Centre for Research and Development (CERAD)
The Federal University of Technology, Akure, Nigeria (www.journal.futa.edu.ng)
JoST 2018

On Some Kernel Based Support Vector Machines

MAKINDE, O.S.* and BODUNWA, O.K.
Department of Statistics, Federal University of Technology, Akure, Nigeria

ABSTRACT: The effectiveness of kernelized support vector machine in classification depends on the choice of kernel function, kernel parameter and soft margin parameter. In practice, there is need for proper guidance on the combination of kernel functions and soft margin parameters to be used. An insight into this is provided in this study. In this paper, we explore the notion of support vector machine and its kernelized version, investigate the performance of some kernel functions and soft margin parameters in support vector classification for some training sample sizes in $\mathbb{R}^d$. We also examine the performance of kernelized support vector machine in functional setting and compare the classifier with maximum functional depth classification methods and centroid classifier based on simulation.

Keywords: Support vection machine, kernel functions, soft margin parameters, classification, error rates

JoST. 2018. 9(2): 21-28.
Accepted for Publication, March 23, 2018

INTRODUCTION

Classification has become one of the most widely used tool to statisticians as well as computer scientists due to a huge amount of online information available these days and it has become paramount to almost all businesses to use these information to target their potential customers (Makinde and Chakraborty, 2015). Several methods have been extensively studied in literature. Bayes rule, linear discriminant analysis and quadratic discriminant analysis are some of the most extensively studied parametric methods in classification. Some nonparametric methods have also been proposed in recent time to solve classification problem. These include distance to mean classifiers (Delaigle and Hall, 2012), nearest neighbour rule (Cover and Hart, 1967, Murty and Devi, 2011), distance to median classification rules (Hall, Titterington and Xue, 2009). Maximum depth classifiers (Ghosh and Chaudhuri, 2005; Li, Cuesta-Albesto and Liu, 2012) is another nonparametric classification method in literature. Data depth measures how central or outlying an observation is to its distribution or data cloud. Maximum depth classifier assigns an observation to the class for which it attains highest depth value.

The foundation of support vector machines (SVM) was developed by Vapnik (1982). SVMs have been successfully applied in solving classification problems in different fields of study. Popularity of SVMs can be attributed to its successful performance in many real applications. Girosi (1998) attributed attractiveness of SVM to the ability to condense information in the training data and provide a sparse representation by using support vectors, a subset of given training data. Cortes and Vapnik (1995) upgraded this method from maximum margin idea to soft margin approach. Park and Liu (2009) proposed use of alternative criterion instead of maximum separation criterion whose solution depends solely on subsets of the training data. Suykens and Vandewalle (1999) proposed least square version of support vector

*Correspondence to: Makinde, O.S.; osmakinde@futa.edu.ng

machine. Li and Yu (2008) proposed functional segment discriminant analysis (FSDA) which combines classical linear discriminant analysis (LDA) as a data reduction tool with support vector machine as classifier. In their proposal, F-statistic is used to select the features on which LDA is applied for data reduction. The first m features with largest statistic values are selected. For sparse functional data (small sample size, large dimention), FSDA uses LDA on short curve segments instead of the whole spectrum.

Kernel trick has been applied in classification for algorithms which solely depends on the inner product of two vectors. This is based on the fact that inner product can be replaced by a kernel function. An example of this is support vector machine. Vapnik (1998), Chapelle et al. (2002), Rossi and Villa (2008) proposed replacing inner product in support vector machine with kernel function. Extension of this approach in infinite-dimensional setting is discussed in literature. See Li and Yu (2008). Amari and Wu (1999) proposed a method of modifying a kernel function to improve the performance of a support vector machine classifier, which is based on the structure of the Riemannian geometry induced by the kernel function. Tong and Koller (2001) introduced a new algorithm for performing active learning with support vector machines. In this study, we explore the notion of kernel based support vector machine and investigate the performance of some kernel functions and soft margin parameters in support vector classification in $\mathbb{R}^d$. We also examine the performance of kernelized SVM in functional data setting. This paper aims to elicit the hidden characteristics and performance of kernel based support vector machine in d-dimensional space and in function space. It answers a big question about which kernel function should be used with any of the soft margin parameters.

SUPPORT VECTOR MACHINE

Suppose $(\boldsymbol{x}, \boldsymbol{y})$ is a pair of random variable in which y, class membership takes values in $\{-1, \quad 1\}$ and $\boldsymbol{x}_i \in S$, where S is a set of training data points in $\mathbb{R}^d$, $i = 1, 2, \ldots, n$. SVM aims at predicting the value of y_i given observed value for $\boldsymbol{x}_i$. SVM separates two different classes of data by a hyperplane

$\{x : < w, x > + b = 0\}$

The corresponding classification rule is

$y_i(x) = sign(< w, x_i > + b)$,

where w is to be estimated and b is a constant scalar. In order to obtain a best separating hyperplane when training data are not linearly separable, $||w||$ is minimised subject to the decision rule for some positive slack variables $\xi_1, \xi_2, \ldots \ldots, \xi_n$ and soft margin parameter C. That is,

$$\min_{\boldsymbol{w}, \boldsymbol{b}, ||\boldsymbol{w}||} \quad \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{n}\xi_i \qquad (1)$$

subject to $y_i(< w, x_i > + b) \geq 1 - \xi_i, \ i = 1, 2, \ldots, n$

The soft margin parameter C, often refer to as penalty parameter of the error term or cost of constraint, controls the trade-off between margin maximisation and error maximisation.

From the geometric perspective, SVM is a large margin classifier. When training data are separable, SVM separates two classes by maximising the margin between them. For non-separable data, the soft-margin SVM chooses a separating hyperplane that splits two classes as cleanly as possible, while still maximising the distance to the support vectors, a subset of the training samples on the separating hyperplane. A desirable property of SVM is that its solution depends only on support vectors. A limitation of SVM is that its decision rule suffers from the presence of redundant variables (Li and Yu, 2008) and extreme outliers (Part and Liu, 2009). Ideas on how to modify SVM to gain robustness are included in Collobert *et al.* (2006) and Wu and Liu (2007).

KERNELIZED SUPPORT VECTOR MACHINE

Kernels are nonlinear mappings of observations in $\mathbb{R}^d$ into feature space. Kernel functions in literature include Gaussian kernel, linear kernel (also called Vanilla kernel), polynomial kernel, hyperbolic tangent kernel (also called sigmoid kernel), exponential kernel, ANOVA kernel, Bessel kernel, Cauchy kernel, Chi square kernel, wavelet kernel, string kernel and Laplace kernel among others. The essence of this is to transform candidate linear algorithms into a non-linear. Those non-linear algorithms are equivalent to their linear originals operating in the range space of a feature space.

Replacing w in (1) by $\sum_i \alpha_i y_i \Phi(x_i)$ subject to $\alpha_i \geq 0$ and $\sum_i y_i \alpha_i = 0$ results in solving a dual problem. There is no need of computing feature function $\Phi(x)$, a kernel function $K(x_i, x_j)$ can be chosen to represent $\Phi(x_i)^T \Phi(x_j)$ in some high dimensional space. The advantage of this is that the resulting classifier is not affected by the presence of noisy observation. The dual formulation of the soft margin problem is defined as

$$\min_{w,b,||w||} \sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i=1}^{n} y_i \alpha_i y_j \alpha_j K(x_i, x_j)$$

subject to $0 \leq \alpha_i \leq C, \ i = 1, 2, \ldots, n$

$$\sum_i y_i \alpha_i = 0$$

Kernelized support vector machine (KSVM) maps training vectors into a feature space using a kernel function that defines an inner product in the feature space. This provides consistent classification in both finite and infinite dimensional spaces (Rossi and Villa, 2008). When no prior information is available about each of the attributes, Chapelle *et al*. (2002) suggested use of spherical kernel, which assigns equal weight to each attribute. Gaussian kernel, most commonly referred to as radial basis function, may be a better choice when attributes have different scales of measurement. Furey *et al*. (2000) applied kernel type in analysing gene expression data.

Accounting for different soft margin parameters

Kernelized support vector machines can use different soft-margin parameters (Cortes and Vapnik, 1995). The soft margin parameters (or costs of constraint) are of three types, which are C support vector classification(C-svc), η support vector classification(η-svc) and bounded-constraint support vector classification(C-bsvc). The parameter sets the upper bound on the training error and the lower bound on the fraction of data points to become support vectors. To the best of our knowledge, there is no literature on kernelized support vector machine where polynomial, linear, hyperbolic tangent, Bessel, ANOVA, spline kernel functions are employed. In the next section, we shall consider the implications of different soft margin parameters on the performance of support vector machine.

NUMERICAL EXAMPLE

As illustration of apparent error rates of the kernelized support vector machine accounting for effect of kernel functions and costs of constraint on the performance of the classifier, we present a simulation study. Let populations π_1 and π_2 be bivariate spherically symmetric with centre of symmetries μ_1 and μ_2, and covariance matrix, Σ_1 and Σ_2, respectively. Assume that the prior probabilities of π_1 and π_2 are equal. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from π_1 and $Y_1, Y_2, \ldots, Y_m$, a random sample from π_2. We simulate a new random sample $Z_1, Z_2, \ldots, Z_m$ from π_1 and

$Z_{m+1}, Z_{m+2}, \ldots, Z_{2m}$ from π_2 and take sample sizes n and m to be 100. The simulation is repeated 1000 times. The competing distributions considered are bivariate normal distributions $\left(N(\mu_1, I), N(\mu_2, I)\right)$, bivariate Laplace distribution $\left(BL(\mu_1, I), BL(\mu_2, I)\right)$, and bivariate t distribution with 3 degrees of freedom $\left(t_3(\mu_1, I), t_3(\mu_2, I)\right)$, where $\mu_1 = \begin{pmatrix}0\\0\end{pmatrix}$ and $\mu_2 = \begin{pmatrix}\delta\\0\end{pmatrix}$ I is identity matrix. For computation, we use R-Package *kernlab* and employ 5-fold cross-validation. The costs of constraint C and η are taken to be 1 and 0.2 respectively. For Gaussian kernel function and Laplace kernel function, the hyperparameter σ is determined automatically by the sigest function in the same library and it returns a value between the 0.1 and 0.9 quantile of $||x_i - x_j||$.

Figure 1 shows the error rate of three bivariate distributions, namely normal, Laplace and t with 3 degrees of freedom, as δ varies in $[-2, 2]$. As expected, error rate is nearly 0.5 when $\delta = 0$ and it decreases as δ goes away from 0 and the separation between the population increases. We observe in Tables 1 and 2 that there is no significant difference between the mean performance of C-svc and C-bsvc for each of the kernel functions used under normal and non-normal settings. For normally distributed samples, mean misclassification error is least when combining either of liner kernel or polynomial kernel with C-svc. Also, an equivalent performance of KSVM is observed when polynomial kernel is combined with C-bsvc as shown in Table 1. Consider the non-normal setting in Table 2, we observe that the optimal performance of KSVM is obtained when linear kernel is combined with C-svc irrespective of the value of δ. In Tables 1 and 2, we observe that c-bsvc does not work with linear kernel and so misclassification errors could not be computed for KSVM with linear kernels irrespective of the distributions of the competing populations. The performance of η-svc is poor compared to C-svc and C-bsvc.

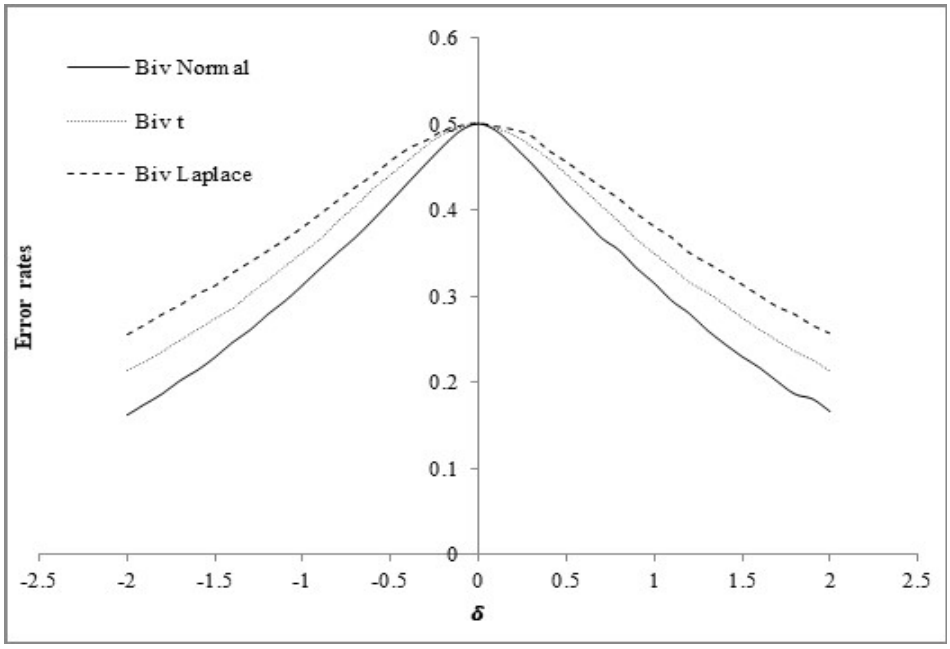

Figure 1: Comparison of error rates of KSVM for different distributions using Gaussian radial basis function

Table 1: Effect of kernel function type and cost function on the performance of KSVM for bivariate normal distributed samples

Kernel function	$\delta = 1$			$\delta = 2$		
	C-svc	η-svc	C-bsvc	C-svc	η-svc	C-bsvc
Gaussian	0.3278	0.4265	0.3245	0.1696	0.2508	0.1686
Polynomial	0.3112	0.4575	0.3127	0.1611	0.3546	0.161
Linear	0.3115	0.4553	-	0.1603	0.362	-
Hyperbolic tangent	0.4249	0.5506	0.4254	0.2823	0.6551	0.2837
Laplace	0.326	0.3862	0.3287	0.1691	0.2071	0.169
Bessel	0.3124	0.4937	0.3134	0.1618	0.4179	0.1614
ANOVA	0.3135	0.5002	0.3149	0.1606	0.3568	0.1615
Spline	0.4216	0.5393	0.4475	0.2859	0.6707	0.3122

Table 2: Effect of kernel function type and cost function on the performance of the error rates associated with KSVM for bivariate Laplace distributed samples

Kernel function	$\delta = 1$			$\delta = 2$		
	C-svc	η-svc	C-bsvc	C-svc	η-svc	C-bsvc
Gaussian	0.3809	0.4675	0.3819	0.2577	0.3726	0.2583
Polynomial	0.3857	0.4972	0.3606	0.2593	0.4597	0.2414
Linear	0.3584	0.4991	-	0.2423	0.4653	-
Hyperbolic tangent	0.4620	0.5199	0.4599	0.3740	0.5784	0.3711
Laplace	0.3826	0.4356	0.3818	0.2563	0.3149	0.2564
Bessel	0.3664	0.5094	0.3670	0.2430	0.4644	0.2441
ANOVA	0.3609	0.5010	0.3635	0.2429	0.4566	0.2416
Spline	0.4389	0.5097	0.4437	0.3532	0.5608	0.3772

Table 3: Effect of training sample sizes on the performance of KSVM for bivariate Laplace distributed samples with $\delta = 2$ when c-svc is used

Kernel function	Sample sizes			
	20	50	100	200
Gaussian	0.288975	0.27018	0.257735	0.250353
Polynomial	0.2593	0.24324	0.2593	0.240098
Linear	0.254725	0.24606	0.242295	0.241553
Laplace	0.284225	0.26702	0.25633	0.250513
Bessel	0.2587	0.25037	0.24296	0.242293
ANOVA	0.266125	0.24928	0.24289	-

Table 4: Effect of training sample sizes on the performance of KSVM for bivariate Laplace distributed samples with $\delta = 2$ when c-bsvc is used

Kernel function	Sample sizes			
	20	50	100	200
Gaussian	0.2927	0.26733	0.258335	0.251005
Polynomial	0.2610	0.2457	0.241405	0.240073
Laplace	0.2828	0.2663	0.256365	0.251643
Bessel	0.2597	0.2486	0.244115	0.243658
ANOVA	0.2639	0.2466	0.24161	-

In Tables 3 and 4, implementation of KSVM for classification using ANOVA kernel is almost practically impossible with large training sample size irrespective of the cost function used and the distributions of competing populations. We also observe from the tables that the larger the training sample size, the less the mean error rate. Misclassification error cannot be estimated for KSVM using ANOVA kernel for large sample size irrespective of competing distributions.

CLASSIFICATION IN FUNCTION SPACE

In function space, kernel based support vector machine can also be applied as discussed in Rossi and Villa (2006, 2008). To examine the performance of kernel based support vector machine for functional data, we compare its performance with depth based procedures and centroid based classification method of Hastie, Tibshirani and Friedman (2001). Centroid classifier assigns an unclassified observation to the class for which the unclassified observation attains least L_2 distance from the class centroid. Consider the functional model in Cuevas, Febrero and Fraiman (2007) and Makinde (2016). The population P_0 consists of trajectories of the process $X(t) = m_0(t) + e(t)$, where $m_0(t) = 30(1-t)t^{1.2}$ and $e(t)$ is a Gaussian process with mean 0 and $cov(X(s), X(t)) = 0.2\,exp(-|s-t|/0.3)$. The process corresponding to P_1 differs from $X(t)$ only in the mean function and is given by, $Y(t) = m_1(t) + e(t)$ with $m_1(t) = 30(1-t)^{1.2}t$. The experiment is repeated 1000 times, mean and standard error of the error rates are computed.

The functions are handled in a discretized version based on 500 equispaced grid points on $[0, 1]$. For maximum functional depth classifier, three functional depths are considered. The depth functions are h-mode depth (HMD), Fraiman-Munic depth (FMD) and random projection depth (RPD). To compute these functional depth functions, we use R package fda.usc with 10% trimming, and assign observations to class with maximum depth value. We choose the sizes of both training samples and validation samples of P_0 and P_1 to be 100 and repeat the simulation for 1000 times. The parameter h in the h-mode depth is chosen as the 15 percentile in the L_2 distances between the functions in the training sample. We choose the number of projections to be 100 for h-mode depth while 50 projections for random projection depth. Here Gaussian kernel is chosen for KSVM with soft margin parameter (C-svc).

Table 5 presents the performance of KSVM, maximal depth classifiers and centroid classifier in term of average misclassification error for the simulation procedures above. We observe that KSVM performs noticeably better than functional depth based classifiers and centroid classifier in this example.

Table 5: Comparison of the performance of KSVM with maximal depth classifiers for functional data

Statistics	Maximum depth classifiers			Centroid Classifier	KSVM
	FM Depth	h-mode Depth	RP Depth		
Mean	0.0712	0.0631	0.0528	0.0240	0.0138
Standard error	0.0183	0.0343	0.0164	0.0107	0.0083

CONCLUSION

In this paper, we have considered some factors affecting the performance of kernelized support vector machine for classification in $\mathbb{R}^d$. We have noted in our examples that the performance of support vector machine in classification does not depend heavily on the training sample sizes of the competing classes. Use of anova kernel with C-svc or C-bsvc performs well for relatively large training sample size. It becomes practically impossible to use when training sample size is larger than 100. The mean error rates of KSVM associated with C-svc and C-bsvc are equivalent and consequentially, KSVM performs well with either of C-svc and C-bsvc except for normally distributed classes when spline kernel functions are used. The generalization of KSVM for functional data is straightforward, so we examine the performance of this extension in function space and compare the performance of KSVM with performance of maximum functional data depth classification rule and centroid classifier based on simulation study.

REFERENCES

Amari S. and Wu S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, **12**, 783 - 789.

Chapelle O, Vapnik V, Bousquet O. and Mukherjee S. (2012). Choosing Multiple Parameters for Support Vector Machines. Machine Learning 46, 131 - 159.

Collobert R., Sinz F., Weston J., and Bottou L. (2006). Large scale transductive svms. Journal of Machine Learning Research 7 1687–1712.

Cortes C. and Vapnik V. N. (1995). Support-Vector Networks. *Machine Learning*. **20**(3), 273 - 297.

Cover T. M. and Hart P. (1967). Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory. **13**(1), 21 - 27.

Cuevas A., Febrero M. and Fraiman R.(2007). Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics. **22**(3), 481 - 496.

Delaigle A. and Hall P. (2012.) Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B*. **74**(2), 267 - 286.

Furey T.S., Cristianini N., Duffy N., Bednarski D.W., Schummer M. and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue sample using microarray expression data. *Bioinformatics*, **16**(10), 906-914.

Ghosh A. K. and Chaudhuri P. (2005). On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, **32**, 327 – 350.

Girosi F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Computation*, **20**, 1455 - 1480.

Hall P., Titterington D.M. and Xue J.(2009). Median Based classifiers for High Dimensional Data. *Journal of the American Statistical Association.* **104**(488), 1597 - 1608.

Hastie T., Tibshirani R. and Friedman J. (2001). The elements of statistical learning. Springer, NY.

Li B. and Yu Q. (2008). Classification of functional data: A segmentation approach. *Computational Statistics and Data Analysis*, **52**, pp. 4790 - 4800.

Li J., Cuesta-Albertos J. A. and Liu R. Y. (2012). DD-Classifier: Nonparametric Classification Procedure Based on DD-plot. *Journal of the American Statistical Association*, **107**, 737-753.

Makinde O.S. (2016). Some classification rules based on distribution functions of functional depth. Statistical Papers. DOI: 10.1007/s00362-016-0841-0

Murty M.N. and Devi V.S. (2011). Nearest Neighbour Based Classifiers. In: Pattern Recognition. Undergraduate Topics in Computer Science, Springer, London

Park S.Y. and Liu Y. (2009). From the support vector machine to the bounded constraint machine. *Statistics and its Interface*, **2**, 285-298.

Rossi F. and Villa N. (2006). Support vector machine for functional data classification. *Neurocomputing*, **69**, 730 - 742.

Rossi F. and Villa N. (2008). Recent advances in the use of SVM for functional data classification. First International Workshop on Functional and Operatorial Statistics, Toulouse.

Suykens J.A.K. and Vandewalle J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, **9**(3), 293-300.

Tong S. and Koller D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 45-66.

Vapnik V. N. (1982). Estimation of dependences based on empirical data. Addendum 1, New York: Springer-Verlag.

Vapnik V.N. (1998). Statistical Learning Theory. John Wiley and Sons, New York.

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