Pinball Twin Bounded Support Vector Clustering

M.Sc. Thesis

By

MOHAMMAD TABISH

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY INDORE

KHANDWA ROAD, SIMROL, INDORE-453552 **INDIA**

JUNE 2021

Pinball Twin Bounded Support Vector Clustering

A THESIS

Submitted in partial fulfillment of the
requirements for the award of the degree

of

Master of Science

by

MOHAMMAD TABISH

(Roll No. 1903141002)

Under the guidance of

Dr. M. Tanveer

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY INDORE

JUNE 2021

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Pinball Twin Bounded Support Vector Clustering** in the partial fulfillment of the requirements for the award of the degree of **Master of Science** and submitted in the **Department of Mathematics**, **Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from July 2020 to June 2021 under the supervision of **Dr. M. Tanveer**, Associate Professor and Ramanujan Fellow, Department of Mathematics, IIT Indore.

The matter presented in this thesis by me has not been submitted for the award of any other degree of this or any other institute.

Tabish 08/06/2021

Signature of the student with date

(**MOHAMMAD TABISH**)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

June 08, 2021

Signature of Thesis Supervisor with date

(**Dr. M. Tanveer**)

MOHAMMAD TABISH has successfully given his M.Sc. Oral Examination held on June 08, 2021.

Signature of Supervisor of M.Sc Thesis

Date: June 08, 2021

Signature of PSPC Member 1

Date: June 08, 2021

Signature of Convener, DPGC

Date: June 08, 2021

Signature of PSPC Member 2

Date: June 08, 2021

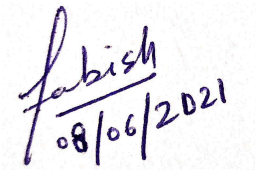
Acknowledgements

I would like to express my sincere thanks to Dr. M. Tanveer, my thesis supervisor, for allowing me to work on a project under his supervision and for being helpful from the start, leading, encouraging, and sharing his brilliant insights with me along the journey, without which it would have been difficult to conduct a thorough research into the given topic.

I am thankful for significant support and guidance of Research Scholars of our lab, Mudasir Ahmad Ganaie, Bharat Richhariya and Ashwani Kumar Malik. Despite being engaged and focused on their research, they were always there to help me in my work. And a particular thanks to Jatin Jangir, whose assistance was of great importance to me.

I would like to thank PSPC members, Dr. Vijay Kumar Sohani and Prof. Ram Bilas Pachori for their valuable remarks, suggestions and questionnaires.

I would also like to express my gratitude to Dr. Md. Aquil Khan, Head, Department of Mathematics, for providing our OPTIMAL research lab with state-of-the-art computing facilities, allowing us to complete the research efficiently.



MOHAMMAD TABISH

1903141002

Department of Mathematics, IIT Indore.

Dedicated to my family...

Abstract

Clustering is a very popular approach in machine learning for unlabelled data. Twin support vector clustering (TWSVC) and the twin bounded support vector clustering (TBSVC), plane-based clustering algorithms introduced recently, work on twin support vector machine (TWSVM) principles and are used in widespread clustering problems. However, both TWSVC and TBSVC are sensitive to noise and suffers from low re-sampling of data stability due to the use of hinge loss. The pinball loss features noise insensitivity and stability for re-sampling of data. Within this thesis, we first present basic formulations of the previous methods in plane based clustering and discuss their shortcomings. Then we propose two plane-based clustering methods, twin bounded support vector clustering using pinball loss (pinTBSVC) and sparse twin bounded support vector clustering using pinball loss (pinSTBSVC) which inherits various attributes from previous plane based clustering algorithms. Sparse solutions help to create better generalized solutions in the clustering problems; hence we attempt to use maximum margin regularization term to propose pinSTBSVC. The proposed pinTBSVC and pinSTBSVC solve the singularity problem and improve the aforementioned plane-based clustering algorithms. Experimental results performed on benchmark UCI datasets indicate that the proposed methods outperform other existing plane-based clustering algorithms. Additionally, we also give the application of the proposed method to biomedical image clustering and marketing science. Numerical experiments on real world benchmark datasets show that the proposed models give better generalization performance.

List of Publications

1. M. Tanveer, M. Tabish, J. Jangir: "Pinball Twin Bounded Support Vector Clustering", IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2021).

2. M. Tanveer, M. Tabish, J. Jangir: "Sparse Pinball Twin Bounded Support Vector Clustering", 2021 (Under Preparation).

Contents

List of Figures

List of Tables

Abbreviations

K.K.T.	**K**arush-**K**uhn-**T**ucker
SVM	**S**upport **V**ector **M**achine
SVC	**S**upport **V**ector **C**lustering
CCCP	**C**oncave **C**onvex **P**rocedure
QPP	**Q**uadratic **P**rogramming **P**roblem
TWSVM	**T**win **S**upport **V**ector **M**achine
TWSVC	**T**win **S**upport **V**ector **C**lustering
TBSVC	**T**win **B**ounded **S**upport **V**ector **C**lustering
PinTSVC	**P**inball **T**win **S**upport **V**ector **C**lustering
PinTBSVC	**P**inball **T**win **B**ounded **S**upport **V**ector **C**lustering
PinSTBSVC	**S**parse **P**inball **T**win **B**ounded **S**upport **V**ector **C**lustering

Chapter 1

Introduction

Machine Learning (ML) is a branch of artificial intelligence wherein we study some advance computer algorithms that improve by themselves with the help of the data provided i.e. by finding some patterns or trends in the data. Vapnik and Cortes [1] developed the concept of support vector machine (SVM), one of the most widely used and easy to interpret machine learning model that works well in solving classification and regression problems [1, 2]. SVM is useful in a number of challenging areas like medical [3], face detection [4], etc. Twin support vector machine (TWSVM) [5] is a variant of SVM which reduces the complexity of SVM and improves the classification accuracy. Both SVM and TWSVM are supervised learning algorithms i.e. they are used for the datasets in which class of each data point is known.

Among the unsupervised learning techniques, clustering is one of the most famous technique that finds the pattern in the data by grouping it into different clusters. It works in the manner that the points having similar properties or attributes are grouped in one cluster. After the formation of clusters in the given dataset, a new datapoint can be easily assigned a label by using some appropriate function. Recently, clustering has gained popularity in a variety of domains, including web analysis [6], facial recognition [7, 8], etc.

In the last few decades, many clustering techniques were proposed such as point-based clustering techniques like k-means [9] that aims to divide the given data into k-clusters by minimizing the Euclidean distance of data points within clusters from the cluster

mean, and k-median [10], which uses median instead of mean to form clusters within the dataset. Although point based clustering algorithms are effective in clustering the data, they fall short in some non-standard datasets. So, plane based clustering were introduced to cluster datasets around cluster center planes instead of clustering around single points like k-plane clustering (kPC) [11] and proximal plane clustering (PPC) [12]. Clustering implementation varies from method to method, for example, the kPC formulates the similarity within a cluster whereas the PPC formulates the dissimilarity between the various clusters. As it turns out, the shortfall of these methods come from the inherent implementation; the kPC algorithm ignores the influence of other nearby clusters, meaning that two non spherical clusters that are placed relatively closeby will not be labeled correctly.

Jayadeva et al. [5] proposed an efficient twin support vector machine (TWSVM) [5] algorithm for the pattern classification. TWSVM uses two non-parallel hyperplanes, each of which is closest to one class and farthest from the other. Shao et al. [13] proposed twin bounded SVM (TBSVM) [13] that added an extra regularization term in primal QPPs and implemented structural risk minimization principle. In 2014, further improvements were proposed by Tanveer in [14], the formulation incorporated the regularization term in each objective function of TWSVM and use two smoothening techniques to solve the proposed formulation. It solves two system of linear equations unlike two QPPs in TWSVM, leading to reduced computation cost and a straightforward and fast algorithm. The robust energy-based least squares twin SVM (RELS-TSVM) [15] proposed in 2016 was an improved approach for implementing the structural risk minimisation principle in TWSVM, with use of a regularization term along with an energy parameter in each problem, and as a result got a positive definite matrix in the dual problem. In the recent comprehensive evaluation of 187 classifiers including eight variants of TWSVM on 90 University of California Irvine (UCI) datasets [16], RELS-TSVM classifier outperformed all other TWSVM variants [17].

Inspired from TWSVM [5], the twin support vector clustering (TWSVC) [18] method was proposed by Wang et al. [18], that uses both the similarity within a cluster and the dissimilarity between clusters. This allowed TWSVC to outperform both the previous plane-based clustering methods. Several developments were made in the

nature of TWSVC to resolve numerous issues and complaints with the procedure, such as computing the inverses of potentially singular matrices and ignoring the large margin concept when generating the proximal planes. Even after disregarding the algorithmic problems, TWSVC takes a long time to converge to an optimal solution [19]. Twin bounded support vector clustering (TBSVC) [20] succeeded TWSVC and introduced a maximum margin regularization terms in the formulation of TWSVC making the gap between parallel planes and proximal plane to be as large as possible. TBSVC also resolves the singularity issue of the matrix in dual problem. Another notable method is the Ramp-based twin support vector clustering (rampTSVC) [21] which introduced the ramp function to measure the within-cluster and between-cluster scatter; this ensured insensitivity towards samples far from the cluster centres. This approach of utilizing within-class and between-class improves noise-insensitivity. Least squares projection twin support vector clustering (LSPTSVC) [22] is another plane based clustering model which clusters the dataset by finding a projection axis for each cluster such that it minimizes the within class scatter. Also, it's solution involves solving system of linear equations which take lesser training time.

The extensions of TWSVC discussed so far use hinge loss as the loss or the cost function. Hinge loss has a central goal of maximizing the shortest distance between clusters. This goal makes hinge loss highly sensitive towards outliers, making it susceptible to noise and re-sampling. A statement can be made that all real-world datasets suffer from some degree of noise, making hinge loss a weaker candidate for real-life clustering applications. Comparing to hinge loss, pinball loss SVM proposed by Huang et al. [23] is robust in both noise insensitivity and re-sampling stability. Pinball loss has been extensively used for classification problems, however not much development has been made for clustering. The pinball loss twin support vector clustering (pinTSVC) [24] was the first approach to incorporate pinball loss in TWSVC. The pinTSVC is a TWSVC-based method that uses pinball loss for optimization. Tanveer et al. [24] concluded that pinTSVC performs better than previous clustering algorithms for datasets with noise. However, pinTSVC does not implement the structural risk minimization principle. Thus, pinTSVC's performance can be further improved by implementing the structural risk minimization principle. A consequence of the pinball loss function is that a penalty

is also imposed on correctly classified points and it negatively influence the sparsity of the solution. Recently, Tanveer et al. introduced the sparse version of pinTSVC known as SPTSVC [25] which make the solution of problem in pinTSVC more sparse. It is shown that classifiers with higher sparsity tend to have better generalization [26].

Motivated by the recent pinTSVC algorithm [24], two novel and efficient algorithms are proposed, pinball loss twin bounded support vector clustering and it's sparse version, pinTBSVC and pinSTBSVC, to improve the performance of clustering algorithms on noisy datasets and make them stable for re-sampling of data. Recognizing the shortcomings of pinball loss, sparse version of pinTBSVC uses pinball loss with a ϵ-insensitive zone, that helps in providing sparsity to the solution of pinTBSVC.

Throughout this thesis, we consider all vectors to be column vectors. Suppose that we have m samples in $\mathbb{R}^n$ (n-dimensional real space), and that these samples are segregated into k-clusters, this is represented by a $m \times n$ matrix $A = (x_1, x_2, ..., x_m)^T$. The data samples in i^{th} cluster is denoted by A_i and those not in the i^{th} cluster is denoted by the matrix $\hat{A}_i \in \mathbb{R}^{(m-m_i)\times n}$ for $i \in 1, 2, ..., k$. $||.||$ denotes the L_2 norm and e represents a vector of ones of appropriate dimension.

Chapter 2

Background

Classical support vector machine constructs two parallel planes for data classification, and maximize the margin between them to make the classification more accurate, by solving one large Quadratic Programming Problem (QPP). Jayadeva et al. [5] proposed twin support vector machine that generates two non-parallel hyperplanes for classification of data, by solving two smaller sized QPPs, unlike SVM. In TWSVM [5], the primary goal is to generate two hyperplanes in such a way that each plane is as close as possible to its class and far away as possible to another. With this idea of generating two planes for two classes of data, TWSVM showed better performance in classifying the datasets.

Based on the principle of TWSVM, a plane based clustering technique was developed by Wang et al. in 2015 known as twin support vector clustering (TWSVC) [18]. In this chapter, we discuss the previous and similar works done in the plane based clustering algorithms.

2.1 TWSVC

Twin support vector clustering (TWSVC) [18], a plane-based clustering method, works on the principles of TWSVM and performs clustering by seeking k-cluster center planes

for k clusters in the data samples as follows:

$$w_i^T x + b_i = 0 \quad \text{for each} \quad i = 1, 2, ..., k \tag{2.1}$$

where $w_i \in \mathbb{R}^n, b_i \in \mathbb{R}$, by solving the following optimization problem for each $i = 1, 2, ..., k$

$$\begin{aligned} \min_{w_i, b_i, \eta_i} \quad & \frac{1}{2}||A_i w_i + b_i e||^2 + c e^T \eta_i \\ \text{s.t.} \quad & |\hat{A}_i w_i + b_i e| \geq e - \eta_i, \quad \eta_i \geq 0, \end{aligned} \tag{2.2}$$

where $c > 0$ is a penalty parameter, η_i is an error bounding variable. The i^{th} problem (2.2) can be decomposed into a series of sub-problems by using CCCP (concave-convex procedure) [27] with initial w_i^0 and b_i^0 as:

$$\begin{aligned} \min_{w_i^{j+1}, b_i^{j+1}, \eta_i^{j+1}} \quad & \frac{1}{2}||A_i w_i^{j+1} + b_i^{j+1} e||^2 + c e^T \eta_i^{j+1} \\ \text{s.t.} \quad & T(|\hat{A}_i w_i^{j+1} + b_i^{j+1} e|) \geq e - \eta_i^{j+1}, \quad \eta_i^{j+1} \geq 0, \end{aligned} \tag{2.3}$$

where $T(.)$ denotes the first order Taylor expansion.

For the expansion of Taylor series, we need sub-gradient of $|\hat{A}_i w_i^j + b_i^j e|$ w.r.t. w_i^j and b_i^j, which is defined as $\nabla(|\hat{A}_i w_i^j + b_i^j e|)$=diag(sign$(\hat{A}_i w_i^j + b_i^j e))[\hat{A}_i, e]$, and also we have that $|\hat{A}_i w_i^j + b_i^j e|$=diag(sign$(\hat{A}_i w_i^j + b_i^j e))(\hat{A}_i w_i^j + b_i^j e)$.

So, we have Taylor expansion as:

$$T(|\hat{A}_i w_i^{j+1} + b_i^{j+1} e|) = |\hat{A}_i w_i^j + b_i^j e| + \nabla(|\hat{A}_i w_i^j + b_i^j e|)([w_i^{j+1}; b_i^{j+1}] - [w_i^j; b_i^j]) \tag{2.4}$$

Solving, we get

$$T(|\hat{A}_i w_i^{j+1} + b_i^{j+1} e|) = D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1} e), \tag{2.5}$$

$$D_i = \text{diag}(\text{sign}(\hat{A}_i w_i^j + b_i^j e)). \tag{2.6}$$

So, problem (2.3) can be written as:

$$\begin{aligned} \min_{w_i^{j+1},b_i^{j+1},\eta_i^{j+1}} \quad & \frac{1}{2}||A_i w_i^{j+1} + b_i^{j+1} e||^2 + c e^T \eta_i^{j+1} \\ \text{s.t.} \quad & \text{diag}(\text{sign}(\hat{A}_i w_i^j + b_i^j e))(\hat{A}_i w_i^{j+1} + b_i^{j+1} e) \geq e - \eta_i^{j+1}, \quad \eta_i^{j+1} \geq 0. \end{aligned} \tag{2.7}$$

Now, using the Karush-Kuhn-Tucker (K.K.T.) conditions [28],we get the dual problem as:

$$\begin{aligned} \min_{\gamma} \quad & \frac{1}{2}\gamma^T G(H^T H)^{-1} G^T \gamma - e^T \gamma \\ \text{s.t.} \quad & 0 \leq \gamma \leq ce \end{aligned} \tag{2.8}$$

where D_i =diag(sign($\hat{A}_i w_i^j + b_i^j e$)), $H = [A_i \quad e]$, $G = D_i[\hat{A}_i \quad e]$ and $\gamma \in \mathbb{R}^{m-m_i}$ is the Lagrange multiplier.

Nonlinear TWSVC:

Linear TWSVC can easily be extended for datasets that are non-linearly separable and center-manifolds are generated instead of center-planes by using kernel trick. So, for k-clusters in the dataset, we get k-cluster center-manifolds as:

$$\text{center-manifold}_i := K(x, A) y_i + b_i e = 0, \tag{2.9}$$

where $K(.,.)$ is an appropriate kernel function chosen according to the problem. So, the optimization problem is as follows:

$$\begin{aligned} \min_{y_i,b_i,\eta_i} \quad & \frac{1}{2}||K(A_i, A) y_i + b_i e||^2 + c_1 e^T \eta_i \\ \text{s.t.} \quad & |K(\hat{A}_i, A) y_i + b_i e| \geq e - \eta_i, \eta_i \geq 0. \end{aligned} \tag{2.10}$$

We can solve the above problem similar to linear case i.e. by first using CCCP to decompose the i^{th} problem into the series of sub-problems and then using K.K.T.

conditions to give the dual as follows:

$$\begin{aligned} \min_{\gamma} \quad & \frac{1}{2}\gamma^T G(H^T H)^{-1} G^T \gamma - e^T \gamma \\ \text{s.t.} \quad & 0 \leq \gamma \leq ce, \end{aligned} \tag{2.11}$$

where $H = [K(A_i, A) \;\; e]$, $G = D_i[K(\hat{A}_i, A) \;\; e]$.

2.2 TBSVC

TWSVC [18] has an assumption of the existence of $(H^T H)^{-1}$ appearing in the dual problem, however, this is not always the case as the matrix can be ill-conditioned. Bai et al. [20] proposed an improved version of TWSVC, known as twin bounded support vector clustering (TBSVC), which implements the structural risk minimization principle by including an extra regularization term in the objective function of TWSVC and resolves the issue of invertibility of matrix $H^T H$. So, the formulation of TBSVC [20] is given as:

$$\begin{aligned} \min_{w_i, b_i, \eta_i} \quad & \frac{1}{2}||A_i w_i + b_i e||^2 + c_1 e^T \eta_i + \frac{1}{2} c_2 ||w_i||^2 \\ \text{s.t.} \quad & |\hat{A}_i w_i + b_i e| \geq e - \eta_i, \quad \eta_i \geq 0, \end{aligned} \tag{2.12}$$

where c_1, c_2 are two positive parameters and η_i is the error bounding variable. Now similar to TWSVC, problem (2.12) is decomposed into series of sub-problems using CCCP and we get the following problem as:

$$\begin{aligned} \min_{w_i^{j+1}, b_i^{j+1}, \eta_i^{j+1}} \quad & \frac{1}{2}||A_i w_i^{j+1} + b_i^{j+1} e||^2 + c_1 e^T \eta_i^{j+1} + \frac{c_2}{2}||w_i||^2 \\ \text{s.t.} \quad & \text{diag}(\text{sign}(\hat{A}_i w_i^j + b_i^j e))(\hat{A}_i w_i^{j+1} + b_i^{j+1} e) \geq e - \eta_i^{j+1}, \\ & \eta_i^{j+1} \geq 0. \end{aligned} \tag{2.13}$$

Consider D_i=diag(sign($\hat{A}_i w_I^j + b_i^j e$)), the Lagrangian of the above problem is as follows:

$$\begin{aligned} L =& \frac{1}{2}||A_i w_i^{j+1} + b_i^{j+1} e||^2 + c_1 e^T \eta_i^{j+1} + \frac{1}{2} c_2 ||w_i^{j+1}||^2 \\ &+ \gamma^T (e - \eta_i^{j+1} - D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1} e)) - \beta^T \eta_i^{j+1}, \end{aligned} \tag{2.14}$$

where γ and β are Lagrange multipliers. Using the K.K.T. conditions, we get the following equations

$$\frac{\partial L}{\partial w_i^{j+1}} = A_i^T (A_i w_i^{j+1} + b_i^{j+1} e) + c_2 w_i^{j+1} - (\gamma^T D_i \hat{A}_i)^T = 0, \tag{2.15}$$

$$\frac{\partial L}{\partial b_i^{j+1}} = e^T (A_i w_i^{j+1} + b_i^{j+1} e) - (\gamma^T D_i e)^T = 0, \tag{2.16}$$

$$\frac{\partial L}{\partial \eta_i^{j+1}} = (c_1 e^T)^T - \gamma - \beta = 0, \tag{2.17}$$

$$\gamma^T (e - \eta_i^{j+1} - D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1} e)) = 0, \tag{2.18}$$

$$\beta^T \eta_i^{j+1} = 0, \tag{2.19}$$

$$\gamma, \beta \geq 0. \tag{2.20}$$

Now again by using K.K.T. conditions, we get the dual for problem (2.13) as:

$$\begin{aligned} \min_{\gamma} \quad & \frac{1}{2} \gamma^T G (H^T H + C)^{-1} G^T \gamma - e^T \gamma \\ \text{s.t.} \quad & 0 \leq \gamma \leq c_1 e, \end{aligned} \tag{2.21}$$

where $C = \begin{pmatrix} c_2 I_n & 0 \\ 0 & 0 \end{pmatrix}$,

D_i=diag(sign($\hat{A}_i w_i^j + b_i^j e$)), $G = D_i[\hat{A}_i \;\; e]$ and $H = [A_i \;\; e]$.

It is easy to verify that the matrix $(H^T H + C)$ is positive definite in comparison to $H^T H$ which is just positive semi-definite [20].

2.3 pinTSVC

The loss function used in the above algorithms is hinge loss whereas pinTSVC [24] is formulated using the pinball loss, an unsymmetrical loss function defined as:

$$\mathcal{L}_\tau(x) = \begin{cases} x, & \text{if } x \geq 0, \\ -\tau x, & \text{if } x < 0, \end{cases} \tag{2.22}$$

where $\tau \in [0, 1]$ is the pinball loss parameter. The pinball loss function assigns a penalty to misclassified data points and correctly classified points. So, the formulation of pinTSVC is given as:

$$\begin{aligned} \min_{w_i, b_i, \eta_i} \quad & \frac{1}{2}||A_i w_i + b_i e||^2 + c e^T \eta_i \\ \text{s.t.} \quad & |\hat{A}_i w_i + b_i e| \geq e - \eta_i, \\ & |\hat{A}_i w_i + b_i e| \leq e + \frac{\eta_i}{\tau}, \end{aligned} \tag{2.23}$$

where η_i is the error bounding variable. Using CCCP, the i^{th} problem can be broken down into a series of convex quadratic sub-problems to give the following problem:

$$\begin{aligned} \min_{w_i^{j+1}, b_i^{j+1}, \eta_i^{j+1}} \quad & \frac{1}{2}||A_i w_i^{j+1} + b_i^{j+1} e||^2 + c e^T \eta_i^{j+1} \\ \text{s.t.} \quad & T(|\hat{A}_i w_i^{j+1} + b_i^{j+1} e|) \geq e - \eta_i^{j+1}, \\ & T(|\hat{A}_i w_i^{j+1} + b_i^{j+1} e|) \leq e + \frac{\eta_i^{j+1}}{\tau}. \end{aligned} \tag{2.24}$$

Using the K.K.T. conditions, we can write

$$v = -(H^T H)^{-1} G^T (\beta - \gamma), \tag{2.25}$$

where $v = \begin{bmatrix} w_i^{j+1} \\ b_i^{j+1} \end{bmatrix}$, $G = D_i[\hat{A}_i \;\; e]$ and $H = [A_i \;\; e]$.
Again using K.K.T. conditions and solving similarly as in TWSVC, we get the dual of

(2.24) as:

$$\begin{aligned} \min_{\gamma-\beta} \quad & \frac{1}{2}(\beta-\gamma)^T G(H^T H)^{-1} G^T(\beta-\gamma) - (\beta-\gamma)^T e \\ \text{s.t.} \quad & \gamma, \beta \geq 0, \quad ce = \gamma + \frac{\beta}{\tau}. \end{aligned} \tag{2.26}$$

For a new data point x_t, the label is assigned as follows:

$$y(x_t) = arg \min_{i=1,2,...,k} |w_i^T x_t + b_i|, \tag{2.27}$$

among the k-clusters in the dataset.

2.4 SPTSVC

The loss function used in pinTSVC [24] has non-zero sub-derivative in the entire domain except for the origin, leading to the solution's loss of sparsity. So, to overcome this problem, sparse version of pinTSVC employing pinball loss (SPTSVC) [25] was introduced in which ϵ-insensitive loss function [23] is used, which is defined as:

$$\mathcal{L}_\tau^\epsilon(x) = \begin{cases} x - \epsilon, & \text{if} \quad x > \epsilon, \\ 0, & \text{if} \quad -\frac{\epsilon}{\tau} \leq x \leq \epsilon, \\ -\tau(x + \frac{\epsilon}{\tau}), & \text{if} \quad x < -\frac{\epsilon}{\tau}, \end{cases} \tag{2.28}$$

which gives us that no penalty is assigned to the data points lying in the width $\epsilon(1+\frac{1}{\tau})$ and we also achieved sparsity as the sub-gradient of the above function is 0 in the range $[-\frac{\epsilon}{\tau}, \epsilon]$. So, SPTSVC can be formulated as:

$$\begin{aligned} \min_{w_i, b_i, \eta_i} \quad & \frac{1}{2}||A_i w_i + b_i e||^2 + ce^T \eta_i \\ \text{s.t.} \quad & |\hat{A}_i w_i + b_i e| \geq e - \eta_i - e\epsilon, \\ & |\hat{A}_i w_i + b_i e| \leq e + \frac{\eta_i}{\tau} + e\frac{\epsilon}{\tau}, \\ & \eta_i \geq 0, \quad \text{for} \quad i = 1, 2, ..., k, \end{aligned} \tag{2.29}$$

where η_i is the error bounding parameter, $\epsilon, \tau \in [0, 1]$, are the various parameters of the loss function used.

Similar to TWSVC, we can use the CCCP to break the i^{th} problem into a sequence of convex quadratic sub-problems as:

$$\begin{aligned} \min_{w_i^{j+1}, b_i^{j+1}, \eta_i^{j+1}} \quad & \frac{1}{2}||A_i w_i^{j+1} + b_i^{j+1} e||^2 + c e^T \eta_i^{j+1} \\ \text{s.t.} \quad & T(|\hat{A}_i w_i^{j+1} + b_i^{j+1} e|) \geq e - \eta_i^{j+1} - e\epsilon, \\ & T(|\hat{A}_i w_i^{j+1} + b_i^{j+1} e|) \leq e + \frac{\eta_i^{j+1}}{\tau} + e\frac{\epsilon}{\tau}, \\ & \eta_i^{j+1} \geq 0, \end{aligned} \tag{2.30}$$

and using K.K.T. conditions, we get the dual of (2.30) as follows:

$$\begin{aligned} \min_{\gamma - \beta} \quad & \frac{1}{2}(\beta - \gamma)^T G (H^T H)^{-1} G^T (\beta - \gamma) - (\beta - \gamma)^T e + e\epsilon(\gamma^T + \frac{\beta^T}{\tau}), \\ \text{s.t.} \quad & \gamma, \beta, \delta \geq 0, \quad ce = \gamma + \frac{\beta}{\tau} + \delta, \end{aligned} \tag{2.31}$$

which can also be written as

$$\begin{aligned} \min_{\lambda} \quad & \frac{1}{2}\lambda^T G (H^T H)^{-1} G^T \lambda - \lambda^T e(\frac{\epsilon}{\tau} + 1) + \gamma^T e(\epsilon + \frac{\epsilon}{\tau}), \\ \text{s.t.} \quad & \gamma, \lambda \geq 0, \quad \gamma(1 + \frac{1}{\tau}) - \frac{\lambda}{\tau} \leq ce, \end{aligned} \tag{2.32}$$

where $\lambda = \gamma - \beta$, Solving the above dual, we get the solution to our problem and any new data point x_t can be assigned a cluster by following equation:

$$y(x_t) = arg \min_{i=1,2,\ldots,k} |w_i^T x_t + b_i|. \tag{2.33}$$

Chapter 3

Proposed Algorithms

In this chapter, we discuss two proposed algorithms given in the below sections and also provide some theoretical justifications of the proposed models.

3.1 Proposed pinball loss TBSVC (pinTBSVC)

In this section, we propose an efficient pinball loss twin bounded support vector clustering (pinTBSVC).

3.1.1 Linear pinTBSVC

The proposed linear pinTBSVC finds k-clusters center-planes with parameters $[w_i, b_i]$ for $i = 1, 2, ..., k$ by solving the following formulation:

$$\begin{aligned} \min_{w_i, b_i, \eta_i} \quad & \frac{1}{2}||A_i w_i + b_i e||^2 + c_1 e^T \eta_i + \frac{1}{2} c_2 ||w_i||^2 \\ \text{s.t.} \quad & |\hat{A}_i w_i + b_i e| \geq e - \eta_i, \\ & |\hat{A}_i w_i + b_i e| \leq e + \frac{\eta_i}{\tau}, \end{aligned} \tag{3.1}$$

where η_i is the error bounding parameter and $\tau \in [0, 1]$ is the pinball loss parameter. The first term of the objective function in problem (3.1) is to minimize the squared

distances of the points in i^{th} cluster from the i^{th} hyperplane. The second term reduces the error caused by data points from clusters other than the i^{th} cluster that is within a unit distance of the i^{th} cluster's hyperplane, as well as correctly classified points that are penalized based on the pinball loss function's parameter τ. The third term in the objective function is the regularization term leading to maximizing the distance between the optimal plane $w^T x + b = 0$ and the parallel planes $w^T x + b = \pm 1$. Our pinTBSVC uses the large margin principle to obtain the proximal and its parallel planes, and the use of pinball loss leads to feature noise insensitivity around the proximal plane.

We can decompose the i^{th} problem (3.1) into a series of convex quadratic sub-problems, with j as the index of sub-problem, by using the concave-convex procedure (CCCP) as:

$$\begin{aligned} \min_{w_i^{j+1}, b_i^{j+1}, \eta_i^{j+1}} \quad & \frac{1}{2}||A_i w_i^{j+1} + b_i^{j+1} e||^2 + c_1 e^T \eta_i^{j+1} + \frac{1}{2} c_2 ||w_i^{j+1}||^2 \\ \text{s.t.} \quad & T(|\hat{A}_i w_i^{j+1} + b_i^{j+1} e|) \geq e - \eta_i^{j+1}, \\ & T(|\hat{A}_i w_i^{j+1} + b_i^{j+1} e|) \leq e + \frac{\eta_i^{j+1}}{\tau}. \end{aligned} \tag{3.2}$$

We can now use the Taylor expansion to write

$$T(\hat{A}_i w_i^{j+1} + b_i^{j+1} e) = D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1} e), \tag{3.3}$$

where $D_i = \text{diag}(\text{sign}(\hat{A}_i w_i^j + b_i^j e))$.

We find the dual of the problem (3.2) by considering the Lagrangian as:

$$\begin{aligned} L = & \frac{1}{2}||A_i w_i^{j+1} + b_i^{j+1} e||^2 + c_1 e^T \eta_i^{j+1} + \frac{1}{2} c_2 ||w_i^{j+1}||^2 \\ & + \gamma^T (e - \eta_i^{j+1} - D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1} e)) + \beta^T (D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1} e) - e - \frac{\eta_i^{j+1}}{\tau}), \end{aligned} \tag{3.4}$$

where $\gamma, \beta \geq 0$ are the Lagrange multipliers. Applying the K.K.T. conditions as:

$$\frac{\partial L}{\partial w_i^{j+1}} = A_i^T(A_i w_i^{j+1} + b_i^{j+1}e) + c_2 w_i^{j+1} - (\gamma^T D_i \hat{A}_i)^T + (\beta^T D_i \hat{A}_i)^T = 0, \tag{3.5}$$

$$\frac{\partial L}{\partial b_i^{j+1}} = e^T(A_i w_i^{j+1} + b_i^{j+1}e) - (\gamma^T D_i e)^T + (\beta^T D_i e)^T = 0, \tag{3.6}$$

$$\frac{\partial L}{\partial \eta_i^{j+1}} = (c_1 e^T)^T - \gamma - \frac{\beta}{\tau} = 0, \tag{3.7}$$

$$\gamma^T(e - \eta_i^{j+1} - D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1}e)) = 0, \tag{3.8}$$

$$\beta^T(D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1}e) - e - \frac{\eta_i^{j+1}}{\tau}) = 0. \tag{3.9}$$

From the above equations, we can write

$$v = -(H^T H + C)^{-1} G^T(\beta - \gamma), \tag{3.10}$$

where $G = D_i[\hat{A}_i \;\; e]$, $H = [A_i \;\; e]$ and $v = [w_i^{j+1} \;\; b_i^{j+1}]^T$ and
$C = \begin{pmatrix} c_2 I_n & 0 \\ 0 & 0 \end{pmatrix}$.
It can easily be shown that the matrix $(H^T H + C)^{-1}$ is non-singular by considering its determinant and showing it to be strictly positive.

Now, using the K.K.T. conditions, we modify our Lagrangian function to get the following dual problem for (3.2) as:

$$\begin{aligned} \min_{\gamma - \beta} \;\; & \frac{1}{2}(\beta - \gamma)^T G(H^T H + C)^{-1} G^T(\beta - \gamma) - (\beta - \gamma)^T e, \\ \text{s.t.} \;\; & \gamma, \beta \geq 0, \;\; c_1 e = \gamma + \frac{\beta}{\tau}. \end{aligned} \tag{3.11}$$

After solving the above dual problem, we can give a label to a new data point x_t by:

$$y(x_t) = arg \min_{i=1,2,\ldots,k} |w_i^T x_t + b_i|. \tag{3.12}$$

3.1.2 Nonlinear pinTBSVC

Linear pinTBSVC is extended to the cases where data is not linearly separable by using the kernel method to get k-clusters manifolds as:

$$\text{Center-manifold}_i := K(x, A)y_i + b_i = 0, \tag{3.13}$$

where $K(.,.)$ is an appropriate kernel function chosen according to the problem, $y_i \in \mathbb{R}^m$, $b_i \in \mathbb{R}$. So, we have our problem as:

$$\begin{aligned} \min_{y_i, b_i, \eta_i} \quad & \frac{1}{2}||K(A_i, A)y_i + b_i e||^2 + c_1 e^T \eta_i + \frac{1}{2}c_2||y_i||^2 \\ \text{s.t.} \quad & |K(\hat{A}_i, A)y_i + b_i e| \geq e - \eta_i, \\ & |K(\hat{A}_i, A)y_i + b_i e| \leq e + \frac{\eta_i}{\tau}, \end{aligned} \tag{3.14}$$

where η_i is the error bounding parameter and $\tau \in [0, 1]$ is the pinball loss parameter. Proceeding similarly to the linear case, i.e. using CCCP to decompose the i^{th} problem into convex quadratic sub-problems and then using K.K.T. conditions, we can get the following equations:

$$v = -(H^T H + C)^{-1} Q^T (\beta - \gamma), \tag{3.15}$$

where $H = [K(A_i, A) \;\; e]$, $Q = D_i[K(\hat{A}_i, A) \;\; e]$, $v = [y_i^{j+1} \;\; b_i^{j+1}]^T$
and $D_i = \text{diag}(\text{sign}(K(\hat{A}_i, X)y_i^j + b_i^j e))$.
So, the dual of (3.14) is obtained as:

$$\begin{aligned} \min_{\gamma - \beta} \quad & \frac{1}{2}(\beta - \gamma)^T Q(H^T H + C)^{-1} Q^T (\beta - \gamma) - (\beta - \gamma)^T e, \\ \text{s.t.} \quad & \gamma, \beta \geq 0, \quad c_1 e = \gamma + \frac{\beta}{\tau}. \end{aligned} \tag{3.16}$$

3.2 Proposed sparse pinball loss TBSVC (pinSTB-SVC)

In this section, we propose sparse pinball twin bounded support vector clustering (pinSTBSVC) for both linear and non-linear cases. The proposed pinSTBSVC is the sparse version of pinTBSVC.

3.2.1 Linear pinSTBSVC

Our proposed linear pinSTBSVC finds k-cluster center-planes i.e. the vectors $[w_i, b_i]^T$ for $i = 1, 2, ..., k$ by solving the following optimization problem:

$$\begin{aligned} \min_{w_i, b_i, \eta_i} \quad & \frac{1}{2}||A_i w_i + b_i e||^2 + c_1 e^T \eta_i + \frac{1}{2} c_2 ||w_i||^2 \\ \text{s.t.} \quad & |\hat{A}_i w_i + b_i e| \geq e - \eta_i - e\epsilon, \\ & |\hat{A}_i w_i + b_i e| \leq e + \frac{\eta_i}{\tau} + e\frac{\epsilon}{\tau}, \\ & \eta_i \geq 0. \end{aligned} \tag{3.17}$$

Here, $\epsilon, \tau \in [0, 1]$ are parameters for the sparse pinball loss function and η_i is the error bounding variable. The first term of the objective function in problem (3.17) minimize the squared distance of the i^{th} hyperplane from the i^{th} cluster points. The second term represents the error term minimization with η_i as the slack variable. The third term is the regularization term which maximizes the margin between the proximal plane and its parallel hyperplanes. The first constraints of our problem minimize the penalty coming from the data points of $\hat{A}_i$ whose distances are at most $(1 - \epsilon)$ from the i^{th} hyperplane. The second constraint minimizes the penalty of the points in $\hat{A}_i$ which are farther than $(1 + \frac{\epsilon}{\tau})$ distance away from the i^{th} hyperplane.

Now, using the convex-concave procedure (CCCP), we can decompose the i^{th} problem

in (3.17) into a series of convex quadratic sub-problems as shown:

$$\begin{aligned} \min_{w_i^{j+1},b_i^{j+1},\eta_i^{j+1}} \quad & \frac{1}{2}||A_i w_i^{j+1} + b_i^{j+1}e||^2 + c_1 e^T \eta_i^{j+1} + \frac{1}{2}c_2||w_i^{j+1}||^2 \\ \text{s.t.} \quad & T(|\hat{A}_i w_i^{j+1} + b_i^{j+1}e|) \geq e - \eta_i^{j+1} - e\epsilon, \\ & T(|\hat{A}_i w_i^{j+1} + b_i^{j+1}e|) \leq e + \frac{\eta_i^{j+1}}{\tau} + e\frac{\epsilon}{\tau}, \\ & \eta_i^{j+1} \geq 0. \end{aligned} \tag{3.18}$$

Expanding the first-order Taylor series, we obtain:

$$T(|\hat{A}_i w_i^{j+1} + b_i^{j+1}e|) = D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1}e), \tag{3.19}$$

where $D_i = \text{diag}(\text{sign}(\hat{A}_i w_i^j + b_i^j e))$.

We now formulate the dual of the primal problem (3.18) by considering it's Lagrangian as:

$$\begin{aligned} L \ = & \frac{1}{2}||A_i w_i^{j+1} + b_i^{j+1}e||^2 + c_1 e^T \eta_i^{j+1} + \frac{1}{2}c_2||w_i^{j+1}||^2 + \gamma^T(e - \eta_i^{j+1} - e\epsilon - D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1}e)) \\ & + \beta^T(D_i(\hat{A}_i w_i^{j+1} + b_i^{j+1}e) - e - \frac{\eta_i^{j+1}}{\tau} - e\frac{\epsilon}{\tau}) - \alpha^T(\eta_i^{j+1}), \end{aligned} \tag{3.20}$$

where $\alpha, \beta, \gamma \geq 0$ are the Lagrange multipliers. Using the K.K.T. conditions, we can write:

$$v = -(H^T H + C)^{-1} G^T(\beta - \gamma), \tag{3.21}$$

where $v = [w_i^{j+1} \quad b_i^{j+1}]^T$, $G = D_i[\hat{A}_i \quad e]$ and $H = [A_i \quad e]$. The matrix $(H^T H + C)^{-1}$ is invertible as it is non-singular, which can easily be shown by considering its determinant and showing it to be strictly positive. We now apply K.K.T. conditions to get our dual problem of (3.18) as follows:

$$\begin{aligned} \min_{\gamma-\beta} \quad & \frac{1}{2}(\beta - \gamma)^T G(H^T H + C)^{-1} G^T(\beta - \gamma) - (\beta - \gamma)^T e + e\epsilon(\gamma^T + \frac{\beta^T}{\tau}), \\ \text{s.t.} \quad & \alpha, \beta, \gamma \geq 0, \quad c_1 e = \gamma + \alpha + \frac{\beta}{\tau}. \end{aligned} \tag{3.22}$$

Solving the above dual we get $[w_i \quad b_i]$ for all $i = 1, 2, ..., k$ i.e. k-cluster center planes and we can easily assign a label to any new data point x_t by the equation:

$$y(x_t) = arg \min_{i=1,2,...,k} |w_i^T x_t + b_i|. \tag{3.23}$$

3.2.2 Nonlinear pinSTBSVC

Using the kernel trick, we can easily extend our model to non-linear cases, i.e. we can generate non-linear surfaces $K(x, A)y_i + b_i = 0; i = 1, 2, ..., k$, where $K(.,.)$ is an appropriate kernel-function, by solving the following optimization problem:

$$\begin{aligned} \min_{y_i, b_i, \eta_i} \quad & \frac{1}{2}||K(A_i, A)y_i + b_i e||^2 + c_1 e^T \eta_i + \frac{1}{2} c_2 ||y_i||^2 \\ \text{s.t.} \quad & |K(\hat{A}_i, A)y_i + b_i e| \geq e - \eta_i - e\epsilon, \\ & |K(\hat{A}_i, A)y_i + b_i e| \leq e + \frac{\eta_i}{\tau} + e\frac{\epsilon}{\tau}, \\ & \eta_i \geq 0. \end{aligned} \tag{3.24}$$

In a similar way to the linear case, we can apply CCCP to break the problem into a series of convex quadratic sub-problems and then apply K.K.T. conditions. We get the following equations:

$$v = -(R^T R + C)^{-1} Q^T (\beta - \gamma), \tag{3.25}$$

where $R = [K(A_i, A) \quad e]$, $Q = D_i[K(\hat{A}_i, A) \quad e]$ and $v = [y_i^{j+1} \quad b_i^{j+1}]^T$.

So, the dual of problem (3.24) can be formulated likewise linear case as:

$$\begin{aligned} \min_{\gamma - \beta} \quad & \frac{1}{2}(\beta - \gamma)^T Q (R^T R + C)^{-1} Q^T (\beta - \gamma) - (\beta - \gamma)^T e + e\epsilon(\gamma^T + \frac{\beta^T}{\tau}), \\ \text{s.t.} \quad & \alpha, \beta, \gamma \geq 0, \quad c_1 e = \gamma + \frac{\beta}{\tau} + \alpha. \end{aligned} \tag{3.26}$$

3.3 Theoretical justifications

In this section, we discuss some theoretical aspects of the proposed algorithms.

3.3.1 Matrix $(H^TH + C)$ is invertible

Proof: The theorem is proved by evaluating the determinant of the above matrix as follows:

$$\begin{aligned}\det\left(H_i^TH_i + C\right) &= \begin{vmatrix} A_i^TA_i + c_2I_n & A_i^Te \\ e^TA_i & e^Te \end{vmatrix} \\ &= |e^Te||A_i^TA_i + c_2I_n - A_i^Te(e^Te)^{-1}e^TA_i|, \\ &= m|A_i^TA_i + c_2I_n - \frac{1}{m}A_i^Tee^TA_i|, \\ &= m\left|A_i^T\left(I - \frac{1}{m}ee^T\right)A_i + c_2I_n\right|. \end{aligned} \tag{3.27}$$

We have $(I-\frac{1}{m}ee^T)$ is a symmetric and idempotent matrix i.e. $(I-\frac{1}{m}ee^T)^2 = (I-\frac{1}{m}ee^T)$. This gives us

$$A_i^T\left(I - \frac{1}{m}ee^T\right)A_i = \left[\left(I - \frac{1}{m}ee^T\right)A_i\right]^T\left(I - \frac{1}{m}ee^T\right)A_i, \tag{3.28}$$

which is positive semi-definite and it gives that $A_i^T(I - \frac{1}{m}ee^T)A_i + c_2I_n$ is positive definite as c_2I_n is positive definite. Hence, the determinant in (3.27) is greater than 0. So, $(H^TH + C)$ is an invertible matrix.

Thus, we get an advantage in pinTBSVC and pinSTBSVC over pinTSVC by adding an extra regularization term.

3.3.2 Noise insensitivity and sparsity

The ϵ-insensitive pinball loss function used in the proposed pinSTBSVC is not differentiable at $x = \epsilon$ and $x = \frac{-\epsilon}{\tau}$. So, to solve the QPP of our problem, we need its

sub-gradient, which is defined as

$$g_\tau^\epsilon(x) = \begin{cases} 1, & x > 0, \\ [0,1], & x = \epsilon, \\ 0, & \frac{-\epsilon}{\tau} < x < \epsilon, \\ [0,1], & x = \frac{-\epsilon}{\tau}, \\ -\tau, & x < \frac{-\epsilon}{\tau}. \end{cases} \tag{3.29}$$

Also we split the data-points belonging to the i^{th} cluster into the five groups as shown below:

$$\begin{aligned} T_1 &= \{l : 1 - |w_i^T x_l + b_i| > \epsilon\}, \\ T_2 &= \{l : 1 - |w_i^T x_l + b_i| = \epsilon\}, \\ T_3 &= \{l : -\frac{\epsilon}{\tau} < 1 - |w_i^T x_l + b_i| < \epsilon\}, \\ T_4 &= \{l : 1 - |w_i^T x_l + b_i| = -\frac{\epsilon}{\tau}\}, \\ T_5 &= \{l : 1 - |w_i^T x_l + b_i| < -\frac{\epsilon}{\tau}\}. \end{aligned}$$

As we increase the pinball loss parameter τ, the number of points in the set T_5 increase, as the points in this set are $1+\frac{\epsilon}{\tau}$ distance far from i^{th} plane as seen from the construction of the sets. Now, since the sub-gradient $g_\tau^\epsilon(x)$ for this interval is non-zero, gives us that data points in set T_5 leads to the final solution. As a result, larger values of τ reduce the model's sensitivity to noise around hyper-plane.

Further, we can also observe that the points lying in the interval $(-\frac{\epsilon}{\tau}, \epsilon)$ i.e. the points in the set T_3, are non-effective in the final solution as the sub-gradient $g_\tau^\epsilon(x) = 0$. So, the solution attains sparsity due to the ϵ-insensitive zone.

3.3.3 Time complexity analysis

Solving the optimization problem of the algorithms is the main source of computation cost. The proposed pinTBSVC has two components to its optimization problem:

solving the QPP and computing the inverse of the matrix $(H^T H + C)$. These two parts are repeated for each iteration until the convergence criteria is met. Let t be the number of iterations required and the number of points in the QPP is m, the convex QPP has a time complexity of $O(m^3/4)$, while the matrix inverse of a $n \times n$ matrix has a time complexity of $O(n^3)$. Thus, the net time complexity of pinTBSVC will be $O(t(m^3/4 + n^3))$. The convergence criteria is reached when $|||[w_i^{j+1}; b_i^{j+1}] - [w_i^j; b_i^j]|||$ is less than our chosen error tolerance value. Thus, for a lower tolerance value, the number of iterations, i.e. t, will increase. In the cases of TWSVC [18] and pinTSVC [24], we have a similar convex QPP with m points, the difference being the Matrix Inversion; Instead of $H^T H + C$, we invert $H^T H$ in TWSVC [18]. The additional C matrix doesn't contribute to the time complexity as the resultant matrix is of the same size. Similarly, in TBSVC [20], the difference comes in the loss function, which doesn't play a significant role in the time complexity. Thus the time complexities of TWSVC [18], TBSVC [20], pinTSVC [24] and pinTBSVC are equivalent.

The pinSTBSVC algorithm, like the previously mentioned pinTBSVC, uses a similar QPP solution, with the exception being in the number of constraints. The number of constraints in pinSTBSVC is double that of TWSVC, resulting in a time complexity $O((2 \times m)^3/4) = O(2 \times m^3)$. The matrix inversion step involves a matrix of the same size; thus, time complexity to solve the matrix inverse remains the same, i.e. $O(n^3)$. Thus, the net time complexity for pinSTBSVC is $O(t(2 \times m^3 + n^3))$.

Chapter 4

Numerical Experiments and Statistical Analysis

In this chapter, we will provide results of numerical experiments of the proposed model and discuss the statistical analysis of the experimental results.

4.1 Experiments

In this section, we compare the performance of the proposed algorithms against the existing plane based clustering algorithms like TWSVC [18], TBSVC [20] and pinTSVC [24] on several benchmark UCI datasets [16]. The results are reported with accuracies for real-world benchmark datasets having different levels of noise. we use the Gaussian noise into the datasets with four values of standard deviation (σ). 5 fold cross validation has been used for all experiments with grid search method to find optimal parameters. For non-linearly separable datasets, Gaussian kernel $K(x, y) = exp(-||x - y||^2/\mu^2)$ has been used. Additionally, the initialization is done via Nearest Neighbour Graph (NNG) [18]. All the methods were implemented on MATLAB® R2017a [29] and have been tested on a High-Performance Computer with a Intel® Xeon® E5-2697 v4 Processor @ 2.30 GHz speed and 128 GB RAM.

4.1.1 Parameters selection

The range and values for the various parameters involved in the various methods have been tabulated in Table 4.1.

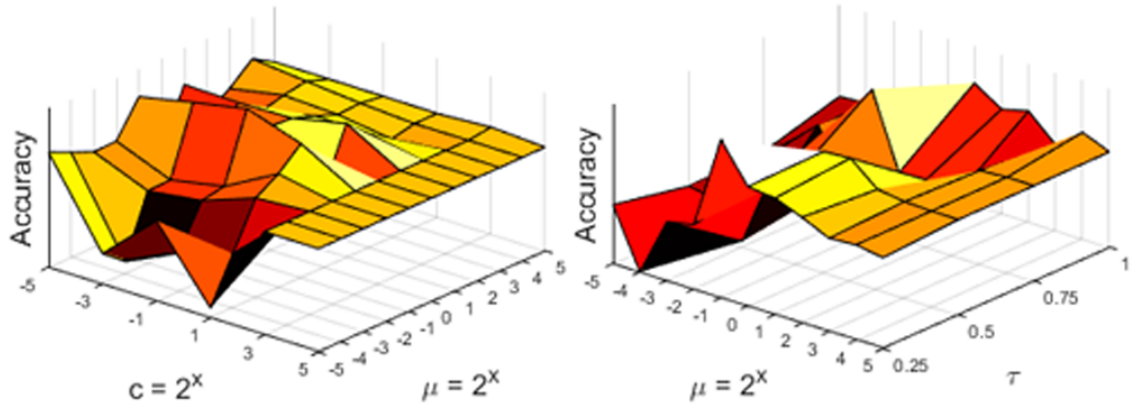
Parameter Name	Symbol	Range/Value
Penalty Parameters	c or c_1	$\{2^i \mid i = -5, -3, -1, 1, 3, 5\}$
	c_2	$\{2^i \mid i = -5, -3, -1, 1, 3, 5\}$
Gaussian Kernel Parameter	μ	$\{2^i \mid i = -5, -4, ...4, 5\}$
Pinball Loss Parameter	τ	$\{0.25, 0.50, 0.75, 1\}$
	ϵ	$\{0.1, 0.3, 0.5\}$
	δ	10^{-4}

TABLE 4.1: Range of parameters for various methods.

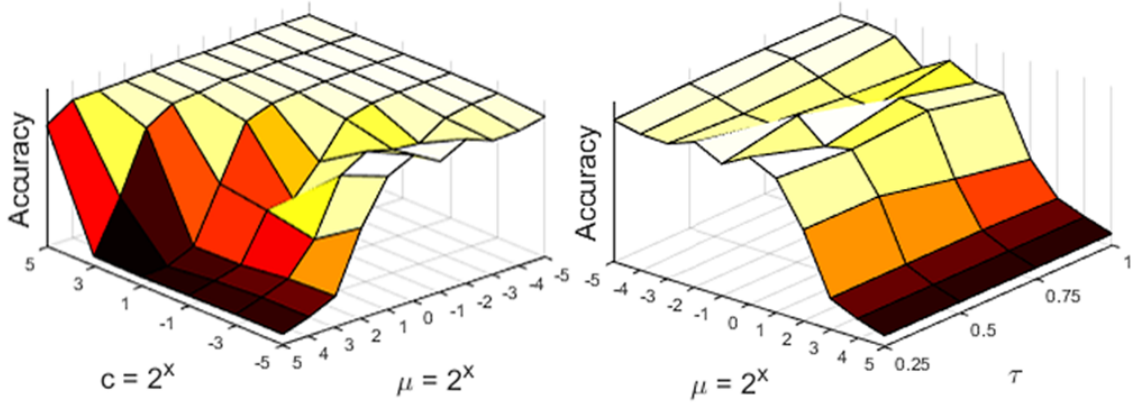
Figures 4.1 and 4.2 give the plot of accuracy for various parameters. For a clear plot with minimal clutter, we have plotted accuracy vs two parameters at a time; for the other parameters (e.g. τ and ϵ in case of the plot: accuracy vs c and μ) we choose the respective overall optimal choice. This was done since we have a total of 4 parameters in pinTBSVC and 5 parameters in pinSTBSVC.

4.1.2 Discussion of the results

The accuracy along with the rank for all the datasets are tabulated in Table 4.2.

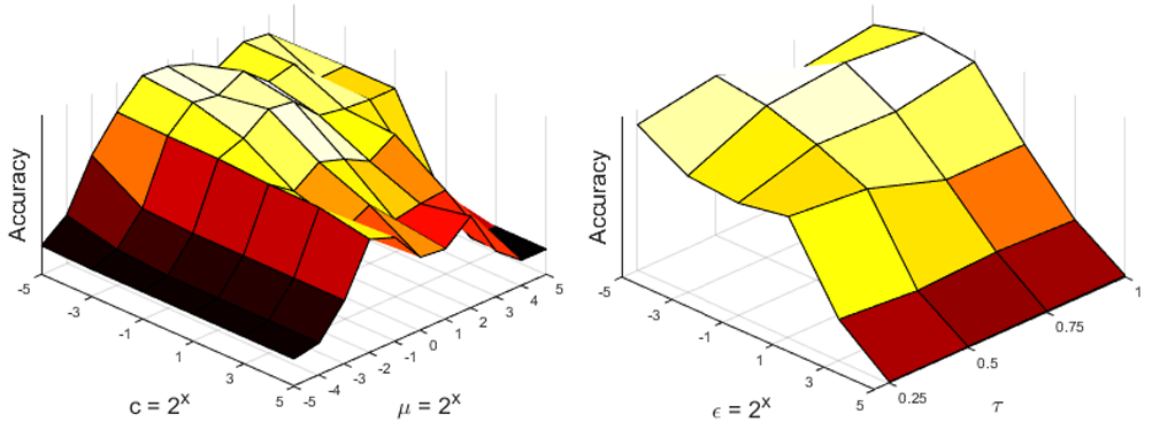


(i) Haberman

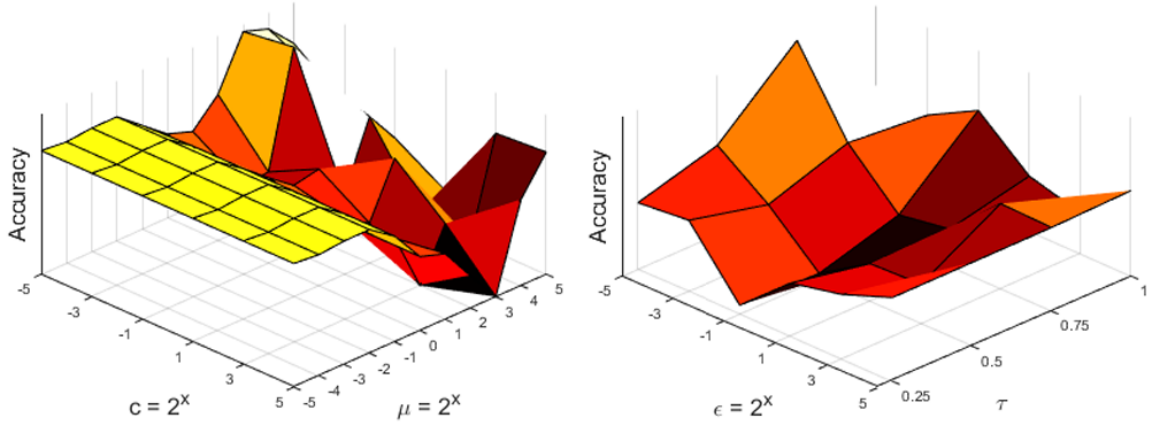


(iii) Libras

FIGURE 4.1: Surface plots illustrating the effectiveness of pinTBSVC with various parameters



(i) Iris



(iii) Haberman

FIGURE 4.2: Surface plots illustrating the effectiveness of pinSTBSVC with various parameters

TABLE 4.2: Accuracies obtained for different methods on various datasets with four noise levels

Data	σ	TWSVC [18]	TBSVC [20]	pinTSVC [24]	pinTBSVC	pinSTBSVC
aa-iris	0	94.9425	93.5632	95.6782	95.7241	**96.5057**
	0.05	**87.2184**	83.5402	85.6552	86.2069	87.0345
	0.075	78.8506	89.8851	87.6782	**91.2644**	90.1149
	0.1	85.1034	85.6092	84.7356	**86.5287**	84
aaa-balloons	0	76.6667	76.6667	76.6667	**90**	**90**
	0.05	90	76.6667	73.3333	73.3333	**100**
	0.075	**86.6667**	73.3333	**86.6667**	76.6667	76.6667
	0.1	76.6667	**86.6667**	76.6667	**86.6667**	**86.6667**
haberman	0	58.8366	59.6249	56.1389	64.6163	**64.7164**
	0.05	61.5561	60.4759	61.5561	65.1627	**65.8135**
	0.075	59.2977	59.1623	61.5561	**64.5725**	64.2665
	0.1	61.5561	60.1135	61.5561	64.318	**64.4632**
hepatitis	0	69.0323	71.6989	66.4516	76.2581	**76.6022**
	0.05	69.3763	70.5806	74.2796	**76.2581**	74.7097
	0.075	69.4624	66.7957	69.4624	**74.6237**	74.5376
	0.1	69.2903	69.6344	64.4731	**75.1398**	73.5914
lense	0	68	82	78	82	**88**
	0.05	71.3333	72	70	66.6667	**80**
	0.075	62	80	80	70	**85.3333**
	0.1	76	78	**86**	84	81.3333

Continuation of Table 4.2 ...

Dataset	σ	TWSVC [18]	TBSVC [20]	pinTSVC [24]	pinTBSVC	pinSTBSVC
new-thyroid	0	77.5858	79.6013	85.6478	85.1606	**88.3278**
	0.05	85.0941	81.5061	85.0941	86.4452	**88.1506**
	0.075	86.0687	85.2492	86.0244	84.5626	**88.1063**
	0.1	84.5404	84.2525	83.6766	**87.1761**	86.2016
pathbased	0	98.1695	95.6497	98.226	**98.7232**	96.4972
	0.05	94.4068	91.2768	93.6723	**94.6893**	90.8814
	0.075	89.5932	88.9153	**94.9266**	93.9774	92.1921
	0.1	91.8192	92.3277	93.6723	**93.8192**	91.5028
spherical-4-3	0	**100**	**100**	99.4367	**100**	99.7152
	0.05	99.2405	99.2785	99.7468	**100**	99.2342
	0.075	99.7342	99.7658	**100**	**100**	97.7658
	0.1	99.7405	**99.7658**	98.5443	**99.7658**	98.7278
spherical-5-2	0	97.1429	93.6816	93.5347	**97.9592**	97.8122
	0.05	87.6571	89.649	89.698	90.0571	**91.9388**
	0.075	85.1592	86.1551	87.3633	86.9388	**88.6367**
	0.1	85.0776	85.9102	86.6449	**87.3143**	85.5673
zz-ionosphere	0	66.4968	57.2618	75.4944	86.8302	**91.2509**
	0.05	67.4843	81.853	78.0524	86.7933	**89.323**
	0.075	72.6321	87.6816	84.6842	88.7681	**89.1573**
	0.1	75.7262	75.8757	81.2338	86.8534	**88.173**
Avg. Accuracy		80.3806	81.2919	82.2982	84.6460	**86.0879**

TABLE 4.3: Ranks of different methods on various datasets based on their performance

Datasets	σ	TWSVC [18]	TBSVC [20]	pinTSVC [24]	PinTBSVC	SPTBSVC
aa_iris	0	4	5	3	2	1
	0.05	1	5	4	3	2
	0.075	5	3	4	1	2
	0.1	3	2	4	1	5
aaa_balloons	0	4	4	4	1.5	1.5
	0.05	2	3	4.5	4.5	1
	0.075	1.5	5	1.5	3.5	3.5
	0.1	4.5	2	4.5	2	2
haberman	0	4	3	5	2	1
	0.05	3.5	5	3.5	2	1
	0.075	4	5	3	1	2
	0.1	3.5	5	3.5	2	1
hepatitis	0	4	3	5	2	1
	0.05	5	4	3	1	2
	0.075	3.5	5	3.5	1	2
	0.1	4	3	5	1	2
lense	0	5	2.5	4	2.5	1
	0.05	3	2	4	5	1
	0.075	5	2.5	2.5	4	1
	0.1	5	4	1	2	3
new-thyroid	0	5	4	2	3	1
	0.05	3.5	5	3.5	2	1
	0.075	2	4	3	5	1
	0.1	3	4	5	1	2
pathbased	0	3	5	2	1	4
	0.05	2	4	3	1	5
	0.075	4	5	1	2	3

TABLE 4.3: Ranks of different methods on various datasets based on their performance

Datasets	σ	TWSVC [18]	TBSVC [20]	pinTSVC [24]	PinTBSVC	SPTBSVC
	0.1	4	3	2	1	5
spherical_4_3	0	2	2	5	2	4
	0.05	4	3	2	1	5
	0.075	4	3	1.5	1.5	5
	0.1	3	1.5	5	1.5	4
spherical_5_2	0	3	4	5	1	2
	0.05	5	4	3	2	1
	0.075	5	4	2	3	1
	0.1	5	3	2	1	4
zz-ionosphere	0	4	5	3	2	1
	0.05	5	3	4	2	1
	0.075	5	3	4	2	1
	0.1	5	4	3	2	1
Average Rank		3.7750	3.6625	3.3375	**2.0250**	**2.2000**

4.2 Applications

To see the applications of the proposed algorithms in real world problems, we apply the proposed model to real-world benchmark datasets and compare it with other baseline models.

1. Breast cancer clustering - For application in breast cancer clustering, we used the Wisconsin Diagnostic Dataset [16]. This dataset was made by extracting features from digitized images of fine-needle aspirates (FNA) of a breast mass. It has 569 instances, of which 357 are benign, and 212 are malignant. The features describe the cell nuclei present in the image in 3d space. From Table 4.4, we see that the best performing model is pinTBSVC with accuracy of 87.839.

2. Marketing data clustering [30]- This dataset consists of data regarding some supermarket mall customers, like their age, gender, annual income, and spending score. We encoded the labels and divided the 0-100 spanning spending score into 4 sections 0-25, 26-50, 51-75 and 76-100; representing low, low-medium, medium-high and high spenders. From Table 4.5, we see that the best performing model is pinSTBSVC with accuracy of 58.5385.

TWSVC [18]	TBSVC [20]	pinTSVC [24]	pinTBSVC	pinSTBSVC
81.9717	86.3883	74.7674	**87.839**	86.9969

TABLE 4.4: Results for breast cancer clustering

TWSVC [18]	TBSVC [20]	pinTSVC [24]	pinTBSVC	pinSTBSVC
53.2564	56	56.4103	56.7692	**58.5385**

TABLE 4.5: Results for marketing data clustering

To illustrate the formation of clusters via different methods, we have also visualized the dataset "spherical-5-2" which only has two features, for the convenience of a 2d graph; the various clusters formed by the different methods are showcased in Figure 4.3.

4.3 Statistical analysis

For statistical analysis, we use the Friedman test [31] and the accompanying post-hoc test to statistically analyse our experimental findings. We need average ranks of the algorithms under consideration based on their accuracies on various datasets for this. Average ranks of algorithms are tabulated in Table 2. First we consider the null hypothesis that all algorithms have similar performance. Now, we calculate the χ^2 as

$$\chi^2 = \frac{12N}{p(p+1)}\left[\sum_{i=1}^{p} R_i^2 - \frac{p(p+1)^2}{4}\right], \tag{4.1}$$

where N is the number of datasets, p is number of algorithms considered and R_i is the average rank of i^{th} algorithm.

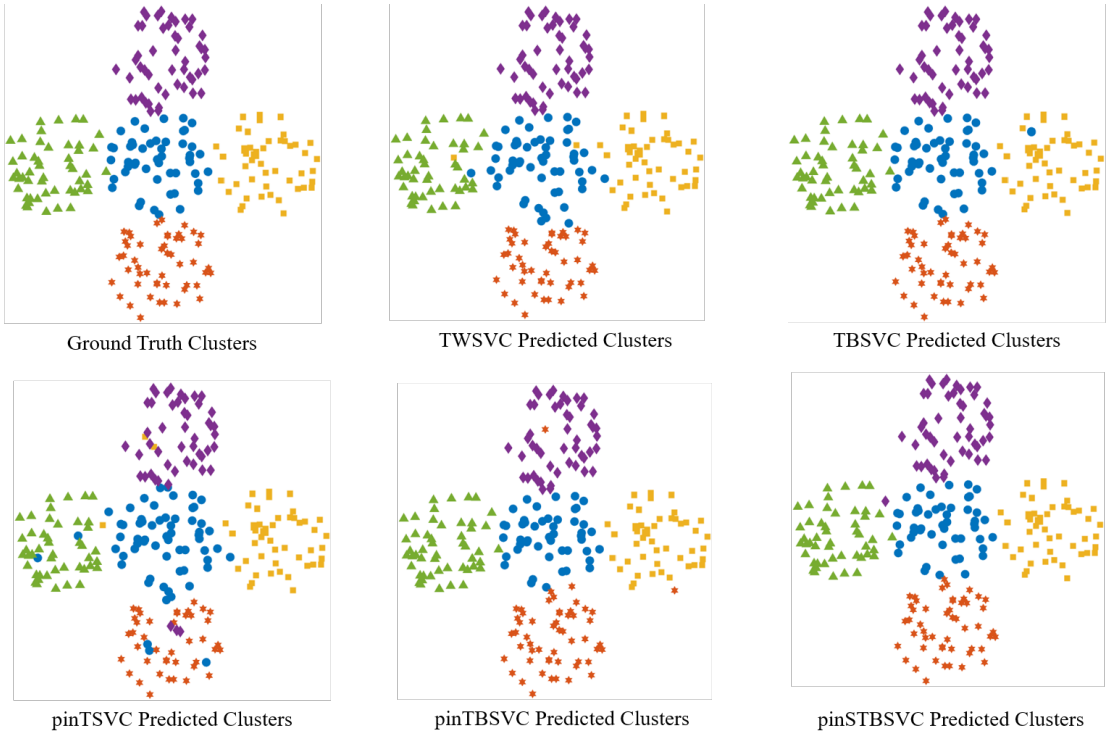


FIGURE 4.3: Formation of clusters by various methods

For $N = 40$, $p = 5$ and R_i from the Table-4.3, we get

$$\begin{aligned}\chi^2 &= \frac{12 \times 40}{5 \times 6}\left[(3.7750)^2 + (3.6625)^2 + (3.3375)^2 + (2.0250)^2 + (2.2000)^2 - \frac{5 \times 36}{4}\right] \\ &\approx 43.905, \\ \text{and} \quad F_F &= \frac{(N-1)\chi^2}{N(p-1) - \chi^2} = \frac{(40-1) \times 43.905}{40 \times 4 - 43.905} \\ &\approx 14.749.\end{aligned} \tag{4.2}$$

F-distribution has degrees of freedom $(p-1, (p-1)(N-1))$. For $F(4, 156)$ with the level of significance $\alpha = 0.05$ the critical value is 2.425. And since $F_F = 14.749 > 2.425$, we reject the null hypothesis.

Further, we use the Nemenyi post-hoc test [32] to look for statistically significant differences between algorithms. For this, we calculate the critical difference (CD) with

critical value as q_α

$$CD = q_\alpha \sqrt{\frac{p(p+1)}{6N}} \tag{4.3}$$
$$CD = 2.728 \times \sqrt{\frac{5 \times 6}{6 \times 40}} \approx 0.9645$$

Therefore, if the average ranks of two algorithms differ by at least $CD \approx 0.9645$, the Nemenyi test [32] indicates that there is a significant difference between them. As a result, we can say that the proposed algorithms differs significantly from existing plane based clustering algorithms.

Chapter 5

Conclusions and Future Directions

In this work, we introduced two novel plane based clustering algorithms, PinTBSVC and PinSTBSVC, which significantly improves the performance of existing algorithms. The pinball loss function used in both the proposed algorithms provides stability for re-sampling of data and benefits the datasets with noise. Furthermore, we incorporated the maximum margin regularization term that made the planes to be as far as possible and improved the accuracy of the algorithms. Numerical experiments on various benchmark datasets demonstrate the advantage of the proposed models over existing plane-based clustering algorithms.

In future, one can work on parameter selection techniques required to find the optimal parameters for the algorithms. Currently, in order to optimize performance, one must evaluate a large number of parameters, due to the lack of a convenient option for searching parameters. Our proposed pinSTBSVC has excellent sparsity and performance, but it comes at the cost of increasing its time complexity which is an impediment when solving for large datasets. Thus, one can also look into ways to decrease this time complexity.

Bibliography

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[2] P. Bradley and O. Mangasarian, "Massive data discrimination via linear support vector machines," *Optimization methods and software*, vol. 13, no. 1, pp. 1–10, 2000.

[3] H. Zhu, X. Liu, R. Lu, and H. Li, "Efficient and privacy-preserving online medical prediagnosis framework using nonlinear SVM," *IEEE journal of biomedical and health informatics*, vol. 21, no. 3, pp. 838–850, 2016.

[4] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in *Proceedings of IEEE computer society conference on computer vision and pattern recognition*. IEEE, pp. 130–136, 1997.

[5] Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905–910, 2007.

[6] R. Ilin, "Unsupervised learning of categorical data with competing models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 11, pp. 1726–1737, 2012.

[7] R. Xu and D. Wunsch, "Data visualization and high-dimensional data clustering," in *Clustering*. Wiley-IEEE Press, pp. 237–261, 2009.

[8] K. Inkim, J. Hyungkim, and Keechuljung, "Face recognition using support vector machines with local correlation kernels," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 16, 11, 2011.

[9] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[10] E. Mahima Jane and E. George Dharma Prakash Raj, "Sbkmeda: Sorting-based k-median clustering algorithm using multi-machine technique for big data," in *Advances in Big Data and Cloud Computing*, E. B. Rajsingh, J. Veerasamy, A. H. Alavi, and J. D. Peter, Eds., pp. 219–225, 2018.

[11] P. S. Bradley and O. L. Mangasarian, "k-plane clustering," *Journal of Global Optimization*, vol. 16, no. 1, pp. 23–32, Jan 2000.

[12] Y.-H. Shao, L. Bai, Z. Wang, X.-Y. Hua, and N.-Y. Deng, "Proximal plane clustering via eigenvalues," *Procedia Computer Science*, vol. 17, pp. 41–47, 2013.

[13] Y.-H. Shao, C.-H. Zhang, X.-B. Wang, and N.-Y. Deng, "Improvements on twin support vector machines," *IEEE Transactions on Neural Networks*, vol. 22, no. 6, pp. 962–968, 2011.

[14] M. Tanveer, "Application of smoothing techniques for linear programming twin support vector machines," *Knowledge and Information Systems*, vol. 45, no. 1, pp. 191–214, 2015.

[15] M. Tanveer, M. Asif Khan, and S. S. Ho, "Robust energy-based least squares twin support vector machines," *Applied Intelligence*, vol. 45, p. 174–186, 07 2016.

[16] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[17] M. Tanveer, C. Gautam, and P. N. Suganthan, "Comprehensive evaluation of twin SVM based classifiers on UCI datasets," *Applied Soft Computing*, vol. 83, p. 105617, 2019.

[18] Z. Wang, Y.-H. Shao, L. Bai, and N.-Y. Deng, "Twin support vector machine for clustering," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 10, pp. 2583–2588, 2015.

[19] S. Moezzi, M. Jalali, and Y. Forghani, "TWSVC+: Improved Twin Support Vector Machine-Based Clustering," *Ingénierie des systèmes d information*, vol. 24, no. 5, pp. 463–471, 2019.

[20] L. Bai, Y.-H. Shao, Z. Wang, and C.-N. Li, "Clustering by twin support vector machine and least square twin support vector classifier with uniform output coding," *Knowledge-Based Systems*, vol. 163, pp. 227–240, 2019.

[21] Z. Wang, X. Chen, Y.-H. Shao, and C.-N. Li, "Ramp-based twin support vector clustering," *Neural Computing and Applications*, vol. 32, no. 14, pp. 9885–9896, Jul 2020.

[22] B. Richhariya, M. Tanveer, and A. D. N. Initiative, "Least squares projection twin support vector clustering (LSPTSVC)," *Information Sciences*, vol. 533, pp. 1–23, 2020.

[23] X. Huang, L. Shi, and J. A. Suykens, "Support vector machine classifier with pinball loss," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 984–997, 2013.

[24] M. Tanveer, T. Gupta, and M. Shah, "Pinball loss twin support vector clustering," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17(2s), 2021.

[25] M. Tanveer, T. Gupta, M. Shah, and B. Richhariya, "Sparse twin support vector clustering using pinball loss," *IEEE Journal of Biomedical and Health Informatics*, 2021, doi= 10.1109/JBHI.2021.3059910.

[26] M. Tanveer, A. Tiwari, R. Choudhary, and S. Jalan, "Sparse pinball twin support vector machines," *Applied Soft Computing*, vol. 78, pp. 164–175, 2019.

[27] A. L. Yuille, A. Rangarajan, and A. Yuille, "The concave-convex procedure (CCCP)," *Advances in Neural Information Processing Systems*, vol. 2, pp. 1033–1040, 2002.

[28] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2013.

[29] MATLAB, *9.2.0.556344 (R2017a)*. Natick, Massachusetts: The MathWorks Inc., 2019.

[30] V. Choudhary, "Mall customer segmentation data," Aug 2018. [Online]. Available: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

[31] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[32] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.