),\\ d_r=-\frac{1}{v}\sum_{i\in T_r}(\alpha_i^r+\beta_i^r-C)k_r(x_i,x_j),\end{array}$$

for some $\alpha_j > 0$. Proceed to the next step, the decision function $f(x)$ can be obtained. The specific algorithm is as follows.

Algorithm 2. (KSVM+)

Step 1. Give a binary linearly nonseparable data set $X \times Y$ and divide it into $t > 1$ groups $X_r \times Y_r, r = 1, \ldots, t$ according to the group information of data.

Step 2. Select proper kernel functions $k : R^m \times R^m \to R$ and $k_r : R^m \times R^m \to R$ for group r, and appropriate parameters $C > 0$ and $v > 0$.

Step 3. Solve the problem (6) and obtain the optimal solution $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$.

Step 4. Calculate $b^* = y_j - \sum_{i=1}^{n} \alpha_i^* y_i k(x_i, x_j)$ for some $\alpha_j^* > 0$.

Step 5. Construct the decision function $f(x) = \sum_{i=1}^{n} \alpha_i^* y_i k(x_i, x) + b^*$.

Step 6. For a new sample $\widetilde{x}$, it can be inferred that $\widetilde{x}$ belongs to class 1 if $f(\widetilde{x}) \ge 0$ and otherwise to class -1.

3 OC-KSVM

Binary KSVMs have been extended to handle one-class classification problems, in which it is assumed that information relative to the class of interest, the target class, is only available. This means that samples from the target class are only used and that no information about the other classes is considered. Unlike binary KSVMs, the task of one-class KSVMs (OC-KSVMs) is to determine a closed domain in feature space with maximal margin from the origin or a closed-sphere that contains almost all the data of the target class but also allows discarding outliers. OC-KSVM determining a closed domain with maximal margin from the origin is called maximum margin OC-KSVM (briefly, MMOC-KSVM) and with closed-sphere is called hyper-sphere OC-KSVM (briefly, HSOC-KSVM). In the following, we briefly review MMOC-KSVM and HSOC-KSVM, for details, see [19-24].

Let $X = \{x_i\}_{i=1}^n \subset R^m$ be a target class and H and $\varphi : R^m \to H$ be the Reproducing Kernel Hilbert Space (RKHS) and the feature mapping of a given kernel $k : R^m \times R^m \to R$, respectively.

3.1 MMOC-KSVM

The aim of MMOC-KSVM is to determine a hyperplane $\langle \omega, \varphi(x) \rangle - \rho = 0$ that separates most of the data in X from the origin by maximizing the distance from the origin to the separating hyperplane. The decision is given by the decision function $f(x) = \langle \omega, \varphi(x) \rangle - \rho \geq 0$, where $\omega \in H$ and $\rho \in R$ result from the following modified optimization problem:

$$\begin{array}{l} \min\limits_{\omega,\rho,\xi_i} \frac{1}{2}\|\omega\|^2 + C \sum\limits_{i=1}^{n} \xi_i - \rho \\ s.t.\ \langle \omega, \varphi(x_i) \rangle \geq \rho - \xi_i, \\ \qquad \xi_i \geq 0, i = 1, \ldots, n, \end{array} \tag{7}$$

where $C > 0$ is a user' parameter and $\{\xi_i\}_{i=1}^n$ are slack variables. By solving the Wolfe dual form of the problem (7):

$$\begin{array}{l} \min\limits_{\alpha_i} \frac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \\ s.t.\ \sum_{i=1}^n \alpha_i = 1, \\ \qquad 0 \leq \alpha_i \leq C,\ i = 1, \ldots, n, \end{array} \tag{8}$$

where $\{\alpha_i\}_{i=1}^n$ are Lagrangian multipliers, it gets

$$\begin{array}{l} \omega = \sum_{i=1}^n \alpha_i \varphi(x_i), \\ \rho = \sum_{i=1}^n \alpha_i k(x_i, x_j), \end{array}$$

for some $\alpha_j \in (0, C)$. Consequently, the decision function $f(x)$ can be obtained. The specific algorithm is as follows.

Algorithm 3. (MMOC-KSVM)

Step 1. Given a target class $X = \{x_i\}_{i=1}^n \subset R^m$.

Step 2. Select a proper kernel function $k : R^m \times R^m \to R$ and an appropriate parameter $C > 0$.

Step 3. Solve the problem (8) and obtain the optimal solution $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$.

Step 4. Calculate $\rho^* = \sum_{i=1}^n \alpha_i^* k(x_i, x_j)$ for some $\alpha_j^* \in (0, C)$.

Step 5. Construct the decision function $f(x) = \sum_{i=1}^n \alpha_i^* k(x_i, x) - \rho^*$.

Step 6. For a new sample $\widetilde{x}$, it can be inferred that $\widetilde{x}$ belongs to class 1 if $f(\widetilde{x}) \geq 0$ and otherwise to class -1.

3.2 HSOC-KSVM

Unlike MMOC-KSVM, the aim of HSOC-KSVM is to determine a hyper-sphere $r^2 - \|\varphi(x) - a\|^2 \geq 0$ that contains almost all the data of the target class, where $r \in R$ and $a \in H$ are the radius and center of the hyper-sphere, respectively. The decision is given by the decision function $f(x) = r^2 - \|\varphi(x) - a\|^2$. the radius r and the center a can be obtained by solving the following modified optimization problem:

$$\begin{array}{l} \min\limits_{r,a,\xi_i} r^2 + C \sum\limits_{i=1}^{n} \xi_i \\ s.t.\ \ \|\varphi(x_i) - a\|^2 \leq r^2 + \xi_i, \\ \qquad \xi_i \geq 0,\ i = 1, \ldots, n, \end{array} \tag{9}$$

where $C > 0$ is a user' parameter and $\{\xi_i\}_{i=1}^n$ are slack variables. By solving the Wolfe dual form of the problem (9):

$$\begin{array}{l} \min\limits_{\alpha_i} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) - \sum\limits_{i=1}^{n} \alpha_i k(x_i, x_i) \\ s.t.\ \sum_{i=1}^n \alpha_i = 1, \\ \qquad 0 \leq \alpha_i \leq C, i = 1, \ldots, n, \end{array} \tag{10}$$

where $\{\alpha_i\}_{i=1}^n$ are Lagrangian multipliers, it gets

$$\begin{array}{l} a = \sum_{i=1}^n \alpha_i \varphi(x_i), \\ r^2 = k(x_k, x_k) + \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \\ \qquad -2 \sum_{i=1}^n \alpha_i k(x_i, x_k), \end{array}$$

for some $\alpha_k \in (0, C)$. Consequently, the decision function can be obtained by

$$\begin{array}{rl} f(x) = & r^2 - \|\varphi(x) - a\|^2 \\ = & 2 \sum_{i=1}^n \alpha_i k(x_i, x) - 2 \sum_{i=1}^n \alpha_i k(x_i, x_k) \\ & + k(x_k, x_k) - k(x, x). \end{array}$$

The specific algorithm is as follows.

Algorithm 4. (HSOC-KSVM)

Step 1 and Step 2 are same as Algorithm 3.

Step 3. Solve the problem (10) and obtain the optimal solution $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$.

Step 4. Calculate

$$\begin{array}{l} r^{*2} = k(x_k, x_k) + \sum_{i,j=1}^n \alpha_i^* \alpha_j^* k(x_i, x_j) \\ \quad -2\sum_{i=1}^n \alpha_i^* k(x_i, x_k) \end{array}$$

for some $\alpha_k^* \in (0, C)$.

Step 5. Construct the decision function

$$\begin{array}{l} f(x) = 2\sum_{i=1}^n \alpha_i^* k(x_i, x) - 2\sum_{i=1}^n \alpha_i^* k(x_i, x_k) \\ \quad +k(x_k, x_k) - k(x, x). \end{array}$$

Step 6. For a new sample $\widetilde{x}$, it can be inferred that $\widetilde{x}$ belongs to class 1 if $f(\widetilde{x}) \geq 0$ and otherwise to class -1.

3.3 Equivalence between MMOC-KSVM and HSOC-KSVM

This subsection mainly discusses the equivalence between MMOC-KSVM and HSOC-KSVM for Gaussian RBF kernel:

$$k(x, y) = \exp(-\tfrac{1}{\sigma^2} \| x - y \|^2), \forall x, y \in R^m.$$

In this case, The problem (10) can be represented as:

$$\begin{array}{l} \min\limits_{\alpha_i} \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) - 1 \\ s.t.\ \sum_{i=1}^n \alpha_i = 1, \\ \qquad 0 \leq \alpha_i \leq C, i = 1, \ldots, n, \end{array} \quad (11)$$

because $k(x, x) = 1$ for all $x \in R^m$. By solving the problem (11), the decision function of HSOC-KSVM can be obtained by

$$f(x) = 2\sum_{i=1}^n \alpha_i^* k(x_i, x) - 2\sum_{i=1}^n \alpha_i^* k(x_i, x_k),$$

which just is the same as the decision function of MMOC-KSVM. So, MMOC-KSVM and HSOC-KSVM are equivalent for Gaussian RBF kernels.

4 OC-KSVM+

This section is devoted to study one-class classification problems with group information of data. Let $X = \{x_i\}_{i=1}^n \subset R^m$ be a target class with t groups ($t > 1$). Let $X_r = \{x_1^r, \cdots, x_{n_r}^r\}$, $T_r = \{1, \cdots, n_r\}$ and $n = \sum_{r=1}^t n_r$, then $X = \cup_{r=1}^t X_r$. The group information hidden in data is used to impose additional constraints on the slack variables in the problems (7) and (9). By means of the group information and kernel functions, each sample can be mapped into two different spaces: a decision space Z via the feature mapping $\varphi : X \to Z(z_i = \varphi(x_i))$ corresponding to a kernel k and a correcting space Z_r via the feature mapping $\varphi_r : X_r \to Z_r(z_i^r = \varphi_r(x_i^r))$ corresponding to a kernel $k_r, r \in \{1, \cdots, t\}$. By using OC-KSVM in the correcting space Z_r, t correcting functions $f_r(x), r = 1, \cdots, t$ can be gotten. The following is a detailed discussion how to impose these correcting functions onto slack variables of the problems (7) and (9).

4.1 MMOC-KSVM+

Firstly, by using MMOC-KSVM in the correcting space Z_r, we obtain t correcting functions

$$f_r(x) =< \omega_r, \varphi_r(x) > -d_r, r = 1, \cdots, t,$$

and then construct the following optimization problem by means of the idea of binary KSVM+:

$$\begin{array}{l} \min\limits_{\omega, \omega_r, \rho, d_r} \frac{1}{2}\|\omega\|^2 + \frac{v}{2}\sum_{r=1}^t \|\omega_r\|^2 - \rho \\ \qquad +C\sum_{r=1}^t \sum_{i \in T_r} \xi_i^r \\ \quad s.t.\ \langle \omega, \varphi(x_i^r) \rangle \geq \rho - \xi_i^r, \\ \qquad \xi_i^r = \langle \omega_r, \varphi_r(x_i^r) \rangle + d_r, \\ \qquad \xi_i^r \geq 0, i \in T_r, r = 1, \ldots, t, \end{array} \quad (12)$$

where $v > 0$ and $C > 0$ are user' parameters. v adjusts the relative weight of decision function and correcting function and C controls the trade-off between complexity and the number of nonseparable samples. By solving the dual problem of the problem (12):

$$\begin{array}{l} \min\limits_{\alpha_i, \beta_i} \frac{1}{2}\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) + \frac{1}{2v}\sum_{r=1}^t \sum\limits_{i \in T_r} (\alpha_i \\ \qquad +\beta_i - C)(\alpha_j + \beta_j - C) k_r(x_i, x_j) \\ s.t.\ \sum_{i=1}^n \alpha_i = 1, \\ \qquad \sum_{i \in T_r}(\alpha_i + \beta_i - C) = 0, \\ \qquad i \in T_r, r = 1, \ldots, t, \end{array} \quad (13)$$

where $\{\alpha_i\}_{i=1}^n$ and $\{\beta_i\}_{i=1}^n$ are Lagrangian multipliers, we can obtain

$$\begin{array}{l} \omega = \sum_{i=1}^n \alpha_i \varphi(x_i), \\ \rho = \sum_{i=1}^n \alpha_i k(x_i, x_k), \\ \omega_r = \sum_{i \in T_r} \alpha_i \varphi_r(x_i), \\ d_r = \sum_{i \in T_r} \alpha_i k_r(x_i, x_k), \end{array}$$

for some $\alpha_k \in (0, C)$. Proceed to the next step, the decision function

$$f(x) = \sum_{i=1}^n \alpha_i k(x_i, x) - \sum_{i=1}^n \alpha_i k(x_i, x_k).$$

The specific algorithm is as follows.

Algorithm 5. (MMOC-KSVM+)

Step 1. Given a target class $X = \{x_1, \cdots, x_n\} \subset R^m$ with group information of data and divide it into t groups $X_r = \{x_1^r, \ldots, x_{n_r}^r\}, r = 1, \ldots, t$.

Step 2. Select proper kernel functions $k : R^m \times R^m \to R$ and $k_r : R^m \times R^m \to R$ and appropriate parameters $C > 0$ and $\upsilon > 0$.

Step 3. Solve the problem (13) and obtain the optimal solution $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$.

Step 4. Calculate $\rho^* = \sum_{i=1}^n \alpha_i^* k(x_i, x_k)$ for some $\alpha_k^* > 0$.

Step 5. Construct the decision function $f(x) = \sum_{i=1}^n \alpha_i^* k(x_i, x) - \sum_{i=1}^n \alpha_i^* k(x_i, x_k)$.

Step 6. For a new sample $\widetilde{x}$, it can be inferred that $\widetilde{x}$ belongs to class 1 if $f(\widetilde{x}) \geq 0$ and otherwise to class -1.

4.2 HSOC-KSVM+

Unlike MMOC-KSVM+, The aim of HSOC-KSVM+ is to train a hyper-plane $b^2 - \|\varphi(x) - a\|^2 \geq 0$ in the feature space by means of the group information of data that contains almost all the data of the target class. The decision is given by the decision function $f(x) = b^2 - \|\varphi(x) - a\|^2$. By using HSOC-KSVM in the correcting space Z_r, t correcting functions can be obtained by

$$f_r(x) = b_r^2 - \|\varphi_r(x) - a_r\|^2, r = 1, \cdots, t,$$

where $b_r \in R$ and $a_r \in Z_r$ are the radius and center of the hyper-sphere in Z_r. Similar to the MMOC-KSVM+, we can construct the following optimization problem:

$$\begin{aligned} \min_{a,b,a_r,b_r} \ & b^2 + \upsilon \textstyle\sum_{r=1}^t b_r^2 + C \sum_{r=1}^t \sum_{i \in T_r} \xi_i^r \\ s.t. \ & \|\varphi(x_i^r) - a\|^2 \leq b^2 + \xi_i^r, \\ & \xi_i^r = b_r^2 - \|\varphi_r(x_i^r) - a_r\|^2, \\ & \xi_i^r \geq 0, i \in T_r, r = 1, \ldots, t, \end{aligned} \tag{14}$$

where $C > 0$ and $\upsilon > 0$ are user' parameters and $\{\xi_i\}_{i=1}^n$ are slack variables. Considering the Lagrangian function of the problem (14):

$$\begin{aligned} L(a, b, a_r, b_r, \alpha_i, \beta_i) &= b^2 + \upsilon \textstyle\sum_{r=1}^t b_r^2 \\ &+ C \textstyle\sum_{r=1}^t \sum_{i \in T_r} (b_r^2 - \|\varphi_r(x_i^r) - a_r\|^2) \\ &+ \textstyle\sum_{r=1}^t \sum_{i \in T_r} \alpha_i (\|\varphi(x_i^r) - a\|^2 \\ &- b^2 - b_r^2 + \|\varphi_r(x_i^r) - a_r\|^2) \\ &- \textstyle\sum_{r=1}^t \sum_{i \in T_r} \beta_i (b_r^2 - \|\varphi_r(x_i^r) - a_r\|^2), \end{aligned} \tag{15}$$

where $\{\alpha_i\}_{i=1}^n$ and $\{\beta_i\}_{i=1}^n$ are nonnegative Lagrangian multipliers, and letting $\frac{\partial L}{\partial a} = \frac{\partial L}{\partial b} = \frac{\partial L}{\partial a_r} = \frac{\partial L}{\partial b_r} = 0$, it can be deduced that

$$\begin{aligned} & \textstyle\sum_{i=1}^n \alpha_i = 1, \\ & \textstyle\sum_{i \in T_r} (\alpha_i + \beta_i - C) = \upsilon, \\ & a = \textstyle\sum_{i=1}^n \alpha_i \varphi(x_i), \\ & a_r = \frac{1}{\upsilon} \textstyle\sum_{i \in T_r} (\alpha_i + \beta_i - C) \varphi_r(x_i^r). \end{aligned} \tag{16}$$

substituting (16) into (15), the Wolfe dual form of the problem (14) can be gotten:

$$\begin{aligned} \min_{\alpha_i,\beta_i} \ & \textstyle\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) - \sum_{i=1}^n \alpha_i k(x_i, x_i) \\ & + \frac{1}{\upsilon} \textstyle\sum_{r=1}^t \sum_{i,j \in T_r} (\alpha_i + \beta_i - C)(\alpha_j \\ & + \beta_j - C) k_r(x_i, x_j) \\ & - \textstyle\sum_{r=1}^t \sum_{i \in T_r} (\alpha_i + \beta_i - C) k_r(x_i, x_i) \\ s.t. \ & \textstyle\sum_{i=1}^n \alpha_i = 1, \\ & \textstyle\sum_{i \in T_r} (\alpha_i + \beta_i - C) = \upsilon, r = 1, \ldots, t, \\ & \alpha_i, \beta_i \geq 0, i = 1, \ldots, n. \end{aligned} \tag{17}$$

By solving the problem (17), it has

$$\begin{aligned} & a = \textstyle\sum_{i=1}^n \alpha_i \varphi(x_i), \\ & b^2 = k(x_k, x_k) + \textstyle\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \\ & \quad -2 \textstyle\sum_{i=1}^n \alpha_i k(x_i, x_k), \\ & a_r = \frac{1}{\upsilon} \textstyle\sum_{i \in T_r} (\alpha_i + \beta_i - C) \varphi_r(x_i), \\ & b_r^2 = k_r(x_k, x_k) - \frac{2}{\upsilon} \textstyle\sum_{i \in T_r} (\alpha_i + \beta_i - C) k_r(x_i, x_k) \\ & + \frac{1}{\upsilon^2} \textstyle\sum_{i,j \in T_r} (\alpha_i + \beta_i - C)(\alpha_j + \beta_j - C) k_r(x_i, x_j), \end{aligned}$$

for some $\alpha_k \in (0, C)$. Consequently, the decision function can be obtained. The specific algorithm is as follows.

Algorithm 6. (HSOC-KSVM+)

Step 1 and Step 2 are the same as in Algorithm 5.

Step 3. Solve the problem (17) and obtain the optimal solution $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$.

Step 4. Calculate

$$\begin{aligned} b^{*2} &= k(x_k, x_k) + \textstyle\sum_{i,j=1}^n \alpha_i^* \alpha_j^* k(x_i, x_j) \\ & -2 \textstyle\sum_{i=1}^n \alpha_i^* k(x_i, x_k), \end{aligned}$$

for some $\alpha_k^* > 0$.

Step 5. Construct the decision function

$$\begin{aligned} f(x) &= 2 \textstyle\sum_{i=1}^n \alpha_i^* k(x_i, x) - 2 \sum_{i=1}^n \alpha_i^* k(x_i, x_k) \\ & + k(x_k, x_k) - k(x, x). \end{aligned}$$

Step 6. For a new sample $\widetilde{x}$, it can be inferred that $\widetilde{x}$ belongs to class 1 if $f(\widetilde{x}) \geq 0$ and otherwise to class -1.

5 OC-KSVM+ for multi-class classification problems

This section considers the applications of MMOC-KSVM+ and HSOC-KSVM+ for multi-class classification problems with group information of data. Let $\{(x_i, y_i)\}_{i=1}^n \subset R^m \times \{1, \cdots, c\}$ be a sample set of a $c(c \geq 3)$ class problem with $t(t \geq 2)$ groups information. Put $X = \{x_i\}_{i=1}^n$ and $Y = \{y_i\}_{i=1}^n$. Let $X^i = \{x_1^i, \cdots, x_{n_i}^i\}$ be the set of samples belonging to the ith class, $Y^i = \{y_1^i, \cdots, y_{n_i}^i\}$ the class index set corresponding to X^i and $n = \sum_{i=1}^c n_i$. Let

$X_r^i = \{x_1^{ir}, \cdots, x_{n_{ir}}^{ir}\}$ be the set of samples belonging to the ith class with rth group information, $Y_r^i = \{y_1^{ir}, \cdots, y_{n_{ir}}^{ir}\}$, $T_{ir} = \{1, \cdots, n_{ir}\}, r = 1, \cdots, t$ and $n_i = \sum_{r \in T_{ir}} n_{ir}$.

The decision function $f_i(x)$ of the ith class can be obtained by using MMOC-KSVM+ or HSOC-KSVM+. For a new sample $\widetilde{x}$, it can be inferred that $\widetilde{x}$ belongs to the jth class if $f_j(\widetilde{x}) = \max_{1 \leq i \leq c} f_i(\widetilde{x})$.

6 Experiments and results analysis

In order to test and evaluate the efficacy of MMOC-KSVM+ and HSOC-KSVM+, a series of comparative experiments with KSVM, KSVM+, OC-KSVM, OAO-KSVM and OAA-KSVM are performed on WBC, Wine and KSC three data sets, which are taken respectively from [32] and [33]. Gaussian RBF kernel $k(x, x') = \exp\{-\frac{1}{\sigma^2}\|x - x'\|^2\}$ is used in all experiments. All the accuracy in experiments is the test accuracy with five-fold cross validation method. In Tables 1-3, CA and TC denote classification accuracy and time consumed, respectively.

6.1 Experiments on WBC data set

WBC data set includes 699 instances composed of two classes (benign and malignant), each of one has 9 attributes. There are 16 instances that contain a single missing (i.e., unavailable) attribute value. We randomly select 600 instances in the rest of 683 instances for comparative experiments with KSVM, KSVM+ and OC-KSVM.

Due to the equivalence of MMOC-KSVM and HSOC-KSVM for Gaussian RBF kernel, they can be unified written as OC-KSVM and the class 'benign' is regarded as the target class. In MMOC-KSVM+ and HSOC-KSVM+, data can be separated as 3 groups by means of the attribute 'Clump Thickness': group 1 contains 170 instances with values of Clump Thickness being less than or equal to 2, group 2 contains 255 instances with values of Clump Thickness belonging to the interval $(2, 5]$ and group 3 contains 175 instances with values of Clump Thickness belonging to the interval $(5, 10]$. Kernel parameter σ is taken as 5.5 in KSVM and OC-KSVM and kernel parameters $\sigma, \sigma_1, \sigma_2$ and σ_3 are taken respectively as 5.5, 15, 1 and 0.1 in KSVM+ and as 15, 5, 10 and 20 in MMOC-KSVM+ and HSOC-KSVM+. Experiment results are shown in Table 1.

6.2 Experiments on Wine data set

Wine data set includes 178 instances composed of three classes, each of one has 13 attributes. We randomly select 45 instances from every class for comparative experiments with OAO-KSVM, OAA-KSVM and OC-KSVM.

In OC-KSVM, the labels of all the data belonging to class of interest are set to 1 and the labels of all other data are set to -1. In MMOC-KSVM+ and HSOC-KSVM+, data are separated 2 groups by means of the attribute 'Hue': group 1 contains 58 instances with values of Hue being less than or equal to 0.9 abd group 2 contains 77 instances with values of Hue being more than 0.9. The kernel parameter σ is taken respectively as 480, 480 and 400 in OAO-KSVM, OAA-KSVM and OC-KSVM and kernel parameters σ, σ_1 and σ_2 are taken respectively as 400, 320 and 480 in MMOC-KSVM+ and HSOC-KSVM+. Experiment results are lasted in Table 2.

6.3 Experiments on KSC data set

KSC data set includes 5211 instances composed of 13 classes, each of one has 176 attributes. We select 5 classes (Spartina marsh, Cattail Marsh, Salt marsh, Mud flats and water) and randomly select 100 instances from each class for comparative experiments with OAO-KSVM, OAA-KSVM and OC-KSVM.

In OC-KSVM, the labels of all the data belonging to class of interest are set to 1 and the labels of all other data are set to -1. In MMOC-KSVM+ and HSOC-KSVM+, data are separated 2 groups by means of the first attribute: group 1 contains 194 instances with values of the first attribute being less than or equal to 10 and group 2 contains 306 instances with values of the first attribute being more than 10. The kernel parameter σ is taken as 100 in OAO-KSVM, OAA-KSVM and OC-KSVM and kernel parameters σ, σ_1 and σ_2 are taken respectively as 100, 20 and 30 in MMOC-KSVM+ and HSOC-KSVM+. Experiment results are shown in Table 3.

6.4 Experiment results analysis

It can be seen from Table 1 that (1) the classification accuracy of KSVM with group information of data (KSVM+, MMOC-KSVM+ and HSOC-KSVM+) is higher than that without group information (KSVM and OC-KSVM) and the time consumed with group information is a lot more than that without group information. This indicates that group information of data can improve the classification accuracy and meanwhile increase the time consumed. (2) the classification accuracy of MMOC-KSVM+ is the highest (95.33%). (3) Although HSOC-KSVM+ is slightly lower than binary KSVM+ in classification accuracy, HSOC-KSVM+ is much faster than binary KSVM+ in time consumed. (4) HSOC-KSVM+ is about 1.6 times faster than MMOC-KSVM+, but the classifica-

Table 1: CA and TC on WBC data set

	KSVM	KSVM+	OC-KSVM	MMOC-KSVM+	HSOC-KSVM+
σ	5.5	5.5	5.5	15	15
σ_1		15		5	5
σ_2		1		10	10
σ_3		0.1		20	20
υ		10		1	1
C	0.01	0.01	0.01	0.00001	0.00001
TC(s)	50.7	648.8	12.1	202.4	126.4
CA(%)	82	92.5	90.83	95.33	91.83

Table 2: CA and TC on Wine data set

	OAO-KSVM	OAA-KSVM	OC-KSVM	MMOC-KSVM+	HSOC-KSVM+
σ	480	480	400	400	400
σ_1				320	320
σ_2				480	480
υ				2	2
C	0.1	0.1	0.1	0.1	0.1
TC(s)	2.8	5.2	0.9	1.99	1.4
CA(%)	69.63	57.78	68.15	71.85	69.63

tion accuracy of HSOC-KSVM+ is about 3.5% lower than MMOC-KSVM+.

It can be seen from Table 2 that (1) The classification accuracy of MMOC-KSVM+ is the highest (71.85%). (2) HSOC-KSVM+ is faster than MMOC-KSVM+, but is 2.22% lower than MMOC-KSVM+ for classification accuracy. (3) Although the classification accuracies of HSOC-KSVM+ and OAO-KSVM are the same (69.63%), HSOC-KSVM+ is 2 times faster than OAO-KSVM. (4) The OC-KSVM method is the fastest one in terms of time consumed.

It can be seen from Table 3 that (1) although the classification accuracies of five classifiers are almost the same, OC-KSVM, MMOC-KSVM+ and HSOC-KSVM+ are much faster than OAO-KSVM and OAA-KSVM. (2) OC-KSVM is faster than MMOC-KSVM+ and HSOC-KSVM+

In order to be more intuitive to compare the classification accuracies of seven classifiers, a histogram is provided in Figure 1.

According to the above analysis, we can conclude that the group information of data really can improve the classification accuracy of OC-KSVMs, but at the same time it increases the time consumed.

7 Conclusion

This paper mainly studies the effect of the group information of data in OC-KSVMs for classification accuracy and time consumed of multi-class classification problems, and presents two new classification methods MMOC-KSVM+ and HSOC-KSVM+. According to the experiment results, we know that the group information of data really can improve the classification accuracy of OC-KSVMs. But in MMOC-KSVM+ and HSOC-KSVM+ all the data are projected into two different spaces (decision space and correcting space), which results in the time consumed of the proposed methods is greatly increased. In addition, The selection of modeling parameters and kernel parameters can also effect the classification accuracy and time consumed of algorithms. Therefore, how to choose suitable parameters and how to develop a fast algorithm for MMOC-KSVM+ and HSOC-KSVM+ are our next study work.

References:

[1] M. Oren, et al. Pedestrain detection using wavelet templates. In proceedings of the computer vision and pattern recognition, 1997, 193-199.

[2] Martina Sattlecker, et al. Support vector machines combined with feature selection for breast cancer diagnosis. Chemometrics and Intelligent Laboratory Systems, 107, 2011, 363-370.

[3] C. Cortes, V. Vapnik. Support-vector network. Mach Learn, 20, 1995, 273-297.

[4] E. Osuna, R. Freund, F. Girosi. Training support vector machines: an application to face detection. In proceedings of the computer vision and pattern recognition, 1997, 130-136.

[5] Xiao-Xiao Niu, Ching Y. Suen. A novel hybrid CNNCSVM classifier for recognizing handwritten digits. Pattern Recognition, 45, 2012, 1318-1325.

Table 3: CA and TC on KSC data set

	OAO-KSVM	OAA-KSVM	OC-KSVM	MMOC-KSVM+	HSOC-SVM+
σ	100	100	100	100	100
σ_1				20	20
σ_2				30	30
υ				2	2
C	200	200	0.1	0.1	0.1
TC(s)	13.84	87.66	2.24	4.26	4.28
CA(%)	95.4	95.6	95.6	95.6	95.6

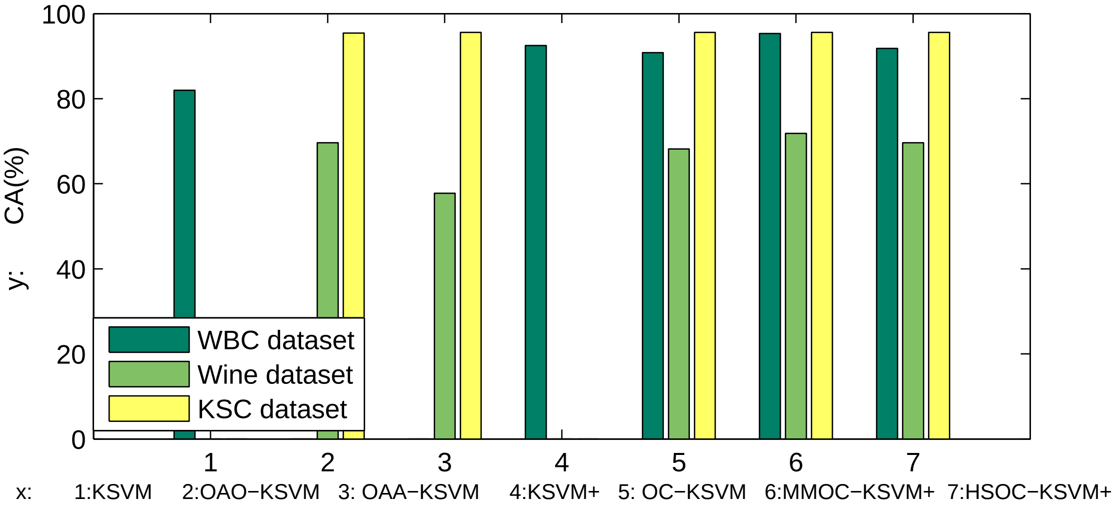

Figure 1: CA comparison of seven classfiers.

[6] V. Vapnik. Statistical learning theory. Wiley, New York, 1998.

[7] D. Tax, R. Duin. Support vector domain description. Pattern Recognition Letters, 20, 1999, 11-13.

[8] D. Tax, R. Duin. Support vector data description. Mach Learn, 54, 2004, 45-66.

[9] Ki Sang, Youn Jung Park, Kar-Ann Toh, Sangyoun Lee. SVM-based feature extraction for face recognition. Pattern Recognition, 43, 2010, 2871-2881.

[10] Minh Hoal, Fernando de la Torre. Optimal feature selection for support vector machines. Pattern Recognition, 43, 2010, 584-591.

[11] Shu Yang and Shuicheng Yan. Bilinear analysis for kernel selection and nonlinear feature extraction. IEEE Trans Neural Networks, 18(5), 2007, 1442-1452.

[12] Kuo-ping Wu, Sheng-De Wang. Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. Pattern Recognition, 42, 2009, 710-717.

[13] Pei-Yi Hao, Jung-Hsien Chiang, Yen-Hsiu Lin. A new maximal-margin spherical-structured multi-class support vector machine. Applied Intelligence, 30(2), 2009, 98-111.

[14] D. Slezak, et al. Intrusion detection system based on multi-class SVM. Lecture Notes in Computer Science, 3642, 2005, 511-519.

[15] Yi Liu, Yuan F.Zheng. One-against-all multi-class SVM classification using reliability measures. Neural Networks, 2, 2005, 849-854.

[16] Gjorgji Madzarov, Dejan Gjorgjevikj, Ivan Chorbev. A multi-class SVM classifier utilizing binary decision tree. Informatica, 33(2), 2009, 233-241.

[17] C.W. Hsu , C.J. Lin. A comparison of methods for multiclass support vector machines. IEEE Trans Neural Networks, 13, 2002, 415-425.

[18] Yong Xu, et al. A fast kernel-based nonlinear discriminant analysis for multi-class problems. Pattern Recognition, 39, 2006, 1026-1033.

[19] M. Moya, R. Hush. Network constraints and multi-objective optimization for one-class classification. Neural Networks, 9, 1996, 463-474.

[20] L. Manevitz, M. Youssef. One-class SVMs for document classification. The Journal of Machine Learning Research, 2, 2002, 139-154.

[21] A. Rabaoui, et al. Improved one-class SVM classified for sounds classification. Advanced Video and Signal Based Surveillance, IEEE, 2007, 117-122.

[22] Mireille Tohme, Redis Lengelle. Maximum margin one-class support vector machine for multiclass problems. Pattern Recognition Letters, 32, 2011, 1652-1658.

[23] Yuhua Li. Selecting training points for one-class support vector machines. Pattern Recognition Letters, 32, 2011, 1517-1522.

[24] Heng Lian. On feature selection with principal component analysis for one-class SVM. Pattern Recognition Letters, 33, 2012, 1027-1031.

[25] V. Vapnik, A. Vashist. A new learning paradigm: learning using privileged information. Neural Networks, 22(5), 2009, 544–557.

[26] Lichen Liang, Feng Cai, Vladimir Cherkassky. Predictive learning with structured (grouped) data. Neural Networks, 22, 2009, 766-773.

[27] Feng Cai, Vladimir Cherkassky. SVM+ regression and multi-task learning. Proceedings of International Joint Conference on Neural Networks, IEEE, 2009, 418-424.

[28] Lichen Liang, Vladimir Cherkassky. Connection between SVM+ and multi-task learning. IEEE International Joint Conference on Neural Networks, IEEE, 2008, 2048-2054.

[29] H. Xiong, M. Swamy, M. Ahmad. Optimizing the kernel in the empirical feature space. IEEE Transactions on Neural Networks, 16, 2005, 460-474.

[30] S.-J. Kim, A. Magnani, S. Boyd. Optimal kernel selection in kernel fisher discriminant analysis. In Proceedings of the 23rd International conference on Machine Learning, 2006, 465-472.

[31] I. Guyon, et al. Gene selection for cancer classification using support vector machines. Machine Learning, 46, 2002, 389-442.

[32] http://archive.ics.uci.edu/ml/.

[33] http://www.csr.utexas.edu/hyperspectral/data/.