IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 24, No. 5, October, 2019, pp. 463-471
Journal homepage: http://iieta.org/journals/isi

TWSVC+: Improved Twin Support Vector Machine-Based Clustering

Sanaz Moezzi, Mehrdad Jalali, Yahya Forghani*

Islamic Azad University, Mashhad branch, Mashhad 9187147578, Iran

Corresponding Author Email: yforghani@mshdiau.ac.ir

https://doi.org/10.18280/isi.240502

Received: 5 June 2019
Accepted: 24 September 2019

Keywords:
plane-based clustering, support vector clustering (SVC), twin support vector clustering (TWSVC), convex

ABSTRACT

Based on twin support vector machines (TWSVM) model, the twin support vector clustering (TWSVC) is a planar clustering model that increases inter-cluster separation. Because the TWSVC is not a standard model for some variables, its solving algorithm consumes lots of time and does not always converge to the optimal solution. To solve the problem, this paper proposes a novel clustering model, denoted as TWSVC+, based on twin support vector machines (TWSVM). The TWSVC+ is convex and standard with respect to each variable. Therefore, it is possible to solve this model rapidly with an algorithm that converges to a global optimal solution relative to each variable. The author presented linear TWSVC+ and non-linear TWSVC+ for clustering linear separable clusters and linear inseparable clusters, respectively. Experimental results on real datasets of UCI repository show that the TWSVC+ was better than TWSVC and support vector clustering (SVC) in accuracy and training time.

1. INTRODUCTION

Clustering is a main procedure for data mining with wide application such as community detection, image processing, gene analysis, text organization, etc. K-Means and its extensions [1-5] are one class of the most famous clustering models. K-Means extracts clusters such that the data of each cluster have minimum distance to its cluster center which is a point. Hence, k-means is a point-based clustering model which tries to increase intra-cluster compactness. Other main category of clustering models are plane-based clustering models, e.g. k-plane clustering (kPC) [6, 7], proximal plane clustering (PPC) [8, 9], support vector clustering (SVC) [10, 11] and twin support vector clustering (TWSVC) [12]. In plane-based clustering model, cluster center or cluster boundary is a plane. SVC [11] tries to increase inter-cluster separation. A faster version of SVC was proposed by K. Zhang, et al. [10]. Their experimental results show that the accuracy of SVC is higher than K-Means. But, when the number of data increases, memory-usage and training time of SVC increases dramatically. SVC is based on a well-known classification model named support vector machine (SVM) [13]. Training time and accuracy of Twin SVM (TWSVM) [14-16] are better than those of SVM. The existing TWSVM-based clustering, i.e. TWSVC [12], is not a standard model with respect to some of its variables. Therefore, a time-consuming algorithm named concave-convex algorithm [17] is used to solve TWSVC with respect to the mentioned variables. Moreover, the concave-convex algorithm does not guarantee a global optimal solution with respect to the mentioned variables.

This paper proposes a novel TWSVM-based clustering model called TWSVC+, which is convex and standard with respect to each model variable. Therefore, there is a fast algorithm for solving TWSVC+ with respect to each model variable which guarantees a global optimal solution of TWSVC+ with respect to each model variable. For simplicity, it assumed that we are dealing with two-cluster clustering. Therefore, the proposed model is explained based on a two-cluster TWSVM: Assuming that we have only two clusters, initial clustering is done by using k-Means and therefore, the membership degrees of data to each initial cluster are determined. Then, cluster centers are corrected using TWSVM because data labels or data membership degrees were determined, previously. Next, data labels are corrected according to cluster centers learned using TWSVM. The process of correcting cluster centers and data labels is continued until convergence. Our experiments on the real dataset of UCI repository, i.e. Sonar, Cancer, PID, Iris and Wine show that the accuracy of TWSVC+ is better than that of k-means, SVC and TWSVC, and the training time of TWSVC+ is less than that of TWSVC and SVC.

In continue, in section 2, TWSVC is explained. As our proposed clustering model is based on TWSVM, it is briefly explained in section 2, too. In section 3, TWSVC+ is proposed. In section 4, TWSVC+ is evaluated using real dataset and in section 5, the conclusion is provided.

2. PRELIMINARIES

2.1 TWSVC for clustering

Let $X = \{x_1, x_2, \dots, x_N\}$ be data which must be clustered into k clusters where $\forall i: x_i \in R^m$. TWSVC [6] assumes that the center of i-th cluster is a hyperplane with the equation $w_i^T x + b_i = 0$, where w_i is the weight vector and b_i is the bias of the mentioned hyperplane. TWSVM finds i-th cluster center or i-th hyperplane such that members of this cluster are in the vicinity of i-th hyperplane and members of other clusters, are away from this hyperplane. For this purpose, the following mathematical model is used:

$$\min_{w_i,b_i,q_i} \frac{1}{2}\|X_i w_i + b_i e\|^2 + c\tilde{e}^T q_i$$
$$\text{subject to } |\tilde{X}_i w_i + b_i \tilde{e}| \geq \tilde{e} - q_i;\ q_i \geq 0. \quad (1)$$

where, $i = 1,2,\dots,k$; X_i is i-th data cluster and $\tilde{X}_i$ is the other data clusters located on both sides of the i-th cluster center. The term $\|X_i w_i + b_i e\|^2$ of the model objective function (1) minimize the Euclidean distance of i-th cluster data to i-th cluster center. q_i is slack vector which exceptionally allows some data of $\tilde{X}_i$ not to be far enough from the i-th cluster center. These data are called outliers. Each of e and $\tilde{e}$ are vectors whose elements are equal to 1. The parameter $c \geq 0$ determines the importance of the second term of the model objective function (1) compared to its first term. If the value of c is large, less number of $\tilde{X}_i$ are allowed not to be far enough from i-th cluster center. Instead, it is possible that i-th cluster data, i.e. X_i, are not placed in the vicinity of i-th cluster center or i-th hyperplane. The model (1) is not a standard model because its first constraint is non-convex. The concave-convex algorithm which is a time-consuming algorithm is used to solve the model (1). Algorithm 1 [12] is an iterative algorithm which is used to solve the model (1) for clustering data into k cluster.

Algorithm 1. An algorithm for solving TWSVC [12], i.e. the model (1).

Input:
$X = \{x_1, x_2, \dots, x_N\}$: Data
k: The number of clusters.
thr: a threshold.
Output:
k cluster centers.

- *Step 1:* Initialized data membership degrees using k-means or randomly.
- *Step 2:* Assign each data to its nearest cluster center, and determine X_i and $\tilde{X}_i$, accordingly.
- *Step 3:* Determine or correct i-th cluster centers by solving the non-standard model (1) using concave-convex algorithm.
- *Step 4:* If the norm difference of the current cluster centers and the previous cluster centers is more than the specified threshold thr, go to step 2.

2.2 TWSVM for two class classification

In SVM, it is assumed that the borders of two data classes which are determined by a mathematical model are parallel. In TWSVM, the borders or centers of two data classes are determined by a twin model. Solving TWSVM is faster than SVM, and the accuracy of TWSVM classification is higher than SVM [8]. TWSVM model is as follows:

$$\min_{w,b,q} \frac{1}{2}\|Xw + b\tilde{e}\|^2 + ce^T q$$
$$\text{subject to } -(\tilde{X}w + be) \geq e - q;\ q \geq 0. \quad (2)$$

$$\min_{\tilde{w},\tilde{b},\tilde{q}} \frac{1}{2}\left\|\tilde{X}\tilde{w} + \tilde{b}e\right\|^2 + \tilde{c}\tilde{e}^T\tilde{q}$$
$$\text{subject to } (X\tilde{w} + \tilde{b}\tilde{e}) \geq \tilde{e} - \tilde{q};\ \tilde{q} \geq 0. \quad (3)$$

where, $w^T x + b = 0$ and $\tilde{w}^T x + \tilde{b} = 0$ are the borders or the center of each of the two data classes, and w and b are the weight vector and the bias of the first class center, and $\tilde{w}$ and $\tilde{b}$ are the weight vector and the bias of the second class center. X is the first data class and $\tilde{X}$ is the second data class. The term $\|Xw + b\tilde{e}\|^2$ in the objective function of the model (2), minimizes Euclidean distance of the first data class from the first class center. The vector q is slack vector which allows some of the second data class not to be far enough from the first class center. Such data are called outliers. Each of e and $\tilde{e}$ is a vector whose elements are equal to 1. The parameter $c \geq 0$ determines the importance of the second term of the objective function (2) compared to its first term. If the value of c is large, less number of the second data class are allowed not to be far enough from the first cluster center. Instead, it is possible that the first data class is not placed in the vicinity of the first class center. The parameter c controls the generalization ability of classifier which can increase the classifier accuracy.

The term $\left\|\tilde{X}\tilde{w} + \tilde{b}e\right\|^2$ in the objective function of the model (3), minimizes Euclidean distance of the second data class from the second class center. The vector $\tilde{q}$ is slack vector which allows some of the first data class not to be far enough from the second class center. Such data are called outliers. The parameter $\tilde{c} \geq 0$ determines the importance of the second term of the objective function (3) compared to its first term. If the value of c is large, less number of the first data class are allowed not to be far enough from the second cluster center. Instead, it is possible that the second data class is not placed in the vicinity of the second class center. Figure 1 shows how TWSVM classifies data.

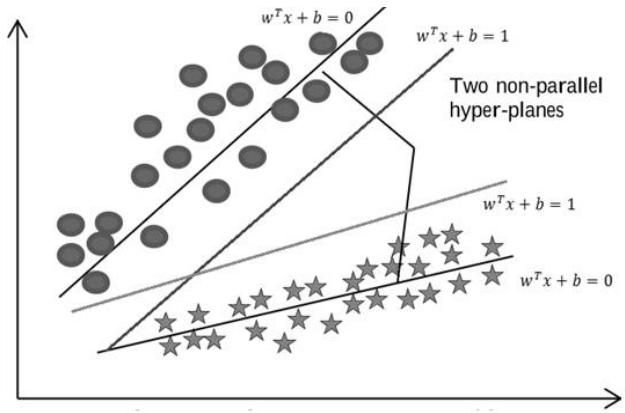


Figure 1. Classification of data using TWSVM

3. OUR PROPOSED MODEL: TWSVC+

As it was mentioned, the existing TWSVC is not a standard model with respect to some of its variables. Therefore, a time-consuming algorithm named concave-convex algorithm is used to solve TWSVC with respect to the mentioned variables. Moreover, the concave-convex algorithm does not guarantee a global optimal solution with respect to the mentioned variables. In this section, we propose a novel TWSVM-based clustering model called TWSVC+, which is convex and standard with respect to each model variable. Therefore, there is fast algorithm for solving TWSVC+ with respect to each model variable which guarantees a global optimal solution of TWSVC+ with respect to each model variable.

In two following sub-sections, linear TWSVC+ and non-linear TWSVC+ is presented for clustering linear separable clusters and linear inseparable clusters, respectively.

3.1 Linear TWSVC+

TWSVC+ is based on the basic TWSVM. The basic TWSVM is a two-class classifier which learns a classifier based on a two-class training data. One can design a multi-class classifier by using some two-class TWSVM models, e.g. DAG approach. In this paper, for simplicity, it is assumed that we want to cluster data into two clusters. Therefore, the proposed model is described based on a two-class TWSVM. Obviously, similarly, by combining several two-cluster clustering model, we can create a multi-cluster clustering model.

Suppose that we want to group $\{x_1, x_2, \dots, x_N\}$ into two clusters n and p. Initialize cluster centers and data membership degree $U_{ip} \in \{0,1\}$ and $U_{in} \in \{0,1\}$, i.e. membership degree of x_i to clusters p and n, respectively, by using k-means or randomly. Then, cluster centers can be determined using TWSVM model because data labels or data membership degrees were determined, previously. To be more precise, to determine the two cluster centers, it is enough to solve the following twin models:

$$\begin{aligned} &\min_{w,b,q} \sum_{i=1}^{N} (w^T x_i + b)^2 U_{ip} + c \sum_{i=1}^{N} q_i U_{in} \\ &\text{subject to } \begin{cases} -(w^T x_i + b) \geq 1 - q_i, \\ q_i \geq 0, \ \ i = 1,2,\dots,N. \end{cases} \end{aligned} \tag{4}$$

$$\begin{aligned} &\min_{\tilde{w},\tilde{b},\tilde{q}} \sum_{i=1}^{N} \left(\tilde{w}^T x_i + \tilde{b}\right)^2 U_{in} + \tilde{c} \sum_{i=1}^{N} \tilde{q}_i U_{ip} \\ &\text{subject to } \begin{cases} \left(\tilde{w}^T x_i + \tilde{b}\right) \geq 1 - \tilde{q}_i, \\ \tilde{q}_i \geq 0, \ \ i = 1,2,\dots,N. \end{cases} \end{aligned} \tag{5}$$

The first model, i.e. model (4), learns the center of cluster p or a hyperplane with the equation $w^T x + b = 0$ such that the data of cluster n are located behind the hyperplane with an appropriate distance from it. In other word, for each x of the cluster n we must almost have $-(w^T x + b) \geq 1$, and for each x of the cluster p we must have $w^T x + b = 0$ which is achieved by minimizing $(w^T x + b)^2$ in the model (4). The variable b is the bias and w is the weight vector of the center of cluster p. q_i is slack variable which allows x_i of the cluster n not to be in the proper distance behind the center of class p. Such data are called outliers. The parameter c controls the number of outliers of class n.

The second model, i.e. the model (5), learns the center of cluster n or a hyperplane with the equation $\tilde{w}^T x + \tilde{b} = 0$ such that the data of cluster p are located in front of the hyperplane with an appropriate distance from it. In other word, for each x of the cluster p we must almost have $\left(\tilde{w}^T x_i + \tilde{b}\right) \geq 1$, and for each x of the cluster n we must have $\tilde{w}^T x + \tilde{b} = 0$ which is achieved by minimizing $\left(\tilde{w}^T x + \tilde{b}\right)^2$ in the model (5). The variable $\tilde{b}$ is the bias and $\tilde{w}$ is the weight vector of the center of cluster n. $\tilde{q}_i$ is slack variable which allows x_i of the cluster p not to be in the proper distance in front of the center of class n. Such data is called outlier. Parameter c controls the number of outliers of class p.

Each of the models (4) and (5) is a standard model, i.e. Quadratic Programming Problem (QPP). There exist efficient and well-known algorithms to solve a QPP. Solving the duals of these two primal models, i.e. models (4) and (5), are faster than solving the primal models because their dual models has less variables and less constraints. Hence, in continue, the duals of each of these primal models is determined. The model (4) can be written as follows:

$$\begin{aligned} &\min_{w,b,q} \tfrac{1}{2}(Xw + eb)^T \mathrm{diag}(U_p)(Xw + eb) + ce^T\left(\mathrm{diag}(U_n)\right)q \\ &\text{subject to } \begin{cases} -(Xw + eb) \geq e - q, \\ q \geq 0. \end{cases} \end{aligned} \tag{6}$$

where, $X = (x_1, x_2, \dots, x_N)^T$; $e \in R^N$ is a vector whose elements are equal to 1; $U_p = \left(U_{1p}, U_{2p}, \dots, U_{Np}\right)^T$; $U_n = (U_{1n}, U_{2n}, \dots, U_{Nn})^T$; and $q = (q_1, q_2, \dots, q_N)^T$. The model (6) can be written as follows:

$$\begin{aligned} &\min_{w,b,q} \tfrac{1}{2}\left\| \sqrt{\mathrm{diag}(U_p)}Xw + \sqrt{\mathrm{diag}(U_p)}\, eb \right\|^2 \\ &\qquad + ce^T\left(\mathrm{diag}(U_n)\right)q \\ &\text{subject to } \begin{cases} -(Xw + eb) \geq e - q, \\ q \geq 0. \end{cases} \end{aligned} \tag{7}$$

Let $\dot{X} = \sqrt{\mathrm{diag}(U_p)}X$, $\dot{e} = \sqrt{\mathrm{diag}(U_p)}\, e$, $\ddot{e} = \mathrm{diag}(U_n)e$, and $\mathrm{diag}(U_n) = \mathrm{diag}(1 - U_p)$. Then, the model (7) can be rewritten as follows:

$$\begin{aligned} &\min_{w,b,q} \tfrac{1}{2}\left\|\dot{X}w + \dot{e}b\right\|^2 + c\ddot{e}^T q \\ &\text{subject to } \begin{cases} -(Xw + eb) \geq e - q, \\ q \geq 0. \end{cases} \end{aligned} \tag{8}$$

Lagrange function of the model (8) is as follows:

$$\mathcal{L} = \tfrac{1}{2}\left\|\dot{X}w + \dot{e}b\right\|^2 + c\ddot{e}^T q - \alpha^T(-(Xw + eb) + q - e) - \beta^T q, \tag{9}$$

where, $\alpha \geq 0$ and $\beta \geq 0$ are Lagrange coefficients. We have at the optimal point of the dual model:

$$\frac{d\mathcal{L}}{dw} = 0 \rightarrow \dot{X}^T\left(\dot{X}w + \dot{e}b\right) + X^T\alpha = 0; \tag{10}$$

$$\frac{d\mathcal{L}}{db} = 0 \rightarrow \dot{e}^T\left(\dot{X}w + \dot{e}b\right) + e^T\alpha = 0; \tag{11}$$

$$\frac{d\mathcal{L}}{dq} = 0 \rightarrow c\ddot{e} - \alpha - \beta = 0; \tag{12}$$

According to Eq. (12) and given that $\beta \geq 0$ and $\alpha \geq 0$, we have:

$$0 \leq \alpha \leq c\ddot{e}. \tag{13}$$

By combining Eq. (10) and Eq. (11) we obtain:

$$\left[\dot{X}^T, \dot{e}^T\right]\left[\dot{X}, \dot{e}\right][w, b] + [X^T, e^T]\alpha = 0. \tag{14}$$

We define $H = \left[\dot{X}, \dot{e}\right]$, $G = [X, e]$, and $V = [w, b]^T$. Then, Eq. (14) can be written as follows:

$$H^T H V + G^T \alpha = 0. \tag{15}$$

Therefore,

$$V = -(H^T H)^{-1} G^T \alpha. \tag{16}$$

The matrix $H^T H$ may be not invertible. In such situation, a

small positive value (λ) is added to its main diagonal. Then, we have:

$$V = -(H^T H + \lambda I)^{-1} G^T \alpha. \tag{17}$$

Moreover, by substituting Eq. (10)-(12) in the Lagrange function (9), the dual of the first model of the proposed twin models can be obtained as follows:

$$\begin{aligned} &\max_{\alpha} -\frac{1}{2}\,\alpha^T G(H^T H)^{-1} G^T \alpha + e^T \alpha \\ &\text{subject to } \ 0 \le \alpha \le c\ddot{e}. \end{aligned} \tag{18}$$

By solving the model (18), and obtaining optimal α and putting it in Eq. (16) or Eq. (17), the value of V, i.e. the weight vector w and the bias b of center of cluster p, is obtained.

Similarly, by setting the gradient of Lagrange function of the second model of the proposed twin model, i.e. the model (5), equal to zero, we obtain:

$$\tilde{V} = (\tilde{H}^T \tilde{H})^{-1} G^T \tilde{\alpha}. \tag{19}$$

where, $\tilde{\alpha}$ is Lagrange coefficient; $\tilde{V} = [\tilde{w} \cdot \tilde{b}]^T$; $\tilde{H} = [\tilde{X}, \bar{e}]$; $\tilde{X} = \sqrt{\text{diag}(U_n)}X$; $\bar{e} = \sqrt{\text{diag}(U_n)}\, e$; and $\bar{\bar{e}} = \text{diag}(U_p)\, e$. Similarly, the dual of the model (5) can be obtained:

$$\begin{aligned} &\max_{\tilde{\alpha}} \ -\frac{1}{2}\,\tilde{\alpha}^T \tilde{H}(G^T G)^{-1} \tilde{H}^T \tilde{\alpha} + e^T \tilde{\alpha} \\ &\text{subject to } \ 0 \le \tilde{\alpha} \le \tilde{c}\bar{\bar{e}}. \end{aligned} \tag{20}$$

By solving the model (20), and obtaining optimal $\tilde{\alpha}$ and putting it in Eq. (19), the value of $\tilde{V}$, i.e. the weight vector $\tilde{w}$ and the bias $\tilde{b}$ of center of class n, is obtained.

After determining the cluster centers using the models (18) and (20), the membership degrees of data to the clusters, or the values of U_{ip} and U_{in} must be modified such that the two clusters of data are as most distinctive as possible. That is, as far as possible, data of cluster p must be on the first hyperplane with the equation $w^T x + b = 0$, namely $\sum_{i=1}^{N}(w^T x_i + b)^2 U_{ip}$ must be minimized, while the data of cluster n must be located far behind the first hyperplane. Also, as far as possible, data of cluster n must be on the second hyperplane with the equation $\tilde{w}^T x + \tilde{b} = 0$, namely $\sum_{i=1}^{N}(\tilde{w}^T x_i + \tilde{b})^2 U_{in}$ must be minimized, while the data of cluster p must be located far in front of the second hyperplane, and the number of clusters outliers must be reduced, or almost equivalently $\sum_{i=1}^{N} q_i U_{in}$ and $\sum_{i=1}^{N} \tilde{q}_i U_{ip}$ must be minimized. To this end, the following model is proposed:

$$\begin{aligned} \min_{U} \ & \sum_{i=1}^{N}(w^T x_i + b)^2 U_{ip} + \sum_{i=1}^{N} q_i U_{in} \\ & + \sum_{i=1}^{N}(\tilde{w}^T x_i + \tilde{b})^2 U_{in} + \sum_{i=1}^{N} \tilde{q}_i U_{ip} \\ \text{subject to } & \begin{cases} U_{ip} + U_{in} = 1, \\ U_{ip}, U_{ip} \in \{0,1\}, \ i = 1,2,\dots,N, \\ -L \le \sum_{i=1}^{N} u_{ip} - \sum_{i=1}^{N} u_{in} \le L. \end{cases} \end{aligned} \tag{21}$$

The constraint $U_{ip} + U_{in} = 1$ states that x_i can be the member of only one cluster. The last constraint which is similar to the constraint of SVC doesn't allow all data are assigned to one cluster. This constraint states that the difference between the number of data of clusters n and p must not be higher than a predefined parameter denoted by L. For the moment, we do not consider the last constraint of the model (21). Note that $(w^T x_i + b)^2 + \tilde{q}_i$ is the coefficient of U_{ip} and $(\tilde{w}^T x_i + \tilde{b})^2 + q_i$ is the coefficienct of U_{in}, and according to the constraints of the model (21), $U_{in} = 1$ exclusive or $U_{ip} = 1$. Therefore, to minimize the objective function (21), U_{ip} must be set to 1 if its coefficient is smaller than the coefficient of U_{in}. In other word,

$$\begin{aligned} U_{ip} &= \begin{cases} 1 & (w^T x_i + b)^2 + \tilde{q}_i \le (\tilde{w}^T x_i + \tilde{b})^2 + q_i \\ 0 & \text{otherwise.} \end{cases} \\ U_{in} &= 1 - U_{ip}. \end{aligned} \tag{22}$$

After determining the membership degrees using Eq. (22), if the last constraint of the model (21) is not satisfied, namely if $\sum_{i=1}^{N} u_{ip} - \sum_{i=1}^{N} u_{in} > L$, some data of cluster p must migrate to cluster n, and if $\sum_{i=1}^{N} u_{ip} - \sum_{i=1}^{N} u_{in} < -L$, some data of cluster n must migrate to cluster p to satisfy the last constraint. Of course, since the objective function (21) must be minimized, the data with the least affect on increasing the objective function (21) must migrate. For this purpose, the algorithm 2 is suggested.

Algorithm 2: Migration of data of cluster p to cluster n or vice versa for satisfying the last constraint of the model (21).

$\text{Diff} = \left[(w^T x_i + b)^2 + \tilde{q}_i\right] - \left[(\tilde{w}^T x_i + \tilde{b})^2 + q_i\right]$
If $\sum_{i=1}^{N} u_{ip} - \sum_{i=1}^{N} u_{in} > L$ then
  $\text{Diff} = \text{Diff} < 0$
  $\text{Diff} = \text{sort}(\text{Diff}, 'descending')$
  $\forall i \in \left\{1 : \frac{\sum_{i=1}^{N} u_{ip} - \sum_{i=1}^{N} u_{in} - L}{2}\right\} : U_{in} = 1, U_{ip} = 0.$
Elseif $\sum_{i=1}^{N} u_{ip} - \sum_{i=1}^{N} u_{in} < -L$ then
  $\text{Diff} = \text{Diff} \ge 0$
  $\text{Diff} = \text{sort}(\text{Diff}, 'ascending')$
  $\forall i \in \left\{1 : \frac{\sum_{i=1}^{N} u_{ip} - \sum_{i=1}^{N} u_{in} - L}{2}\right\} : U_{in} = 1, U_{ip} = 0.$
End if

After determining the membership degrees using equation (22) and using algorithm 2, if the current membership degrees is different from the previous membership degrees, the center of each cluster must be corrected using models (18) and (20), and then this process must be continued until convergence. In fact, the proposed iterative algorithm tries to solve the following models:

$$\begin{aligned} \min_{\substack{w,b,q,\\ \tilde{w},\tilde{b},\tilde{q},U}} \ & \tfrac{1}{2}\sum_{i=1}^{N}(w^T x_i + b)^2 U_{ip} + c\sum_{i=1}^{N} q_i U_{in} \\ & + \tfrac{1}{2}\sum_{i=1}^{N}(\tilde{w}^T x_i + \tilde{b})^2 U_{in} + \tilde{c}\sum_{i=1}^{N} \tilde{q}_i U_{ip} \\ \text{subject to } & \begin{cases} -(w^T x_i + b) \ge 1 - q_i, \\ q_i \ge 0, \ i = 1,2,\dots,N, \\ (\tilde{w}^T x_i + \tilde{b}) \ge 1 - \tilde{q}_i, \\ \tilde{q}_i \ge 0, \ i = 1,2,\dots,N, \\ U_{ip} + U_{in} = 1, \\ U_{ip}, U_{ip} \in \{0,1\}, \ i = 1,2,\dots,N, \\ -L \le \sum_{i=1}^{N} u_{ip} - \sum_{i=1}^{N} u_{in} \le L. \end{cases} \end{aligned} \tag{23}$$

If membership degrees are considered to be fixed, the model (23) is transformed into the models (4) and (5), and if the clusters centers are considered to be fixed, the model (23) is transformed into the model (21). Our proposed algorithm for solving the model (23) can be summarized as algorithm 3.

Algorithm 3. An algorithm for solving TWSVC+, i.e. the model (23).

Input:

$S = \{x_1, x_2, \dots, x_N\}$: data

Output:

The members of each of two clusters.

- Step 1: Initialize membership degrees using K-Means or randomly.
- Step 2: Determine cluster centers using the models (18) and (20).
- Step 3: Determine membership degrees using Eq. (22) and then algorithm 2.
- Step 4: If the current membership degrees are different from the previous membership degrees, go to step 2.

3.2 Non-linear TWSVC+

If clusters are not linear separable, no hyperplane can be found to separate the clusters. To address this problem, first, data are mapped into a high-dimensional space using a mapping function denoted by φ. The mapping function φ is selected such that the data is linear separable in the high dimensional space. We have

$$w = \sum_{i=1}^{N} s_i \varphi(x_i),$$
$$\tilde{w} = \sum_{i=1}^{N} \tilde{s}_i \varphi(x_i),$$

where, $s_i, \tilde{s}_i \in R$. Thus, the first and the second cluster centers with the equations $w^T x + b = 0$ and $\tilde{w}^T x + \tilde{b} = 0$ can be written as follows:

$$k(x_i, X)s + b = 0,$$
$$k(x_i, X)\tilde{s} + b = 0,$$

where, $k(x, z) = \phi^T(x)\phi(z)$ is a kernel function. Thus, the model (4) and (5) can be restated as follows:

$$\min_{w,b,q} \frac{1}{2}\sum_{i=1}^{N}(k(x_i, X)s + b)^2 U_{ip} + c\sum_{i=1}^{N} q_i U_{in} \quad \text{s.t.} \begin{cases} -(k(x_i, X)s + b) \geq 1 - q_i, \\ q_i \geq 0, i = 1,2, \dots, N. \end{cases} \tag{24}$$

$$\min_{\tilde{w},\tilde{b},\tilde{q}} \frac{1}{2}\sum_{i=1}^{N}\left(k(x_i, X)\tilde{s} + \tilde{b}\right)^2 U_{in} + \tilde{c}\sum_{i=1}^{N} \tilde{q}_i U_{ip} \quad \text{s.t.} \begin{cases} \left(k(x_i, X)\tilde{s} + \tilde{b}\right) \geq 1 - \tilde{q}_i, \\ \tilde{q}_i \geq 0, i = 1,2, \dots, N. \end{cases} \tag{25}$$

The models (24) and (25) are also QPPs. Hence, there are efficient algorithms to solve them. Solving the duals of these primal models, i.e. models (24) and (25), are faster than solving the primal models because their dual models has less variables and less constraints. Hence, in continue, the duals of each of these primal models is determined. The model (25) can be written as follows:

$$\min_{w,b,q^2} \frac{1}{2}\left\|\sqrt{\text{diag}(U_p)}k(X, X)s + \sqrt{\text{diag}(U_p)}\, eb\right\|^2 + ce^T(\text{diag}(U_n))q + ce^T\text{diag}(U_n)q \quad \text{subject to } -(k(X, X)s + eb) \geq e - q, q > 0. \tag{26}$$

Let $\dot{k}(X, X) = \sqrt{\text{diag}(U_p)}k(X, X)$; $\dot{e} = \sqrt{\text{diag}(U_p)}\, e$; and $\ddot{e} = \text{diag}(U_n)e$. Thus, the model (26) can be written as follows:

$$\min_{w,b,q} \frac{1}{2}\left\|\dot{k}(X, X)s + \dot{e}b\right\|^2 + c\ddot{e}^T q \quad \text{subject to } -(k(X, X)s + eb) \geq e - q, q \geq 0. \tag{27}$$

Similarly, by setting the gradient of the Lagrange function of model (27) equal to zero, we obtain:

$$\bar{V} = -(\bar{H}^T\bar{H})^{-1}\bar{G}^T\bar{\alpha}. \tag{28}$$

where, $\bar{\alpha}$ is Lagrange coefficient, $\bar{V} = [s, b]^T$, $\bar{H} = \left[\dot{k}(X, X), \dot{e}\right]$, and $\bar{G} = [k(X, X), e]$. Similarly, the duals of the model (27) can be obtained:

$$\max_{\bar{\alpha}} -\frac{1}{2}\bar{\alpha}^T\bar{G}(\bar{H}^T\bar{H})^{-1}\bar{G}^T\bar{\alpha} + e^T\bar{\alpha} \quad \text{subject to } 0 \leq \bar{\alpha} \leq c\ddot{e}. \tag{29}$$

The matrix $\bar{H}^T\bar{H}$ may be not invertible. In such situation, a small positive value (λ) is added to its main diagonal. In this case, we have:

$$\bar{V} = -(\bar{H}^T\bar{H} + \lambda I)^{-1}\bar{G}^T\bar{\alpha}. \tag{30}$$

After solving the dual model (29) and obtaining optimal $\bar{\alpha}$ and putting it into Eq. (30), $\bar{V}$, i.e. the weight vector s and the bias b of the class center p, is obtained.

Similarly, by setting the gradient of Lagrange function of the model (25) equal to zero, we obtain

$$\bar{\bar{V}} = -(\bar{\bar{H}}^T\bar{\bar{H}})^{-1}\bar{G}^T\bar{\bar{\alpha}}. \tag{31}$$

where, $\bar{\bar{\alpha}}$ is Lagrange coefficient, $\bar{\bar{V}} = [\tilde{s}, \tilde{b}]^T$, $\bar{\bar{H}} = \left[\ddot{k}(X, X), \bar{\bar{e}}\right]$, and $\ddot{k}(X, X) = \sqrt{\text{diag}(U_n)}k(X, X)$. Similarly, the dual of the model (25) can be obtained:

$$\max_{\bar{\bar{\alpha}}} -\frac{1}{2}\bar{\bar{\alpha}}^T\bar{G}(\bar{\bar{H}}^T\bar{\bar{H}})^{-1}\bar{G}^T\bar{\bar{\alpha}} + e^T\bar{\bar{\alpha}} \quad \text{subject to } 0 \leq \bar{\bar{\alpha}} \leq \tilde{c}\bar{\bar{e}}. \tag{32}$$

where, $\bar{\bar{e}} = \text{diag}(U_p)\, e$. The matrix $\bar{\bar{H}}^T\bar{\bar{H}}$ may be not invertible. In such situation, a small positive value (λ) is added to its main diagonal. In this case, we have:

$$\bar{\bar{V}} = -\left(\bar{\bar{H}}^T\bar{\bar{H}} + \lambda I\right)^{-1}\bar{G}^T\bar{\bar{\alpha}}. \tag{33}$$

After solving the dual model (32), and obtaining the optimal $\bar{\bar{a}}$ and putting it into Eq. (33), $\bar{\bar{V}}$, i.e. the weight vector $\tilde{s}$ and the bias $\tilde{b}$ of the cluster center n, is obtained.

4. EXPERIMENTAL RESULTS

In this section, the proposed model called TWSVC+ is compared with k-means, linear SVC [10], nonlinear SVC [10] and TWSVC [12]. We used Gaussian kernel function in the nonlinear clustering models. The parameter of Gaussian kernel function is σ. The parameters c and $\tilde{c}$ were considered to be the same in this paper. The optimal value of these parameters, i.e. σ and c, were chosen by grid search method from the sets

$\{0.1,0.2,\ldots,2\}$, $\{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$, respectively. It should be noted that each of the eight real datasets, i.e. Iris, Wine, Sonar, Cancer, PID, Ecoli, Haberman, and Parkinsons were normalized. For random initialization in k-means, each experiment was repeated 20 times and the average of results was reported in Table 1. As can be seen, the accuracy of TWSVC+ is a bit better than SVC because the accuracy of TWSVM is a bit better than SVM [8], and TWSVC+ is TWSVM-based model and SVC is SVM-based model.

The accuracy of TWSVC+ is much better than TWSVC because TWSVC+ is a standard model with respect to each variable, while TWSVC is not a standard model with respect to some variables. The concave-convex algorithm is used to solve TWSVC with respect to the mentioned variables. The concave-convex algorithm does not guarantee a global optimal solution with respect to the mentioned variables.

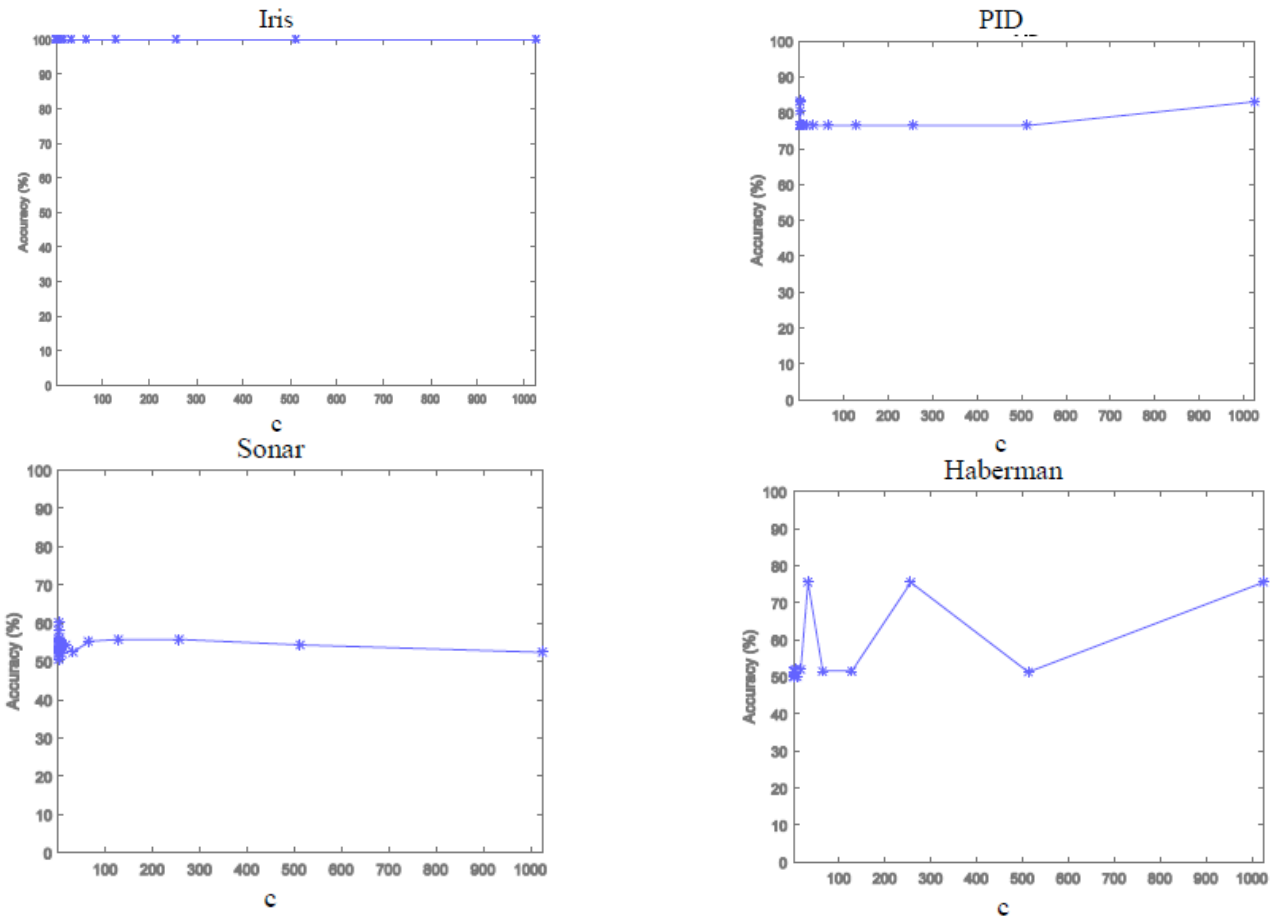
Table 2 shows training times. As it can be seen, k-means has the least training time, and TWSVC+ has the second least training time. The training time of TWSVC+ is less than that of SVC because the training time of TWSVM is less than SVM [8]. The training time of TWSVC+ is less than that of TWSVC because concave-convex algorithm which is used to solve TWSVC is too time consuming.

Table 1. The accuracy of clustering models (%)

	Iris	**Wine**	**Sonar**	**Cancer**	**PID**	**Ecoli**	**Haberman**	**Parkinsons**	**Mean**
TWSVC (linear)	100	54.57	52.88	94.69	65.10	71.65	74.83	59.11	71.60
TWSVC (Non-linear)	100	65.73	52.40	95.70	76.43	73.81	51.63	65.63	72.67
SVC (linear)	96.00	95.51	66.73	97.13	79.01	95.17	74.72	85.09	86.17
SVC (Non-linear)	100	95.51	66.81	97.13	76.02	94.37	75.98	84.37	86.27
K-Means	100	54.49	54.80	97.70	82.29	73.51	51.96	86.36	75.14
TWSVC+ (linear)	100	94.28	57.40	95.55	83.48	98.21	75.49	86.36	**86.35**
TWSVC+ (Non-linear)	100	89.93	69.85	94.09	87.67	95.96	71.43	84.09	**86.63**

Table 2. Training time of clustering models (second)

	Iris	**Wine**	**Sonar**	**Cancer**	**PID**	**Ecoli**	**Haberman**	**Parkinsons**	**Mean**
TWSVC (linear)	0.81	1.90	2.89	46.62	102.14	5.67	7.62	1.45	21.14
TWSVC (Non-linear)	1.74	2.35	4.89	103.15	204.54	9.70	16.03	2.54	43.12
SVC (linear)	0.73	1.26	1.57	45.09	26.52	4.89	4.80	0.22	10.64
SVC (Non-linear)	0.68	1.52	2.13	88.18	37.03	.470	3.60	0.18	17.25
K-Means	0.02	0.07	0.03	0.02	0.04	0.01	0.01	0.01	**0.03**
TWSVC+ (linear)	0.52	0.87	1.01	14.18	22.16	2.50	3.33	0.41	5.62
TWSVC+ (Non-linear)	0.87	1.02	1.91	30.11	31.41	2.92	2.52	0.37	8.89



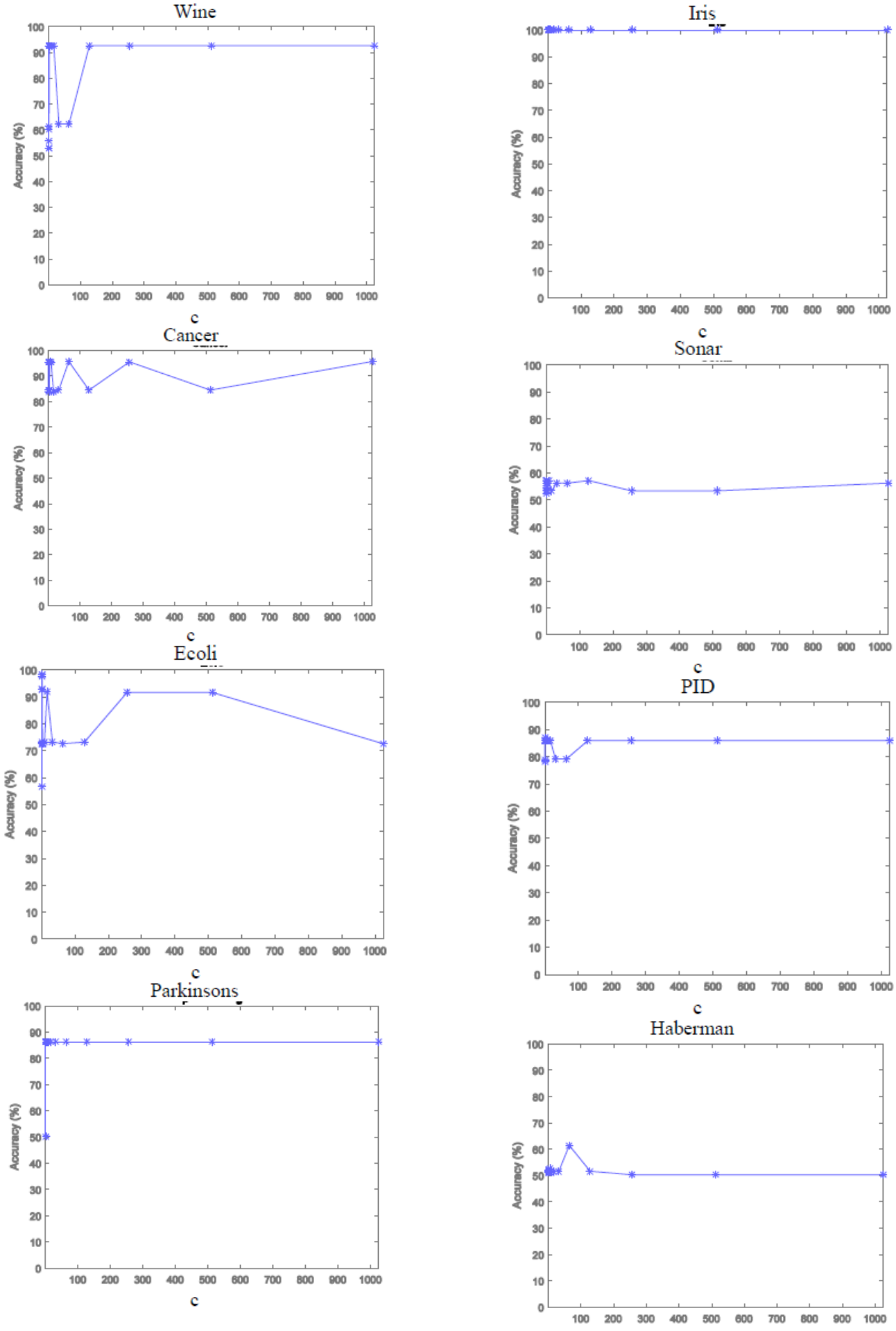


Figure 2. Sensitivity of linear TWSC+ to the hyper-parameter c

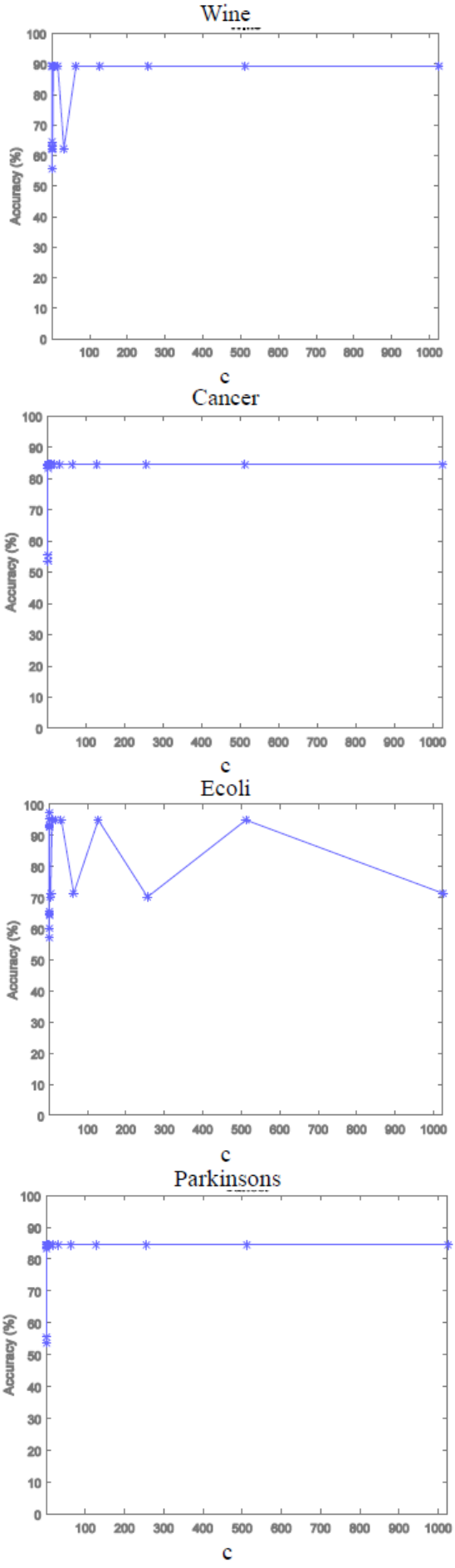


Figure 3. Sensitivity of non-linear TWSC+ to the hyper-parameter c

Figure 2 and Figure 3 depict the sensitivity of linear TWSVC+ and non-linear TWSVC+ to the hyper-parameter c, respectively. As it can be seen, the sensitivity of TWSVC+ to the hyper-parameter c depends on input dataset. For some datasets, when the hyper-parameter c changes the accuracy does not changes; for some datasets, when the hyper-parameter c changes for a wide range the accuracy does not changes; and for some datasets, when the hyper-parameter c changes the accuracy changes.

5. DISCUSSION AND CONCLUSION

(1) In this paper, a TWSVM-based clustering model called TWSVC+ was proposed which is an improved version of TWSVC. Our experiments on five real datasets showed that TWSVC+ has the highest accuracy compared with SVC, TWSVC and k-means, and has the less training time compared with SVC and TWSVC.

(2) The accuracy and training time of TWSVC+ is better than SVC because SVC is SVM-based clustering model while TWSVC+ is TWSVM-based clustering model, and training time and accuracy of TWSVM is better than those of SVM [8].

(3) The accuracy and training time of TWSVC+ is better than TWSVC because TWSVC is not a standard model with respect to some of its variables while TWSVC+ is a standard model (QPP) with respect to the mentioned variables. A time-consuming algorithm named concave-convex algorithm is used to solve the non-standard model TWSVC with respect to the mentioned variables while there are efficient algorithms to solve TWSVC+ with respect to the mentioned variables. Moreover, the concave-convex algorithm does not guarantee a global optimal solution with respect to the mentioned variables while QPP does.

REFERENCES

[1] Huang, X., Ye, Y., Zhang, H. (2014). Extensions of kmeans-type algorithms: a new clustering framework by integrating intracluster compactness and intercluster separation. IEEE Transactions on Neural Networks and Learning Systems, 25(8): 1433-1446. https://doi.org/10.1109/TNNLS.2013.2293795

[2] Huang, X., Yang, X., Zhao, J., Xiong, L., Ye, Y. (2018). A new weighting k-means type clustering framework with an l2-norm regularization. Knowledge-Based Systems, 151: 165-179. https://doi.org/10.1016/j.knosys.2018.03.028

[3] Yao, M., Wu, Q., Li, J., Huang, T. (2016). K-walks: clustering gene-expression data using a K-means clustering algorithm optimised by random walks. International Journal of Data Mining and Bioinformatics, 16(2): 121-140. https://doi.org/10.1504/IJDMB.2016.080039

[4] Cui, X., Wang, F. (2015). An improved method for K-means clustering. 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, pp. 756-759. https://doi.org/10.1109/CICN.2015.154

[5] Suryavanishi, A.S., Gujra, A.D. (2016). A new framework for k-means algorithm by combining the dispersions of clusters. International Journal of Advance Research in Computer Science and Management Studies,

4(1).
[6] Bradley, P.S., Mangasarian, O.L. (2000). K-plane clustering. Journal of Global Optimization, 16(1): 23-32. https://doi.org/10.1023/A:1008324625522
[7] Tabatabaei-Pour, M., Salahshoor, K., Moshiri, B. (2006). A modified k-plane clustering algorithm for identification of hybrid systems. 2006 6th World Congress on Intelligent Control and Automation, Dalian, China. https://doi.org/10.1109/WCICA.2006.1712564
[8] Shao, Y.H., Bai, L., Wang, Z., Hua, X.Y., Deng, N.Y. (2013). Proximal plane clustering via eigenvalues. Procedia Computer Science, 17: 41-47. https://doi.org/10.1016/j.procs.2013.05.007
[9] Liu, L.M., Guo, Y.R., Wang, Z., Yang, Z.M., Shao, Y.H. (2017). K-proximal plane clustering. International Journal of Machine Learning and Cybernetics, 8(5): 1537-1554. https://doi.org/10.1007/s13042-016-0526-y
[10] Zhang, K., Tsang, I.W., Kwok, J.T. (2009). Maximum margin clustering made practical. IEEE Transactions on Neural Networks, 20(4): 583-596. https://doi.org/10.1109/TNN.2008.2010620
[11] Xu, L., Neufeld, J., Larson, B., Schuurmans, D. (2005). Maximum margin clustering. In Advances in Neural Information Processing Systems, pp. 1537-1544.
[12] Wang, Z., Shao, Y.H., Bai, L., Deng, N.Y. (2015). Twin support vector machine for clustering. IEEE transactions on Neural Networks and Learning Systems, 26(10): 2583-2588. https://doi.org/10.1109/TNNLS.2014.2379930
[13] Gunn, S.R. (1998). Support vector machines for classification and regression. ISIS Technical Report, 14(1): 5-16.
[14] Khemchandani, R., Chandra, S. (2007). Twin support vector machines for pattern classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(5): 905-910. https://doi.org/10.1109/TPAMI.2007.1068
[15] Tian, Y., Qi, Z., Ju, X., Shi, Y., Liu, X. (2014). Nonparallel support vector machines for pattern classification. IEEE Transactions on Cybernetics, 44(7): 1067-1079. https://doi.org/10.1109/TCYB.2013.2279167
[16] Chen, S.G., Wu, X.J. (2018). A new fuzzy twin support vector machine for pattern classification. International Journal of Machine Learning and Cybernetics, 9(9): 1553-1564. https://doi.org/10.1007/s13042-017-0664-x
[17] Yuille, A.L., Rangarajan, A. (2002). The concave-convex procedure (CCCP). In Advances in Neural Information Processing Systems, pp. 1033-1040. https://doi.org/10.1162/08997660360581958

NOMENCLATURE

X	training set
N	the number of training data
x_i	i-th training data
X_i	i-th data cluster
$\bar{X}_i$	the other data clusters located on both sides of the i-th cluster center
w_i	weight vector of i-th cluster
b_i	bias of i-th cluster
k	the number of cluster
thr	threshold parameter
e, $\tilde{e}$	vectors of which elements are equal to 1
q, $\tilde{q}$	slack vectors
U_{ip}	membership degree of x_i to clusters p
w, $\tilde{w}$	weight vectors of hyperplanes
b, $\tilde{b}$	biases of hyperplanes
λ	a small positive scalar
L	A threshold parameter
α, β	lagrange coefficients
$c, \tilde{c}$	penalty parameter