Communications for Statistical Applications and Methods
2023, Vol. 30, No. 3, 245–258

https://doi.org/10.29220/CSAM.2023.30.3.245
Print ISSN 2287-7843 / Online ISSN 2383-4757

Modifying linearly non-separable support vector machine binary classifier to account for the centroid mean vector

Mubarak Al-Shukeili [1,a], Ronald Wesonga[ab]

[a]Department of Statistics, Sultan Qaboos University, Muscat, Oman
[b]Data Science Analytics Lab, Department of Statistics, Sultan Qaboos University, Muscat, Oman

Abstract

This study proposes a modification to the objective function of the support vector machine for the linearly non-separable case of a binary classifier $y_i \in \{-1, 1\}$. The modification takes into account the position of each data item $\mathbf{x_i}$ from its corresponding class centroid. The resulting optimization function involves the centroid mean vector, and the spread of data besides the support vectors, which should be minimized by the choice of hyper-plane β. Theoretical assumptions have been tested to derive an optimal separable hyperplane that yields the minimal misclassification rate. The proposed method has been evaluated using simulation studies and real-life COVID-19 patient outcome hospitalization data. Results show that the proposed method performs better than the classical linear SVM classifier as the sample size increases and is preferred in the presence of correlations among predictors as well as among extreme values.

Keywords: support vector machine, quadratic cost function, misclassification rate, linearly non-separable, centroid mean vector

1. Introduction

Support vector machine (SVM) classifiers have a sustainable performance on regression and classification, especially for high-dimensional features extraction (Izenman, 2008). These classifiers use support vectors, which provide efficient information, in constructing a linear hyperplane when data sets are completely separable. However, in the case of non-separable data, the classifier becomes more difficult and complex to build. Distinct types of SVM's algorithms deal with two situations of non-separable data. Firstly, when the data are partially overlapped but can be separated linearly ,and secondly, when the data of one class are mixed or surrounded due to the differences in classes' homogeneities. As a consequence of this data structure, linear separable hyperplane is impossible or at least works poorly under classification, even when the kernel trick that uses a specific transformation function to convert the non-linear separable data to linearly separable is applied (Izenman, 2008). Despite these differences of data distributions, the SVM classifier still utilizes the same quadratic programming (QP) objective function regardless of the situation to solve constrained optimization problems (Izenman, 2008; Jiang *et al*., 2014). Consequently, this leads to higher misclassification rates.

Our target was to modify the objective function of the classical SVM so as to minimize the misclassification rate. In addition, we investigated its properties in terms of consistency and convergence

[1] Corresponding author: Department of Statistics, College of Science, Sultan Qaboos University, Muscat, Oman. Email: mubaraklife10@gmail.com

Published 31 May 2023 / journal homepage: http://csam.or.kr

(Steinwart, 2001, 2002; Zhang, 2004; Vert *et al*., 2006). To minimize the criteria of the misclassified rate and outliers, some studies have introduced a marginal loss for the data by computing the location of the data point and its corresponding marginal hyperplane; and thereby attempted to reduce the misclassification rate (Bühlmann and Yu, 2003). To deal with the outlier sensitivity, the ramp loss classification approach was discussed during the proceedings of the 23^{rd} international conference on machine learning (Brooks, 2011; Shen *et al*., 2003; Collobert *et al*., 2006). They attempted to solve the optimization problem with ramp loss so as to obtain solutions that did not guarantee to global minima. The idea of ramp loss was to fix the misclassified points, especially for the outliers; and thereby reduce the model's sensitivity towards them, hence minimizing the misclassification rate. The other method used has been the quadratic optimization with the hard margin loss, which is used to minimize the number of misclassification points by giving 1 to misclassified points or to those that lie in the margin and 0 loss to the correctly classified (Wang *et al*., 2021). Further, a discrete SVM classifier has been formulated using the hard margin loss with the linear kernel (Orsenigo and Vercellis, 2003).

Incidentally, many of these cost functions can replace the costs needed to be minimized for every data item $\mathbf{x_i}$; for example, the step loss function based SVM classifier for binary classification (Jarray *et al*., 2018). This method overcomes the issue of the hard and ramp losses by introducing different costs for each instance according to where the points are located. Considerations are made as to whether the points lie within or outside the margin, and also whether they tend to be correctly classified or misclassified (Jarray *et al*., 2018; Brooks, 2011).

One of the most frequent problems in efficient classification or prediction for regression is having collinearity among predictors (Dormann *et al*., 2013; Siqueira *et al*., 2018; Han *et al*., 2013). This phenomenon reflects difficulty in finding a linear separable hyperplane as well as increased computational time because of the large sets of support vectors. However, although deriving principal components for some datasets could reduce the association (Rencher, 2003), this is still not enough to accelerate the computation time and minimize the misclassification. The main focus of our study was to improve the objective function of the classical SVM classifier by incorporating the centroid and marginal costs. Other specific contributions of the study are reflected in the theoretical derivation of the appropriate centroid cost function. This study proposes a novel SVM classifier that aggregates centroid and marginal costs for the non-separable data. The performance of our novel approach was evaluated against the classical SVM using simulation studies and an application on multivariate classification of COVID-19 patient hospitalization outcomes. Further validation is done to check the robustness of the proposed method in the presence of outliers or extreme data on the misclassification rate.

2. Method

2.1. Formulation of the centroid cost

The linear classifier $f(x_i) = \beta^T x_i + \beta_0, i = 1, 2, \ldots, n.$ of the support vector machine, SVM takes the data point $\mathbf{x_i} \in \Re^p$ to one of the two classes $y_i \in \{-1, 1\}$. In the non-separable case, data points are mainly overlapped rendering it difficult to find a linear hyperplane β, which can easily separate the data points, and therefore lead to high misclassification rates. One possibility was to introduce slack variables for each data point to measure the distance between each data point and its corresponding marginal plane. These distances that need to be minimized are associated with a free penalized parameter (C).

In our case, we introduced a cost function that measures all the individual distances from their

corresponding centroids. Accordingly, the most suitable cost function should take into account the variation of distances from the centroid. We used the quadratic cost function $(z_i - E(z_i))^2$, where $E(z_i)$ is the expected value of the data value $i, i = 1, \ldots, n$ (Here $E(z_i) = \bar{z}, \bar{z} = \sum_{i=1}^{n} z_i/n$), because of its desirable properties. The enhanced optimization system for this compounding cost helps us to find the optimal separable hyperplane β. The enhanced optimization system was set up as below:

1. The marginal distance between the two hyperplanes $y = -1$ and $y = +1$ is $2/\| \beta \|^2$ that needs to be maximized, which as a convex minimization problem becomes:

$$\text{Min}_{\beta} \frac{\|\beta\|^2}{2}.$$

2. The classical *linear non-separable hyperplane for the SVM* classifier should be subjected to the following well known optimization problem:

$$\text{classical}_{\text{SVM}} = \begin{cases} \text{Min}_{\beta,\zeta_i} \dfrac{1}{2} \|\beta\|_2^2 + C \displaystyle\sum_{i=1}^{n} \zeta_i \\ \quad s.t., \\ y_i \left(\beta_0 + \beta^T x_i\right) \geq 1 - \zeta_i, \quad i = 1, \ldots, n \\ \quad \zeta_i \geq 0. \end{cases} \tag{2.1}$$

3. It is required to minimize the compounding cost, which measures the risk that a data point is misclassified. There are two kinds of costs responsible for this risk; the marginal cost, $C \sum_{i=1}^{n} \zeta_i$, and the centroid cost $(z_i - \bar{z}_i)^2$, where $z_i = \beta_0 + \beta^T x_i$.

4. Compounding the two costs in one minimization problem, gives the proposed SVM:

$$\text{proposed}_{\text{SVM}} = \begin{cases} \text{Min}_{\beta,\zeta} \dfrac{1}{2} \|\beta\|_2^2 + C \displaystyle\sum_{i=1}^{n} \zeta_i + \text{Centroid} \\ \quad \mathbf{s.t.}, \\ y_i \left(\beta_0 + \beta^T x_i\right) \geq 1 - \xi_i, \quad i = 1, \ldots, n \\ \quad \zeta_i \geq 0. \end{cases} \tag{2.2}$$

5. The proposed SVM works on the assumption that these constituent costs, that is, the *marginal* and *centroid* costs, are independent and not identically distributed, especially when the sample size is small.

2.2. Defining the centroid cost function in terms of β

Our proposed centroid cost function is the *quadratic centroid cost* that replaces the **centroid** term in model 2.2. It represents the sum of all corresponding losses for each $\mathbf{x_i}$. Quadratic cost function is a good choice in measuring the losses because it represents very well the sum of two within covariance matrices, thereby the estimated β may minimize the association between predictors. Since we are seeking to minimize that expression with respect to β, it is a good idea also to express the centroid cost in term of β so that it can be minimized.

We define the algebraic function for the **cost** as follows:

$$\text{Centroid}_{\text{cost}} = \sum_{j=1}^{2} \sum_{i=1}^{n_j} \left(z_{ij} - E\left(z_{ij}\right)\right)^2, \tag{2.3}$$

where $z_{ij} = \beta_0 + \beta^T x_{ij}, E(z_{ij}) = \bar{z_j} = \bar{z_j} = \sum_{i=1}^{n_j} z_{ij}/n_j = \beta^T \bar{x}_j, j = 1, 2$ (ignoring β_0 for simplicity). On substituting these individual expressions in the main expression of centroid cost function (2.3) yields the cost function in terms of β as follows:

$$\begin{aligned}
\text{Centroid}_{\text{cost}}\left(\beta \mid X_{n\times p}\right) &= \sum_{j=1}^{2} \sum_{i=1}^{n_j} \left(\beta^T x_{ij} - \beta^T \bar{x}_j\right)^2 \\
&= \sum_{j=1}^{2} \sum_{i=1}^{n_j} \beta^T \left(x_{ij} - \bar{x}_j\right) \beta^T \left(x_{ij} - \bar{x}_j\right) \\
&= \beta^T \sum_{j=1}^{2} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right) \left(x_{ij} - \bar{x}_j\right)^T \beta \\
&= \beta^T A \beta.
\end{aligned} \tag{2.4}$$

Ultimately, the optimization system for the *linear non-separable support vector machine* can be defined as follows :

$$\text{Proposed}_{\text{SVM}} \;=\; \begin{cases} \text{Min}_{\beta,\,\zeta} \dfrac{1}{2} \|\beta\|_2^2 + C \sum_{i=1}^{n} \zeta_i + \beta^T A \beta \\ \qquad \textbf{s.t.}, \\ y_i(B_0 + \beta^T x_i) \geq 1 - \zeta_i, \quad i = 1, \ldots, n \\ \qquad \zeta_i \geq 0, \\ \qquad \mathbf{A} \;> \mathbf{0}, \end{cases} \tag{2.5}$$

where $\mathbf{A}$ is $p \times p$ (p is number of attributes) symmetric matrix representing the sum of within covariance matrices: $\mathbf{A} = \mathbf{\Sigma_1} + \mathbf{\Sigma_2}$ and C is the parameter that tells the SVM optimization how much you want to avoid misclassifying each training data point. This last model 2.5 requires to be minimized by finding the optimum hyperplane β under the influence of the quadratic centroid cost function.

2.3. Finding the optimal hyperplane β

In order to estimate the parameters $\{\beta, \zeta\}$, we need to construct the **lagrangian primal function** for the constrained optimization system as follows:

$$F_D = \frac{1}{2} \|\beta\|_2^2 + \beta^T A \beta - \sum_{i=0}^{n} \left[\alpha_i \left(y_i \left(\beta_0 + \beta^T x_i\right) - (1 - \zeta_i)\right)\right] - \sum_{i=1}^{n} \eta_i \zeta_i, \tag{2.6}$$

where α_i and η_i are the lagrange multipliers greater than 0. To minimize Equation (2.6), we need to differentiate it with respect to β_0, β and ζ_i; and then, substitute these quantities in the original primal function to be reduced in α and consequently solve for α to obtain its final estimate. Algebraic steps

are shown as follows:

$$\frac{\partial F_p}{\partial \beta_0} = -\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{2.7}$$

$$\frac{\partial F_p}{\partial \beta} = \beta + 2A^T\beta - \sum \alpha_i y_i \mathbf{x_i} = \mathbf{0_{p\times 1}}. \tag{2.8}$$

Solving Equation (2.8) for β yields:

$$\beta = \left(\mathbf{I_p} + 2\mathbf{A^T}\right)^{-1} \sum_{i=0}^{n} \alpha_i y_i \mathbf{x_i}, \tag{2.9}$$

where the expression $(I_p + 2\,A^T)^{-1}$ can be denoted by the $p \times p$ matrix $\mathbf{E}$ under the condition that $\mathbf{A}$ is a non-singular matrix since it is positive definite. Thus, $\beta = E\sum_{i=0}^{n} \alpha_i y_i x_i$.

$$\begin{aligned} \frac{\partial F_p}{\partial \zeta_i} &= C - \alpha_i - \eta_i = 0 \\ &\Rightarrow \alpha_i = C - \eta_i. \end{aligned} \tag{2.10}$$

Now, we substitute Equations (2.7), (2.9) and (2.10) into Equation (2.6) to yield:

$$\begin{aligned} F_D = {} & \frac{1}{2}\left(\sum_{i=0}^{n} \alpha_i y_i x_i^T\right) E^T E \left(\sum_{i=0}^{n} \alpha_i y_i x_i\right) + C\sum_{i=1}^{n} \zeta_i \\ & + \left(\sum_{i=0}^{n} \alpha_i y_i x_i^T\right) E^T A E \left(\sum_{i=0}^{n} \alpha_i y_i x_i\right) - \sum_{i=0}^{n} \alpha_i y_i B_0 \\ & - \left(\sum_{i=0}^{n} \alpha_i y_i x_i^T\right) E \left(\sum_{i=0}^{n} \alpha_i y_i x_i\right) + \sum_{i=0}^{n} \alpha_i - \sum_{i=0}^{n} \alpha_i \zeta_i - \sum_{i=0}^{n} (C - \alpha_i)\,\zeta_i\,, \end{aligned} \tag{2.11}$$

where $E = (I + 2A^T)^{-1}$ is a $p \times p$ symmetric positive definite matrix since $\mathbf{A}$ is symmetric positive definite matrix.

However, Equation (2.11) can be written in vector and matrix form as follows:

$$\begin{aligned} F_D = {} & \frac{1}{2}\alpha_{1\times n}^T \overbrace{\operatorname{diag}\{y_1, y_2, \ldots, y_n\}\ X_{n\times p} E^2 X_{p\times n}^T \operatorname{diag}\{y_1, y_2, \ldots, y_n\}}^{D_{n\times n}} \alpha \\ & + \alpha_{1\times n}^T \overbrace{\operatorname{diag}\{y_1, y_2, \ldots, y_n\}\ X_{n\times p} E^T A E X_{p\times n}^T \operatorname{diag}\{y_1, y_2, \ldots, y_n\}}^{F_{n\times n}} \alpha \\ & - \alpha_{1\times n}^T \overbrace{\operatorname{diag}\{y_1, y_2, \ldots, y_n\}\ X_{n\times p} E X_{p\times n}^T \operatorname{diag}\{y_1, y_2, \ldots, y_n\}}^{G_{n\times n}} \alpha + \sum \alpha_i\,. \end{aligned} \tag{2.12}$$

$$F_D(\alpha) = \frac{1}{2}\alpha^T D_{n\times n}\alpha + \alpha^T F_{n\times n}\alpha - \alpha^T G_{n\times n}\alpha + \alpha^T 1_n. \tag{2.13}$$

Differentiating Equation (2.13) with respect to α and setting it equal to $\mathbf{0}$, we find:

$$\begin{aligned}\frac{\partial F_D}{\partial \alpha} &= D^T\alpha + 2F^T\alpha - 2G^T\alpha + \mathbf{1_n} = \mathbf{0_{n\times 1}} \\ &\Rightarrow \hat{\alpha} = [2G - D - 2F]^{-1}\,\mathbf{1_n}. \end{aligned} \tag{2.14}$$

Equation (2.14) is solved with the Karush-Kuhn-Tucker condition $\alpha^T y = \mathbf{0}$ (Gordon and Tibshirani, 2012), where matrices G, D and F are symmetric positive definite. Since α is the set of (n) lagrangian multipliers, all its entries are greater than 0. In order to show that matrix $H = [2G^T - D^T - 2F^T]$ is invertible, we need to make sure its expression is positive definite and all its eigenvalues are positive. Thus, the expression of matrices is invertible. The following steps illustrate that idea.

Referring to Equation (2.12), we let:

$$Q = YXE \;\Rightarrow\; Q^T = EX^TY.$$

When E is symmetric and Y is an $n \times n$ diagonal matrix, it implies that:

$$D_{n\times n} = QQ^T, \quad F_{n\times n} = QAQ^T,$$

and

$$\begin{aligned} G_{n\times n} &= \operatorname{diag}\{y_1, y_2, \ldots, y_n\}\, XEE^{-1}EX^T \operatorname{diag}\{y_1, y_2, \ldots, y_n\} \\ &= QE^{-1}Q^T \\ &= Q\left[(I + 2A)^{-1}\right]^{-1}Q^T \\ &= Q\,(I + 2A)\,Q^T \\ &= QIQ^T + 2QAQ^T \\ &= D + 2F. \end{aligned}$$

After substituting the quantities of (G, D, F), we obtain:

$$\begin{aligned} H = 2G - D - 2F &= 2\,(D + 2F) - D - 2F \\ &= 2D + 4F - D - 2F \\ &= D + 2F > 0. \end{aligned}$$

We can see from the expressions of **D**, **F** and **G** that these are symmetric matrices confirming that the inverse of matrix H always exists.

2.4. Assessing performance of the proposed$_{\text{SVM}}$ classifier

The linear classifier of x_i is

$$y_i = \operatorname{sign} f\,(\hat{x_i}) = \operatorname{sign}\left(\hat{\beta_0} + \hat{\beta}^T x_i\right) = \begin{cases} +1 & \text{if } \hat{\beta_0} + \hat{\beta^T} x_i > 0, \\ -1 & \text{if } \hat{\beta_0} + \hat{\beta^T} x_i < 0. \end{cases} \tag{2.15}$$

Two criteria were used to validate the performance of the proposed SVM classifier in comparison with the classical SVM classifier. The following table illustrates these proportions:

Table 1: Confusion matrix to assess performance of SVM classifiers

True/Predicted	class 1	class 2	total
class 1	n_{11}	n_{12}	n_1
class 2	n_{21}	n_{22}	n_2
True/Predicted	class 1	class 2	total
class 1	O_{11}	O_{12}	O_1
class 2	O_{21}	O_{22}	O_2

Table 2: Performance Comparisons of two classification methods based on simulated data

Data set	**n**	**a**	$H_0: \mu_1 = \mu_2$	**MCR**	$\mathbf{MCR_c}$	$\mathbf{H_0: MCR = MCR_c}$	**F1**	**F1**
NO.	sample size	variation	p-value	SVM	SVMC	p-value	SVM	SVMC
Variation effect a								
1.	40	2	0.0000	0.2588	0.2451	0.3830	0.7997	0.7939
2.	40	3	0.0011	0.2898	0.2760	0.4369	0.7532	0.7514
3.	40	6	0.0177	0.3131	0.3198	0.7574	0.6858	0.6844
4.	40	12	0.1098	0.3631	0.3623	0.9066	0.6402	0.6387
Sample size effect n								
5.	40	3	0.0081	0.2502	0.2642	0.1583	0.7705	0.7567
6.	60	3	0.0014	0.2618	0.2700	0.2615	0.7578	0.7492
7.	100	3	0.0000	0.3113	0.3253	0.4811	0.7458	0.7449
8.	200	3	0.0000	0.3775	0.3658	0.2720	0.7521	0.7479
9.	350	3	0.0000	0.4275	0.4258	0.3120	0.7521	0.7499

It can be defined that n_{11}, n_{22} are the correctly classified data values from class 1 and class 2, respectively. Also, n_{12} and n_{21} are the misclassified data values from class 1 and class 2, respectively. The confusion matrix to the right of Table 1 measures the proportion of misclassification for the extreme data (outliers). Generally, p_{11}, p_{22} refer to the proportions of the correctly classified extreme values from class 1 and class 2, respectively, whereas p_{12} and p_{21} are the proportions of misclassified extreme values from class 1 and class 2, respectively. We utilized these measures to validate the performance of our novel modified SVM.

2.5. Algorithm to validate the proposed$_{\text{SVM}}$ classifier

In **Algorithm 1**, we setup the sample sizes for both groups needed for the simulation ($N/2$) with the number of iterations *(iter)* required to calculate misclassification rates (MCR) for each. Then, inside the subroutine, the mean vectors of the two groups were tested in order to relate between the behaviour of how the data are separated as well as the resulting **MCR**. We calculate **E**, a matrix needed to find other matrices **D, F**, and **G**. Because of the high-dimensionality data issue of matrix **H**, in most cases, many recommendations have been made to estimate **H** such as *shrinkage method* (Wang *et al*., 2015). However, in our case, we instead used the Moore-Penerose generalized inverse (Shinozaki *et al*., 1972) to find the solution for the linear system that yields the lagrangian estimator (α).

3. Results

3.1. Simulation study 1

In this simulation study, we sought to validate the effect of sample size and variation in the data on the misclassification rates for the different samples under the proposed method using the quadratic centroid cost function **SVMc**. The samples were generated and tested for their homogeneity in both groups, and were controlled by a positive value (a), such that $\sum_1 = \sum_2 = aI_p$, where **I** is the identity

Data: simulated Data.txt
Result: Calculate MCR for SVM-Cost classifier
initialization N, $iter$, $\tau = 1$;
while $\tau <$ $iter$ **do**
    Simulate N;
    Test $\mu_1 = \mu_2$ & $\sum_1 = \sum_2$;
    Find $\mathbf{E} = (\mathbf{I} + \mathbf{2A})^{-1}$;
    Find matrices **G,D,F**;
    Find **H=2G-D-2F**;
    Solve $H\alpha = 1, \alpha^T y = 0$;
    Find hyperplane β, β_0;
    Calculate MCR;
    $\tau \leftarrow \tau + 1$;
end

Algorithm 1: Simulation studies to validate the proposed$_{\text{SVM}}$ classifier

Table 3: Comparison of MCR based on simulated data

	Misclassification results			
Sample	Original		PC	
n	SVM	SVMC	SVM	SVMC
70	0.3428	0.3285	0.3428	0.3714
80	0.3750	0.4500	0.3750	0.4000
100	0.4600	0.4500	0.4300	0.4500
120	0.4000	0.4250	0.4000	0.4083
200	0.4150	0.4100	0.4150	0.4150
250	0.4000	0.4040	0.4000	0.4120
280	0.4714	0.4571	0.4714	0.4500
370	0.4567	0.4594	0.4567	0.4594
470	0.4127	0.3978	0.4127	0.3936
720	0.4458	0.4486	0.4458	0.4500

matrix of the dimension equal to the number of predictors. In each iteration for generating samples, we changed the parameters of the sample size (n), and the value of (a) to see how these affect the performance of each method in terms of the **MCR**. Furthermore, we evaluated the performance using the $F1$-*score*, which is based on the concept in 1. Some numerical results are summarized in Table 2.

We generated 100 datasets for each pair (n, a). In Table 2, it can be seen that by fixing the sample size and increasing the variation in the data, the MCRs for the two SVM classifiers did not vary significantly. Secondly, although the performance of both methods revealed increases in the MCRs as the sample size increased, there was no significant difference on the predictive classification efficiency of the two methods. Furthermore, the $F1$-*score* confirmed that the performances of the classical SVM and the proposed SVMC were approximately the same, with SVMC performing slightly better, especially for larger sample sizes.

3.2. Simulation study 2

In this simulation study, we sought to validate the proposed classifier, where two multivariate normal samples from populations have significant association between predictors. We generated covariance matrices that have positive eigenvalues $\lambda_j > 0, j = 1, \ldots, p$. We noted that minimizing them affected the cost of classifying data items, which measures the distance from its corresponding centroid. In order to assess the effect within covariances in the objective function, we used the principal components (PCs) method instead of the original variables and compared MCR for both methods of SVM

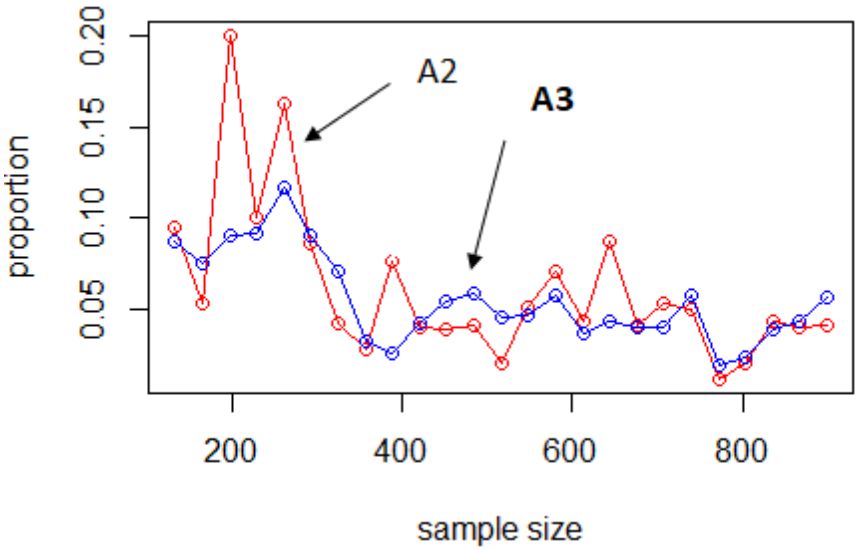

Figure 1: *Effect of changes in the proportion of extreme values with sample size on MCR.*

classifiers. In Table 3, results from the simulated data for the two multivariate groups show lower misclassification rates for both the proposed SVMC and SVM classifiers using the original variables compared to the one where the PCs were used. However, this difference was not significant, and this difference tended to be equal as the sample size n increased.

3.3. Simulation study 3

In this study, we validated the influence of extreme values and sample size on the misclassification rate for the proposed SVM classifier. We used the *minimum covariance determinant* estimator to get the estimates of the Mahalanobis distance. The robust estimators for location and covariance were calculated by using subset J of observations of size h, which minimizes the determinant of the sample covariance matrix, which were only computed from only these h points. Then, these estimators were used to find the robust Mahalanobis distances for each data point (Cabana *et al.*, 2017). Results in Table 3 show that the proportion of extreme values or outliers decreased as the sample size increased. In other words, the effect of outliers is negligible with a large sample size.
Further, the proportion of extreme values that contributed to misclassification decreased when the sample size increased. Figure 1 compares the misclassified extreme value rates of the proposed SVMC and SVM classifiers. Both proportions decreased as the sample size increased. The proportion of misclassified extreme values (outliers) among all misclassified points using the proposed SVMC classifier (A3) is lower than the corresponding proportion under the classical SVM classifier (A2) and tend to be consistent.

3.4. Application on COVID-19 hospitalized patient survival

In this application, we sought to investigate the performance of the proposed SVMC classifier using real-life data on the COVID-19 patient hospitalization outcome. The data, (Khamis *et al.*, 2021) were

Table 4: Comparison of misclassification rates based on COVID-19 survival data

True/Predicted	Original data				Principal components			
	SVM		SVMc		SVM		SVMc	
	Died	Survived	Died	Survived	Died	Survived	Died	Survived
Died	81	88	118	51	169	0	115	54
Survived	18	16	14	20	33	1	6	28
MCR	52.2%		32%		16.2%		29.5%	

collected from The Royal Hospital in Oman. The final data for the application included information on 203 COVID-19 confirmed patients for only complete data. We sought to build a linear classifier for the death outcome due to COVID-19 based on some given features.

Using the COVID-19 data, we constructed a linear classifier based on the two classification methods: The classical SVM method and the proposed SVMC method. Before classifying, we tested for separation between means of groups ($p < 0.001$), that indicated possible separation and homogeneity ($p < 0.001$), to confirm the existence of large differences between covariances in the groups which may have been due to the unbalanced group sample sizes.

Results in Table 4 show that the misclassification rate of SVM (52.2%) was relatively higher than that for the proposed SVMC (32%), which reflects a better and more efficient performance for the proposed SVM classification method. Furthermore, results based on the PCs were as expected since PCs usually work to eliminate the correlation between the original variables. For this reason, MCR of the classical SVM was lower than that for the proposed method since the SVMC is expected to perform better in the presence of more significant correlations among variables.

4. Discussion

The nature of a data structure is a key determinant for how efficient an SVM classifier can be. When data are linearly separable, the problem centers around how to determine the most optimal hyperplane (Al-Shukeili and Wesonga, 2021). However, when data are linearly non-separable, the key problem expands to examining the nature relationships among the data parameters and covariances. Distinct types of algorithms for the SVM deal with two situations for the non-separable data, namely the partially overlapped, but linearly separable and the perfectly mixed data. As a consequence, identifying a linear separable hyperplane is difficult and probably the reason for high misclassification. The kernel trick has been used whereby a specific transformation function is developed to convert the non-linear separable dataset to linearly separable (Izenman, 2008). Despite the differences in data structures, the SVM classifier still employs the same objective function in the quadratic programming regardless of these situations. In this study, we have modified the objective function for the linearly non-separable problem by introducing a centroid mean vector that demonstrated our method's ability to minimize the squared distances and consequently the misclassification rates. Figure 2 demonstrates the principles of our proposed method for the linearly non-separable SVM classification problem.

To validate the proposed modification for the classical SVM objective function, we focused the simulation studies to examine the effects of separation between classes, the large sample size and the extreme or outliers data points on the rate of misclassification. The findings are further discussed below.

4.1. Separation between classes and misclassification

At least a minimum level of separation between classes is required for classification to be efficient. Under the linearly non-separable classification problem, this separability should be compounded with

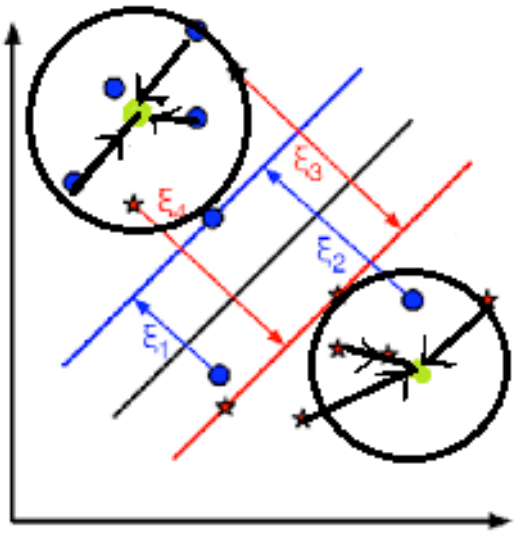

Figure 2: *Modifying SVM objective with a centroid mean vector.*

the centrality of the data for the separate classes. The separation in our solution is measured by examining the squared distances of the multivariate data items from their respective mean vectors. Findings from the simulation studies show that in order to achieve minimum misclassification rates, it is necessary to have meaningful separation between the two classes. This is in agreement with other studies such as the one that explored classification under high dimensional data (Ghaddar and Joe, 2018; Zhang *et al*., 2012; Özcan and Gürgen, 2010; Zhang *et al*., 2022).

4.2. Effect of large sample size on misclassification

In statistics, large sample size has been known to influence both theoretical development and applied statistical analyses. Sufficient sample size is required, without necessarily being larger than the number of instances, so as to avoid non-invertibility of the covariance matrices. Results from our application on the COVID-19 patient data with the sample size of 203 and nine parameters, would not have been successful if there was a high-dimensional challenge. Indeed, results in Table 4 show a higher misclassification rate for the classical SVM (52.2%) than for the proposed SVM (32%). This can be explained by the original data having correlated parameters. However, when data do not have significant correlations among parameters, such as the principal components, the classical SVM produces a lower misclassification rate. A related study that attempted to resolve the SVM classification with significantly correlated parameters proposed a doubly regularized SVM, which treated the L_1 norm together with the standard L_2 norm (Wang *et al*., 2006).

4.3. Effect of extreme or outlier values on misclassification

Outliers or extreme values may influence the rate of classification efficiency. It is also known that the quadratic cost may be influenced by extreme values or outliers. This effect may be minimized by increasing the sample size, though the classification problem is not limited to large samples only. Indeed, findings from our third simulation study show that the proportions of misclassified outliers

among all misclassified points using the proposed SVMC were lower than the corresponding proportions using the classical SVM in terms of their resultant mean and variance. The effect of outliers has been demonstrated in applications that show its effect on the accuracy of the SVM classifier (Debruyne, 2009). Regarding the imbalanced sample size of the binary classification that was discussed in the study (Tang *et al*., 2008; Pérez-Cruz *et al*., 2005), we also found also that our proposed SVMC performs well, even in the presence of the imbalanced data issue.

5. Conclusions

Our proposed SVMC classifier presents a competitive approach to predict the class membership for the binary groups when separation between them is reasonable with significant associations among class predictors. We considered the quadratic centroid cost function, which involve two within the covariance matrices. The validation results from all three simulation studies demonstrate the superiority of the proposed method and, in most cases, its convergence to the classical SVM as the sample size increases. Despite the convexity and parabolic structure of the centroid mean vector or cost function, the misclassification of extreme values or outliers decreases as the sample size increases. Moreover, based on the application with COVID-19 patient hospitalization, the proposed SVMC method is superior to the standard SVM, even when the imbalanced data are used.

6. Acknowledgements

The authors would like to acknowledge the Department of Statistics, Sultan Qaboos University, especially for providing a conducive working research environment. Further appreciation goes to The Royal Hospital, Sultanate of Oman for availing the data we used to validate our novel SVM classifier.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

There is currently no financial support for this research.

Ethical conduct

This work represents the original work of the authors and has not been published anywhere else. Permission was sought for any data used in the method validation and testing.

7. Data availability statements

The data used for the study were mainly simulation data and are available by request from the authors. The data used for the application may be obtained from the original cited paper.

Notes on contributor(s)

- Mubarak Al-Shukeili holds an MSc in statistics and is currently a graduated Ph.D. student. His area of research is to investigate methods that can result in the minimization of classification rates. He is also interested in doing research in medical science, mathematical modelling and computational statistics.

- Ronald Wesonga holds a PhD in statistics. Currently, he is an associate professor in the department of statistics at Sultan Qaboos University. He is also Chair of the Data science Analytics Lab, a lab that promotes international collaborative research. He has published widely in high impact journals, and has inspired many students. Recently, his focus is on challenging problems related to multivariate statistical modelling and error minimization.

References

Al-Shukeili M and Wesonga R (2021). A novel minimization approximation cost classification method to minimize misclassification rate for dichotomous and homogeneous classes, *RMS: Research in Mathematics & Statistics*, **8**, 1–11.

Bühlmann P and Yu B (2003). Boosting with the L_2 loss: Regression and classification, *Journal of the American Statistical Association*, **98**, 324–339.

Brooks JP (2011). Support vector machines with the ramp loss and the hard margin loss, Operations Research, **59**, 467–479.

Cabana Garceran del Vall E, Henry LR, and Lillo Rodríguez RE (2017). Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators, Available from: http://hdl.handle.net/10016/24613

Collobert R, Sinz F, Weston J, and Bottou L (2006). Trading convexity for scalability, *Proceedings of the 23rd International Conference on Machine Learning*, 201–208.

Debruyne M (2009). An outlier map for support vector machine classification, *The Annals of Applied Statistics*, **3**, 1566–1580.

Dormann CF, Elith J, Bacher S *et al.* (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance, *Ecography*, **36**, 27–46.

Ghaddar B and Joe N-S (2018). High dimensional data classification and feature selection using support vector machines, *European Journal of Operational Research*, **265**, 993–1004.

Gordon G and Tibshirani R (2012). Karush-Kuhn-Tucker conditions, *Optimization*, **725**, 10–36.

Han L, Han L, and Zhao H (2013). Orthogonal support vector machine for credit scoring, *Engineering Applications of Artificial Intelligence*, **26**, 848–862.

Izenman AJ (2008). *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*, Springer New York, New York.

Jarray F, Boughorbel S, Mansour M, and Tlig G (2018). A step loss function based SVM classifier for binary classification, *Procedia Computer Science*, **141**, 9–15.

Jiang P, Missoum S, and Chen Z (2014). Optimal SVM parameter selection for non-separable and unbalanced datasets, *Structural and Multidisciplinary Optimization*, **50**, 523–535.

Khamis F, Awaidy SA, Shaaibi MA *et al.* (2021). Epidemiological characteristics of hospitalized patients with moderate versus severe COVID-19 infection: A retrospective cohort single centre study, *Diseases*, **10**, 1–16.

Shuxia L, Xizhao W, Guiqiang Z, and Xu Z (2015). Effective algorithms of the Moore-Penrose inverse matrices for extreme learning machine, *Intelligent Data Analysis*, **19**, 743–760.

Orsenigo C and Vercellis C (2003). Multivariate classification trees based on minimum features discrete support vector machines, *IMA Journal of Management Mathematics*, **14**, 221–234.

Özcan NÖ and Gürgen F (2010). Fuzzy support vector machines for ECG arrhythmia detection, In *Proceedings of 20th IEEE International Conference on Pattern Recognition*, Istanbul, Turkey, 2973–2976.

Pérez-Cruz F, Bousoño-Calzón C, and Artés-Rodríguez A (2005). Convergence of the IRWLS proce-

dure to the support vector machine solution, *Neural Computation*, **17**, 7–18.

Rencher AC (2003). Methods of Multivariate Analysis (2nd ed), Wiley Hoboken, New Jersey.

Shen X, Tseng GC, Zhang X, and Wong WH (2003). On ψ-learning, *Journal of the American Statistical Association*, **98**, 724–734.

Shinozaki N, Masaaki S, and Kunio T (1972). Numerical algorithms for the Moore-Penrose inverse of a matrix: Direct methods, *Annals of the Institute of Statistical Mathematics*, **24**, 193–203.

Siqueira LFS, Morais CLM, Júnior RFA, Araújo AA, and Lima KMG (2018). SVM for FT-MIR prostate cancer classification: An alternative to the traditional methods, *Journal of Chemometrics*, **32**, e3075.

Steinwart I (2001). On the influence of the kernel on the consistency of support vector machines, *Journal of Machine Learning Research*, **2**, 67–93.

Steinwart I (2002). Support vector machines are universally consistent, *Journal of Complexity*, **18**, 768–791.

Tang Y, Zhang Y-Q, Chawla NV, and Krasser S (2008). SVMs modeling for highly imbalanced classification, *IEEE Transactions on Systems, Man, and Cybernetics*, **39**, 281–288.

Vert R, Vert JP, and Schölkopf B (2006). Consistency and convergence rates of one-class SVMs and related algorithms, *Journal of Machine Learning Research*, **7**, 817–854.

Wang C, Pan G, Tong T, and Zhu L (2015). Shrinkage estimation of large dimensional precision matrix using random matrix theory, *Statistica Sinica*, **25**, 993–1008.

Wang H, Shao Y, Zhou S, Zhang C, and Xiu N(2021). Support vector machine classifier via $L_{0/1}$ soft-margin loss, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 7253–7265.

Wang L, Zhu J, and Zou H (2006). The doubly regularized support vector machine, *Statistica Sinica*, **16**, 589–615.

Wu Y and Liu Y (2007). Robust truncated hinge loss support vector machines, *Journal of the American Statistical Association*, **102**, 974–983.

Zhang M, Rubio F, and Palomar DP (2012). Calibration of high-dimensional precision matrices under quadratic loss, In *Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 3365–3368.

Zhang T (2004). Statistical behavior and consistency of classification methods based on convex risk minimization, *The Annals of Statistics*, **32**, 56–85.

Zhang J, Li Y, Zhao N, and Zheng Z (2022). L_0-regularization for high-dimensional regression with corrupted data, *Communications in Statistics-Theory and Methods*, 1–17.

Received June 10, 2022; Revised September 17, 2022; Accepted September 20, 2022