Weighted support vector machine algorithm for efficient classification and prediction of binary response data

A W Banjoko¹, W B Yahya¹, M K Garba¹ & K O Abdulazeez¹

¹ Department of Statistics, University of Ilorin, P.M.B. 1515, Ilorin, Kwara State, Nigeria.

Abstract. This paper proposes a weighted Support Vector Machine (w-SVM) method for efficient class prediction in binary response data sets. The proposed method was obtained by introducing weights which utilizes the point biserial correlation between each of the predictors and the dichotomized response variable into the standard SVM algorithm to maximize the classification accuracy. The optimal value of the proposed w-SVM cost and each of the kernels parameters were determined by grid search in a 10-fold cross validation resampling method. Monte-Carlo Cross Validation method was employed to examine the predictive power of the proposed method by partitioning the data into train and test samples using different sampling splitting ratios. Application of the proposed method on the simulated data sets yielded high prediction accuracy on the test sample. Results from other performance indices further gave credence to the efficiency of the proposed method. The performance of the proposed method was compared with three of the state-of-the-art machine learning methods including the standard SVM and the result showed the superiority of this method over others. Finally, the results generally show that the modified algorithm with Radial Basis Function (RBF) Kernel perform excellently and achieved the best predictive performance than any of the existing classifiers considered.

1. Introduction
Classification of binary response data can be useful for medical personnel if it is computerized for the purpose of fast diagnosis with accurate result. Support Vector Machine (SVM) [1-8] is well known for its classificatory features in binary and multiclass response data. Although, SVM is a very powerful machine learning tool in the field of statistical learning and pattern recognition compare with other machine learning methods. A major drawback of the SVM and other machine learning methods is that its classificatory power depends on the quality of features supply to the algorithm [2-8].

Data quality is a factor that can be an effective tool for machine learning success. If there is duplication and ambiguous data, classification would be difficult. Several modifications of the SVM algorithm has been proposed in literature [4-10], but little has considered the quality of data that are supplied into the algorithm for efficient classification of response with high predictive measures [11]. Effect of data quality on the use of Machine Learning (ML) tools has also received attentions in literature [12]. Since each of the features have different strength of relationship with the response [13, 14], it is therefore necessary to consider this relationship and provide an alternative method that will provide a good quality data to be supplied to the SVM algorithm for better classification/prediction.

This paper presents a new Weighted Support Vector Machine (w-SVM) method that improves on the predictive accuracy of the standard SVM method for two-class response data classification problems.
The predictive performance of this proposed method was compared with three of the state-of-the-art methods including the standard SVM on the simulated data.

2. Methods

2.1. Data Description

2.1.1 Simulated Dataset. Binary response data were simulated following the scheme adopted in [14]. For a typical binary response data with low dimensional structure \( n = 100 \) observations with \( p = 10 \) features (genes) were simulated to yield a \( p = 10 \) data matrix of 100 samples by 10 features. The first 50 samples \( (n_1 = 50) \) represent the first sample group, the tumor group each coded \((-1)\) and the second 50 samples \( (n_2 = 50) \) represent the normal (tumor-free) sample group coded \((+1)\) such that \( n_1 + n_2 = n \).

All the 10 features were simulated from the mixture of two multivariate normal densities having covariance matrix \( \Sigma_1 \) and \( \Sigma_2 \) with respective mean vectors \( \mu_1 \) and \( \mu_2 \). These 10 features have expression levels that are related to the response groups and were thus labeled \( X_1, X_2, \ldots, X_{10} \) that is;

\[
([X_1, X_2, \ldots, X_{10}] | y) \sim \left[ \pi_1 \ast N (\mu_1, \sum_1) + \pi_2 \ast N (\mu_2, \sum_2) \right]
\]

and the mixing parameters \( \pi_1 \) and \( \pi_2 \) was each taken to be 0.5 respectively. Generally, the covariance matrix \( \Sigma \) defined as \( \Sigma = \{ \sigma_{ij} \} \), has a block structure such that

\[
\sigma_{ij} = \begin{cases} 
0.2, & \text{if } |i - j| \leq 5 \\
0, & \text{otherwise}
\end{cases}
\]

2.2 Formulation of the Weighted SVM (w-SVM) Method

The basic idea is to assign different weight to different features in the data such that w-SVM learns the decision surface according to the relative importance of each feature in the training set. The weights used in w-SVM are generated by utilizing the point biserial correlation between each of the features and the response. Also, the different kernels of the SVM are considered for the classification task with different splitting ratios for the training and test data.

Given \( n \) sample points such that each point in \( X \) has \( D \) attributes and is in one of the two classes \( y_i = \pm 1 \).

Let \( X \) be a \( n \times p \) data matrix for the features in the original dataset such that

\[
X = \begin{pmatrix}
X_{11} & X_{12} & \ldots & X_{1p} \\
X_{21} & X_{22} & \ldots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & \ldots & X_{np}
\end{pmatrix}
\]

Let \( \omega_{p \times p} \) be a diagonal matrix representing the weight attached to each feature such that

\[
\omega = \begin{pmatrix}
\omega_{11} & 0 & \ldots & 0 \\
0 & \omega_{22} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \omega_{pp}
\end{pmatrix}
\]
with \( tr(\omega) = 1 \). Each weight \( \omega_{ij} \) in \( \omega \) is computed by

\[
\omega_{ij} = \frac{|r_{xy_i}|}{\sum_{j=1}^{p} |r_{xy_i}|} \quad \text{for } j = 1, 2, \ldots, p, i = 1 \tag{5}
\]

where \( r_{xy_i} \) is the point biserial correlation between the \( i^{th} \) feature \( X_i \) and the binary response \( Y_i \). The \( r_{xy_i} \) is computed for each \( X_i \) as [15, 16]:

\[
r_{xy_i} = \frac{\bar{X}_{+1} - \bar{X}_{-1}}{S_X} \sqrt{\frac{np_{+1}p_{-1}}{n-1}} \tag{6}
\]

where \( \bar{X}_{+1} \) and \( \bar{X}_{-1} \) are the mean values of the continuous (predictor) variable \( X \) for all data points in groups +1 and -1 respectively, \( p_{+1} \) and \( p_{-1} \) are the proportions of data points in groups +1 and -1 respectively with \( p_y = n_y/n \), \( y = -1, +1 \), and

\[
S_X = \left( \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right)^{1/2} \tag{7}
\]

the sample standard deviation of the \( i^{th} \) feature \( X_i \). Let \( Z \) be a new data set derived from the original data matrix \( X \) in (3) and the weight matrix in (4) such that;

\[
Z = X \omega \tag{8}
\]

Hence, the new data matrix \( Z \) of the form

\[
\begin{pmatrix}
   z_{11} & z_{12} & \ldots & z_{1p} \\
   z_{21} & z_{22} & \ldots & z_{2p} \\
   \vdots & \vdots & \ddots & \vdots \\
   z_{n1} & z_{n2} & \ldots & z_{np}
\end{pmatrix}
\]

with \( z_{ij} = X_{ij} \omega_{ij} \) for \( i = 1, 2, \ldots, n \) and \( j = 1, 2, \ldots, p \). Thus, we have \( Z = (Z_1, Z_2, \ldots, Z_p) \). Hence, Given \( n \) sample points such that each input \( z_i \) has \( D \) attributes and is in one of the two classes \( y_i = \pm 1 \). Thus, the training dataset is of the form \( \{z_i, y_i\} \), where \( i = 1, 2, \ldots, n \). Let \( z_i \) be the nearest data point to the hyperplane and \( w \) be the weight vector orthogonal to the hyperplane with \( w'z = 0 \). For any bias \( b \), the equation of the hyperplane can therefore be described by

\[
w'z + b = 0 \tag{10}
\]

to normalize \( w \) with minimum \( z_i \), it is required that

\[
|w'z_i + b| = 1 \tag{11}
\]

The implementation of SVM boils down to selecting \( w \) and \( b \) so that the training data \( \{z_i, y_i\} \) can be describe by the equation:
\[ y_i(z_iw + b) - 1 \geq 0 \] (12)

Hence, the objective of the SVM is to maximize the distance that separate the two sample groups, \( y_i = -1 \) and \( y_i = +1 \). The formulation and optimization of the Quadratic Programming (QP) problem in (12) therefore follows the same pattern as observe in [3] with the objective function

\[
\min_{\alpha} \left( \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j z_i z_j - \sum_{i=1}^{n} \alpha_i \right)
\] (13)

where \( \alpha_i \geq 0 \) is the Lagrange multiplier as explain in [2], with \( w = \sum_{z_i \in SV} \alpha_i y_i z_i \).

The transformed (weighted) data set is then passed to the SVM algorithm for better classification using the Monte-Carlo Cross Validation (MCCV) with 1000 iterations using different splitting ratios for the train \( (n_{tr}) \) and test \( (n_{te}) \) data. Different performance measures such as Prediction Accuracy (ACC), Misclassification Error Rate (MER), sensitivity (SEN), specificity (SPEC), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Jaccard Index (JI) were used to assess the performance of the proposed method.

All analysis was performed using R software (www.cran.r-project.org).

3. Results and Discussion

3.1. Determining the Weight of Each Feature

The result for simulated data set using the proposed method is presented in this session. Table 1 show the results for the absolute correlation coefficient between the dependent variable with each individual independent variables and their respective weight.

Table 1. Table of absolute point biserial correlation coefficient between each of the features and the binary response variable for simulated data set and the respective weight of each of the features.

| \( X_i \) | \( X_1 \) | \( X_2 \) | \( X_3 \) | \( X_4 \) | \( X_5 \) | \( X_6 \) | \( X_7 \) | \( X_8 \) | \( X_9 \) | \( X_{10} \) |
|---|---|---|---|---|---|---|---|---|---|---|
| \( r_{X, Y} \) | 0.2443 | 0.4309 | 0.4587 | 0.464 | 0.3653 | 0.4307 | 0.4168 | 0.4306 | 0.4049 | 0.4623 |
| \( \omega_j \) | 0.0595 | 0.1049 | 0.1116 | 0.1129 | 0.0889 | 0.1048 | 0.1015 | 0.1048 | 0.0986 | 0.1125 |

From Table 1, the quality of the original data sets was modified using the point biserial correlation coefficient and obtaining the weight for each feature in the data set. Much attention is not on the value of the correlation coefficient as it is basically needed for computing the corresponding weight for each of the feature.

3.2. Determining the Kernel Function and Splitting Ratio for Classification with W-SVM

As earlier affirm, the weighted data was splitted into train and test set using different splitting ratios (95:5, 90:10, 80:20, 75:25 and 50:50) in percentages. The train set was used to train the algorithm while the test set was used to validate the result in an out-of-sample scenario. Four different kernels (Linear, Radial Basis Function (RBF), Polynomial and Sigmoid) of the SVM algorithm were also considered for the proposed method and the results are presented in Tables 2 to 4.
Table 2. Showing average misclassified error rate for different kernels with different splitting ratios (in %) using MCCV with 1000 iterations for the proposed method.

| Kernel         | Splitting Ratio | 95:5 | 90:10 | 80:20 | 75:25 | 50:50 |
|----------------|----------------|------|-------|-------|-------|-------|
| Linear         |                | 0.1002 | 0.0942 | 0.0951 | 0.1356 | 0.1138 |
| Polynomial     |                | 0.0818 | 0.0833 | 0.0896 | 0.1279 | 0.1065 |
| RBF            | **0.0786**     | **0.0814** | **0.0915** | **0.1102** | **0.1064** |
| Sigmoid        |                | 0.0789 | 0.0819 | 0.0926 | 0.9142 | 0.1083 |

3.3. Assessing the Performance of w-SVM

As mentioned earlier, some performance indices were used to assess the performance of the proposed w-SVM. The tables below present the summary of the results obtained.

Table 3. Showing the results of the performance indices used in assessing the proposed w-SVM with RBF kernel using MCCV with 1000 iterations.

| Performance Index (%) | Splitting ratios (%) | 95 : 5 | 90 : 10 | 80 : 20 | 75 : 25 | 50 : 50 |
|-----------------------|----------------------|-------|--------|--------|--------|--------|
| ACC                   |                      | 92.14 | 91.86  | 90.85  | 88.92  | 89.36  |
| MER                   |                      | 7.86  | 8.14   | 9.15   | 11.08  | 10.64  |
| SEN                   |                      | 92.81 | 92.50  | 92.08  | 87.66  | 90.54  |
| SPEC                  |                      | 92.12 | 91.58  | 90.18  | 87.22  | 88.87  |
| PPV                   |                      | 92.01 | 91.50  | 90.25  | 87.37  | 89.16  |
| NPV                   |                      | 92.36 | 92.39  | 91.64  | 87.63  | 90.35  |
| JI                    |                      | 85.43 | 80.39  | 83.49  | 77.37  | 81.08  |

w-SVM Algorithm

**Step 1:** Determine the pairwise point Biserial correlation between each predictors and response variable $r_{x_i,y_1}$

**Step 2:** Use $r_{x_i,y_1}$ to determine the weight $\omega_{ij}$ of each predictors

**Step 3:** Multiply the each of the predictors by their corresponding weight above to obtain a new data $Z = X\omega$

**Step 4:** Split the new data into train and test dataset using an appropriate splitting ratio

**Step 5:** The new data train is then pass to the traditional SVM algorithm using the appropriate kernel

**Step 6:** Obtain the performance indices of the w-SVM on the test dataset
3.4. Assessing the Performance of w-SVM.

The proposed w-SVM was compared with three existing classifiers, Naïve Bayes (NB), Random Forest (RF) [17] and the traditional SVM (SVM) [1] using the MER of the classifiers. The results obtained are shown below

**Table 4.** Showing the MER (%) result of each classifier using different splitting ratio on the simulated data.

| Splitting Ratio | Classifiers | NB  | RF  | SVM | w-SVM |
|-----------------|-------------|-----|-----|-----|-------|
| 95 : 5          |             | 14.09 | 18.69 | 10.84 | 7.86 |
| 90 : 10         |             | 13.32 | 18.81 | 10.69 | 8.14 |
| 80 : 20         |             | 14.91 | 16.20 | 11.03 | 9.15 |
| 75 : 25         |             | 16.22 | 17.11 | 12.79 | 11.08 |
| 50 : 50         |             | 15.78 | 16.22 | 11.56 | 10.64 |

**Figure 1.** The graph of MER result of the performance of each classifier presented in Table 4.

From Table 1, the point biserial correlations between the response variable and each of the features were determined respectively. It is observed that though the value of the point biserial correlation differs for each variable in the data, the significance of each value is not important in this study as the values were used in determining the weight for each of the variables which indicates the strength each variable has with the response variable.

Result from Tables 2 to 3 shows the RBF kernel produced the least MER’s among all the SVM kernel functions considered. The reason is supported by [2, 3] when determining the choice of kernel for SVM under different data structure. As the case in this study, the number of features \( (p) \) is far less than the number of subjects \( (n) \), i.e. low dimensional data \( (n \gg p) \) in the simulated data. Also, it can be observed that for all the splitting ratios considered, the same trend of result was achieved, which indicates that splitting ratio of the train and test data have no significant effect on the proposed method. Although, the results generally
supported the literature [14] position to have more data points in the training set for better prediction accuracy.

The assessment of the proposed method in Table 3 using different splitting ratios indicates the efficiency of the proposed method. The least prediction accuracy yielded by the proposed method is 89.36% while the maximum accuracy for the simulated data is 92.14%. The same trend of result was also achieved for all other performance indices considered as presented in Tables 3.

Finally, results in Table 4 shows that the proposed method performs best when compared to three other existing classifiers. This can be viewed by considering the MER of the classifiers considered as presented and selecting the classifier with the minimum MER. Similarly, the MER results of the classifiers are also presented in Figures 1. It can be observed that the proposed method consistently maintains the least MER at all the splitting ratio considered.

4. Conclusion
This study has proposed a weighted Support Vector Machine (w-SVM) for efficient classification and prediction of binary response data. The proposed method uses the point biserial correlation between the response variable and each of the features in determining the weight for each of the features which simply measures the quality of each feature in the data. Each feature is then multiplied by its corresponding weight before being passed to the SVM algorithm for classification using the appropriate kernel function.

From the results obtained in this study, it is observed that the proposed method is best for efficient classification and prediction of binary response data as often the case in machine learning.

The method may also be extended to the case where high dimensional data ($n \ll P$) is inevitable.

References
[1] Vapnik V N 1995 The Nature of Statistical Learning Theory (New York: Springer Verlag)
[2] Banjoko A W, Yahya W B, Garba M K, Olaniran O R, Amusa L B, Gatta N F, Dauda K A and Olorede K O 2017 Proc. Int. Conf. of the Nigeria Statistical Society (NSS) vol 1 p 104-109
[3] Phan A V, Nguyen M L and Bui L T 2016 J. Applied Intelligence 46(2) 1-15
[4] Ji Y, Chen Y, Fu H, et. al. 2017 J. Pattern Recognition 62(Complete) 202-213
[5] Xuehao Yin et al. 2019 J. Phys.: Conf. Ser. 1237 022140
[6] Banjoko A W, Yahya W B, Garba M K, Olaniran O R, Olorede K O and Dauda K A 2015 Annals. Computer Science Series 13 (2) 69-79
[7] Mohamad M S, Omatu S, Deris S, Misman M F and Yoshioka M 2009 Int. J. of Artif. Life and Robotics 13(2) 414-417
[8] Mohamad M S, Safaai D and Rosli MD 2005 International Journal of Computational Intelligence and Applications 5(1) 91-107
[9] Mahata P and Mahata K 2007 Journal of Biomedical Informatics 40(6) 775 – 786
[10] Pelekmans K, Suykens J A K, Gestel T V, Brabanter J D, Hamers B, Moor D and Vandewalle J 2002 LS-SVMlab: a MATLAB/C toolbox for Least Squares Support Vector Machines (Neural Information Processing Systems)
[11] Lee Y W, Strong D M, Kahn B K, and Wang R Y 2002 Inform. Manage 40(2) 133 - 460
[12] Pipino L L, Lee Y W, and Wang R Y 2002 Communications of the ACM 45(4) 211-218
[13] Fan R E, Chang K W, Hsieh C J, Wang X R and Lin C J 2008 J. Mach. Learn. Res. 9 1871–1874
[14] Yahya W B 2012 Gene Selection and Tumour Classification in Cancer Research. A New Approach (Germany: Lambert)
[15] Tate R F 1954 Ann. Math. Statist. 25 603-607
[16] Sheskin D J 2011 Handbook of Parametric and Non-Parametric Statistical Procedures (London: Chapman & Hall/CRC)
[17] Breiman L 2001 Random Forests Machine Learning 45 5 – 32.