Multiclass Support Matrix Machines by Maximizing the Inter-class Margin for Single Trial EEG Classification

Imran Razzak, Michael Blumenstein, and Guandong Xu,

***Abstract*—Accurate classification of Electroencephalogram (EEG) signals, plays an important role in diagnoses of different type of mental activities. One of the most important challenges, associated with classification of EEG signals is how to design an efficient classifier consisting of strong generalization capability. Aiming to improve the classification performance, in this paper, we propose a novel multiclass Support Matrix Machine (M-SMM) from the perspective of maximizing the inter-class margins. The objective function is a combination of binary hinge loss that works on C matrices and spectral elastic net penalty as regularization term. This regularization term is a combination of Frobenius and nuclear norm, which promotes structural sparsity and shares similar sparsity patterns across multiple predictors. It also maximizes the inter-class margins that helps deal with complex high dimensional noisy data. The extensive experiment results supported by theoretical analysis and statistical tests show the effectiveness of the M-SMM for solving the problem of classifying EEG signals associated with motor imagery in Brain-computer Interface (BCI) applications.**

***Index Terms*—SVM, Matrix classification, multiclass support matrix machines, SMM, M-SMM, nuclear norm.**

I. Introduction

Brain-Computer Interface (BCI) is an advanced approach to establish the direct communication between a human brain and machine [1], [2], [3]. Electroencephalogram (EEG) has been used by clinicians as a standard neuroimaging tool to study the neuronal dynamics within the human brain. EEG data reflect the process of an individual's information processing [4] and is widely used for the diagnosis of neurological disorders such as epilepsy, sleep apnoea [5]. EEG signals offer an inexpensive and relatively easy way of reading the brain activity as compared to other neuroimaging methods such as Magnetoencephalography (MEG), Functional Magnetic Resonance Imaging (fMRI) and Electrocorticography (ECoG). Setting up an experiment with EEG can also be done without too much hassle, it is sometimes easy as placing a headset on and checking the data quality. These reasons make the EEG a popular and an important resource for identification and classification the brain activity.

Recent technological advancement in EEG data acquisition by using dense groups of electrodes attached to the cranium have increased the scope of EEG recording abilities. Visual inspection of such massive data sets is cumbersome and prone to error. In the field of machine learning and statistics, classification is an important research task, aiming to identify to which set of categories a new observation belongs to. This classification of unseen data is made from information learnt from a training set containing known observations. To classify EEG data efficiently, not only optimized set of features but also classifier optimization is essential. Approaches that perform little optimization of features or classifier are guaranteed to provide badly over-fitted and useless models when applied to complex data, posing great difficulties for further data analysis tasks. In addition to this, selection of discriminant features, maximizing the hyper-plane and training point margins is the key to success of support vector machines. This helps to select patterns important for classification and also overcome the computational complexity.

Generally, the high-dimensional EEG data poses several serious challenges, especially due to its limited size. One of the most important issues for the classification of EEG signals is how to design a powerful classifier with strong generalization capability?. For certain classification tasks, the data has to be reshaped into vectors for dimensionality reduction and classification which could destroy the structural information embedded within. An ad hoc solution for this issue is to concatenate the matrix into vectors, however, this results in an increase in dimensionality leading to model over-fitting. For example BCI IIIa dataset has small sample size (288 samples with 750×22 temporal spatial matrix) that results in high dimensional features (750×22=16,500). Nevertheless, representation of such data in the form of matrices can preserve its structural information i.e. EEG signal which consists of voltage fluctuations at several electrodes during a time period, has a strong correlation between certain frequency band and channels. Furthermore, reshaping of high dimensional data to vectors could result in increase of dimensionality [6] [7].

Recently, several efforts have been made to classify directly from matrix without converting it into vectors, which exploits the correlation between the columns or rows of matrix. Although, these methods takes full advantage of low rank assumption to exploit the strong correlation between columns and rows of each matrix and able to extract useful features, however, all these matrix classifiers expect MSMM are originally built for binary classication problems. Although, they could be used for multiclass classification by breaking multiclass problem into series of binary class classification problem such as one-vs-rest (OvR) or one-vs-one (OvO) strategies (e.g. In OvsR, the mutli-class problem is solved by splitting it into n binary class classification problems,

Razzak and Xu are with the AAI, UTS, Sydney, Austraia, (Corresponding Author: Guandong Xu) email:imran.razzak@ieee.org, michael.blumenstien@uts.edu.au, guandong.xu@uts.edu.au,

whereas OvsO approach splits the problem into $\frac{c(c-1)}{2}$ binary classification problems.) but are computationally expensive and may results in unbalanced distribution of input samples. Thus, there is need to train multiple binary classifier or multiclass classifier into single optimization.

Aiming to address the multiclass classification for matrix data, in this paper, we present a novel multiclass Support Matrix Machine (M-SMM) approach by utilizing the maximization of the inter-class margins (i.e. margins between pairs of classes). The proposed model is a combination of binary hinge loss for models fitting, and elastic net penalty as a regularization on regression matrix. The binary hinge loss uses C matrices to simulate one-vs-one classifier of all classes rather than $\frac{c(c-1)}{2}$ models. The regularization term which promotes the structural sparsity and shares similar sparsity patterns across multiple predictors, is a combination of Frobenius and nuclear norm. Thus, the proposed objective function not only maximizes the inter-class margins but is a spectral extension of conventional elastic net that combines the property of low rank and joint sparsity together, to deal with complex high dimensional noisy data. We can describe the theoretical and empirical **key contributions** of this work as follows: A novel classifier M-SMM which works by effectively combining the binary hinge loss function (to maximize the inter-class hyper plane margin for model fitting) and elastic net penalty (to promote low-rank plus sparsity), as a regularization on regression matrix. Unlike one vs one classification strategy, we have used C matrices to simulate the binary classification that not only helps to overcome the complexity issue but also maximizes the inter class margin. Since the optimization is convex and one of the major challenges is how to efficiently solve non smooth optimization?, thus, we devised an efficient algorithm for solving the proposed objective functions.

II. Recent Advancement on EEG Classification

EEG measures the electrical activity of the brain via electrodes placed on the scalp. It provide the information from the surface measurements, how active the brain is and can be used to measure the abnormal activity, such as with epilepsy. EEG classification is challenging task due complex nature of the data such as the low signal-to-noise ratio, the non-stationary of signals and the presence of noises etc. An ad hoc approach for EEG classification is to measure the different between left and right electrodes, however, it provides poor accuracy. Several research efforts has been performed to improve the EEG classification accuracy. To classify EEG data efficiently, not only optimized feature extraction of relevant EEG data but also optimization of classifier is essential to improve the quality of cognitive performance evaluations. With respect to the data, EEG classification approaches are divided into two categories such as vector based methods (requires data to be in the form of vectors) and matrix based methods (works on matrix data without reshaping the data into vectors).

Vector based methods such as linear discriminant analysis (LDA) [8], [9], [10], support vector machines (SVM) [11], [12], [13], K nearest neighbor (KNN) [14], [15] and neural network [16], [17], [18], [16], [19] have been successively applied for EEG classification. In most cases, the data has to be reshaped into vectors for further classification which could in-turn destroy the structural information embedded in it. Recently, some efforts have been made to suppress the matrix into vectors using common spatial patterns [20], [21], [22], [23], [24], [25]. However, these methods ignore the topological structure embedded in the matrix data, whereas considering structural information is of great interest and helps to improve the classification.

To utilize the information properly that were lost due to reshaping of matrix into vector, recently, matrix based classifier are extensively being used for classification of EEG task [3], [26], [11], [27]. Wolf et al. used rank- k SVM by regularizing the regression matrix as the sum of k rank-one orthogonal matrices to capture the global structure of matrix data [28]. Pirsiavash et.al. [29] and Dyrholm et al. [30] presented a bi-linear classifier by applying the hinge loss for model, fitting through factorization of regression matrix into low rank matrix. Zhang et. al. devised low rank linearization to transform the non-linear SVM to linear one through kernel map computed from low rank approximation of matrices [31]. However, these methods are not able to exploit the correlation between rows and columns within each single trial EEG data.

For classification of high dimensional data, not only dimensional reduction or feature selection, it is also important to find salient features that belong to specific part of image as projection procedure involves all the original features and it may have redundant or irrelevant features. To select such salient patterns, projection matrix should consist of sparse element with respect to such features. Recently, Luo et al present support matrix machines by using hinge loss, nuclear norm and Frobenius norm [27]. SMM is not only able to select features but also leverage the structural information by learning the low rank regularization from the noisy EEG features. Zheng et al. extended this work and presented sparse support matrix machines (SSMM) by combining the ℓ_1-norm, nuclear norm and hinge loss [3]. SSMM is robust against noisy data due to ℓ_1 robustness against outliers. The above methods take full advantage of the low-rank assumptions to exploit the correlation between rows and columns within each single trial EEG data and able to extract important features by discarding irrelevant and redundant features, however, works only for binary classication problems.

Recently, Zheng et. al. extended the problem into multiclass EEG classification problem and presented a mulitcalss support matrix machines that is a combination of hinge loss, nuclear and Frobenius norm [26]. However, it is computationally complex and is sensitive to noisy data. To overcome aforementioned issues in earlier works, in this paper, we presented a novel multiclass EEG classification approach that maximizes the inter-class margins to deal with complex high dimensional noisy data.

III. Notations and Preliminaries

We started by establishing the notation and preliminaries used throughout this paper. Scalar, vector and matrix are represented by lowercase letter (e.g. x), lowercase bold letter

(e.g. $\mathbf{x}$) and uppercase letter (e.g. X) respectively. We let I_p denoted by $p \times p$ matrix. For a matrix $X \in \mathbb{R}^{p\times q}$, its singular value decomposition is denoted as $X = U\Sigma V^T$, where U is the unitary matrix, $\Sigma = diag(\sigma_1, \sigma_2 ... \sigma_r)$ is the rectangular diagonal matrix and V^T is the conjugate transpose of the unitary matrix. The Frobenius norm of a matrix X is denoted as $||X|| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q} x_{pq}^2}$. The nuclear norm of a matrix X is $||X||_* = \sum_{i=1}^{r} \sigma_i$.

As we know, the nuclear norm $||X||_* = \sum_{i=1}^{r} \sigma_i$ of a matrix X as a function from $\mathbb{R}^{p,q}$ to $\mathbb{R}$ can not differentiated. Alternatively, we have to consider the sub-differential of X_* that is denoted by $\partial||A||_*$. It is a set of sub-gradients. For a matrix X of dimension $p \times q$ of rank r,

$$\partial||A||_* = \big\{U_X V_X^T + Z : Z \in \mathbb{R}^{p\times q},$$

$$U_X^T Z = 0, ZV_X = 0, ||Z||_2 \leq 1 \quad (1)$$

For any $\tau \geq 0$, the singular value thresholding operator (SVT) is defined as

$$\mathbb{D}_\tau(X) = U\Sigma_\tau V^T$$

where $\Sigma_\tau = diag([\sigma_i(X) - \tau)]_+, ...[\sigma_r(X) - \tau)]_+)$

IV. The Problem Formulation

In this section, we provide the motivation, brief description, and formalization of matrix classification problem. Practically, it has been noticed that the selection of features and suitable model design is far more important than the choice of classifier itself [32]. Hence, in this paper, we are focusing on selection of useful features and making the classifier robust.

We are given a set of training samples $T = \{X_i, y_i\}_{i=1}^n$, where $X_i \in \mathbb{R}^{p\times q}$ is the i^{th} input sample matrix and $y_i \in \{1, -1\}$ is its corresponding class label. Generally, the data needs to be transformed/stacked into vectors in order to fit a classifier. Let $x_i = vec(X_i^T) = ([X_i]_{11}, [X_i]_{12}, ...[X_i]_{1q}, [X_i]_{21}, [X_i]_{22}, ...[X_i]_{pq})^T \in \mathbb{R}^{pq}$. We have n number of training samples and c number of classes. Thus, we are required to build c number of binary SVM classifiers.

The classical multiclass soft margin SVM is defined as

$$arg \min_{w_j, b_j} \frac{1}{2} tr(w_j^T w_j) + C\sum_{i=1}^{n} \xi_i^j \quad (2)$$

such that

$$w_j^T x_i + b \geq 1 - \xi_i^j, \ if \ y_i = j$$
$$w_j^T x_i + b \leq -1 + \xi_i^j, \ if \ y_i \neq j$$
$$\xi_i^j \geq 0$$

Where $\xi_i^j = 1 - y_i[tr(W^T X_i) + b]_+$ is the hinge loss, $W \in \mathbb{R}^{pq}$ is the vector of regression coefficients, $b \in \mathbb{R}^{pq}$ is an offset term and C is a regularization parameter. This problem is considered unbalanced even though the number of training samples in class are balanced due to one-vs-all strategy. This property affects the classification performance, and was resolved through one-vs-one classification strategy. To classify unseen data, voting strategy is used and the class with maximum votes is considered as output. In result, it is required to build $\frac{c(c-1)}{2}$ number of classification models in total and can be defined as follow

$$arg \min_{w_{jk}, b_{jk}} \frac{1}{2} tr(w_{jk}^T w_{jk}) + C\sum_{i=1}^{n} \xi_i^{jk} \quad (3)$$

such that

$$w_{jk}^T x_i + b_{jk} \geq 1 - \xi_i^{jk}, \ if \ y_i = j$$
$$w_{jk}^T x_i + bjk \leq -1 + \xi_i^{jk}, \ if \ y_i \neq k$$
$$\xi_i^{jk} \geq 0$$

Later on, Guermeur formulated a theoretical SVM framework for multiclass classification [33] which can be written as

$$arg \min_{w^{d\times c}, b^c} \frac{1}{2}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c} ||w_j - w_k||_2^2 + \sum_{j=1}^{c} ||w||_2^2 + C\sum_{i=1}^{n}\sum_{j\neq y_i} \xi_i^{jk} \quad (4)$$

such that

$$w_{y_i}^T x_i + b_{y_i} \geq w_j^T x_i + bj + 1 - \xi_i j$$
$$\xi_i j \geq 0, \ \forall i \in 1, ...c_i$$

Xu et.al. extended the above Eq. 4 to multiclass binary SVM and proposed c vectors to simulate one-vs-one binary classifiers [34]

$$arg \min_{w^{d\times c}, b^c} \frac{1}{2}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c} ||w_j - w_k||_2^2 + \sum_{j=1}^{c} ||w||_2^2 + \frac{1}{2}\sum_{j=1}^{c} b_j^2 + C\sum_{j=1}^{c}\sum_{k=j+1}^{c}\sum_{y_i \in j,k} \xi_i^{jk} \quad (5)$$

such that

$$y_i^{jk} f_{jk}(x_i) \geq 1 - \xi_i^{jk}, \ \forall y_i \in j, k$$
$$\xi_i^{jk} \geq 0$$

Here, it is required to reshape the date into vectors which results in losing losing the correlation among columns or rows in the matrix. To be benefited from rich structural information hidden in the data, recently support matrix machine has been proposed. By directly transforming the Eq. 2 for matrix, we get

$$argmin \frac{1}{2} tr(W^T W) + C\sum 1 - y_i[tr(W^T X_i) + b]_+ \quad (6)$$

It is an established fact that $tr(WW^T) = vec(W)vec(W^T)$ and $tr(W^T X_i) = vec(W)^T vec(X_i)$, thus the above objective function can not capture the intrinsic structure of each input matrix efficiently due to the loss of structural information during the process of matrix reshaping into vectors. To take the advantage of intrinsic structural information within each matrix, one intuitive way is to capture the correlation within each matrix through low rank constraints on the regression parameters. Results showed that exploiting the correlation

information improved the classification performance. The Eq. 2 can be rewritten for matrix classification as

$$argmin\frac{1}{2}tr(W^TW) + C\sum 1 - y_i[tr(W^TX_i) + b]_+ \quad (7)$$

Motivated by this, Luo et. al. presented sparse matrix machine shown in Eq. 8 [27]. The objective function in Eq. 8 consists of hinge loss plus nuclear norm and Frobenius norm as reguarlizer.

$$argmin\frac{1}{2}tr(W^TW)+\tau||W||_*+C\sum 1 - y_i[tr(W^TX_i) + b]_+ \quad (8)$$

The spectral elastic net regularization $\frac{1}{2}tr(W^TW)+\tau||W||_*$ captures the correlation with in each matrix individually. Furthermore, the nuclear norm in reguarlizer is used to control the rank of W (NP-hard problem) and provides the best approximation of rank of matrix W. The objective function shown in Eq. 8 is capable of capturing the latent structure within each matrix and perform the classification based on all entities of each matrix, which effects the classification performance and makes the model complicated. Although, these methods for matrix data take full advantage of low rank assumption to exploit the strong correlation between columns and rows with in each matrix and are able to extract useful features, however, there is still need to train multiple binary classifier or multiclass classifier into single optimization.

V. Maximizing Inter-Class Margins for SMM

In this section, we introduce the proposed approach for maximizing the inter-class margin for support matrix machine. Figure 2 illustrate the proposed mulitclass support matrix machines for classification of EEG signals. It is in principal novel classifier being able to, maximize the inter-class margins, select the discriminant patterns by removing the redundant information, and to consider the strong correlation of rows and columns in the matrix. Figure 1 shows the motivation of M-SMM. The objective function in Eq. 9 is combination of sparse and low rank properties aiming at efficient capture of the correlations with each input matrix and further maximization of the inter-class hyperplane margin for better multiclass classification.

A. Objective Function

Given a c-class ($c \geq 2$) matrix form training data $\{X_i, y_i\}_{i=1}^n \in \{X, Y\}$, where $X_i \in \mathbb{R}^{pq}$ is the i^{th} feature matrix and $y_i \in \{1, 2, 3....c\}$ is the corresponding class label. The support matrix classifier ($arg\min \frac{1}{2}||W||_F^2 + C\sum_{i=1}^n \xi$) focuses on binary classification and hence incapable of dealing with multiclass problems. We devised a novel objective function that maximizes the margin between inter-class.

To maximize the inter class margin,

$$arg \min_{w^{d\times c, b^c}} \frac{1}{2}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c} ||W_j - W_k||_F^2 + \tau\sum_{j=1}^{c}||W_j - W_k||_* C\sum_{j=1}^{c}\sum_{k=j+1}^{c}\sum_{y_i\in j,k}\xi_i^{jk} \quad (9)$$

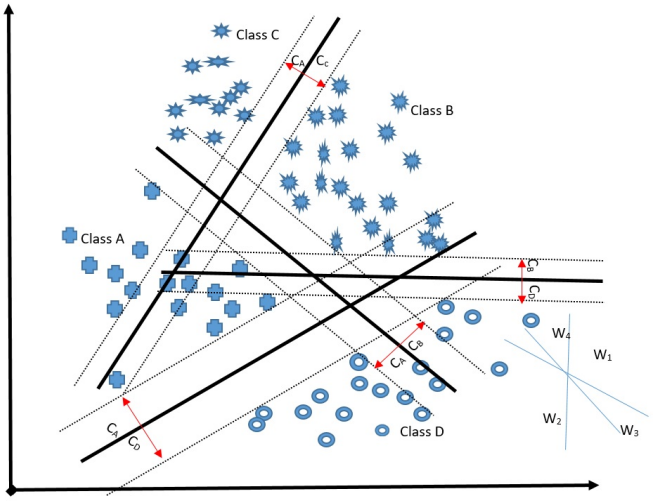

Fig. 1. Illustration of multiclass support matrix machine: For four classes, we need three parameters W_1, W_2, W_3, and W_4 to maximize the inter-class margins

such that

$$y_i^{jk} f_{jk}(X_i) \geq 1 - \xi_i^{jk},\ \forall y_i \in j, k$$
$$\xi_i^{jk} \geq 0$$

Where $W \in R^{pq}$ denotes the regression parameter in the form of tensor and $||X||_F$ is the Frobenius norm of W. The objective function in Eq. 9 resulted in multiple optimal solutions. In order to reach the objective function which provides single global optima, we further added constraints in the objective function as follow

$$arg \min_{W^{d\times c, b^c}} \frac{1}{2}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c} ||W_j - W_k||_F^2 + \tau\sum_{j=1}^{c}||W_j - W_k||_* + \frac{1}{2}\sum_{j=1}^{c} b_j^2 + C\sum_{j=1}^{c}\sum_{k=j+1}^{c}\sum_{y_i\in j,k}\xi_i^{jk} \quad (10)$$

such that

$$y_i^{jk} f_{jk}(x_i) \geq 1 - \xi_i^{jk},\ \forall y_i \in j, k$$
$$\xi_i^{jk} \geq 0$$

Whereas as $f_{jk}x_i = (W_j - W_k)^T x_i + b_j - b_k$ and $y_i^{jk} = \{1, -1\}$.

For classification of unseen data object, we follow the same voting strategy as in one-vs-one multiclass classification and simulated using C matrices. Thus, M-SMM does not require to compute $\frac{c(c-1)}{2}$ decision function, it only needs to compute the decision function c times and decided based on the largest value. Surprisingly, the problem could be solved using a simple yet efficient algorithm as shown in table 1.

B. Learning Algorithm

The objective function in Eq.10 consists of four terms and all of them are convex i.e. the Nuclear and Frobenius norm, that satisfies the triangle and homogeneity properties.

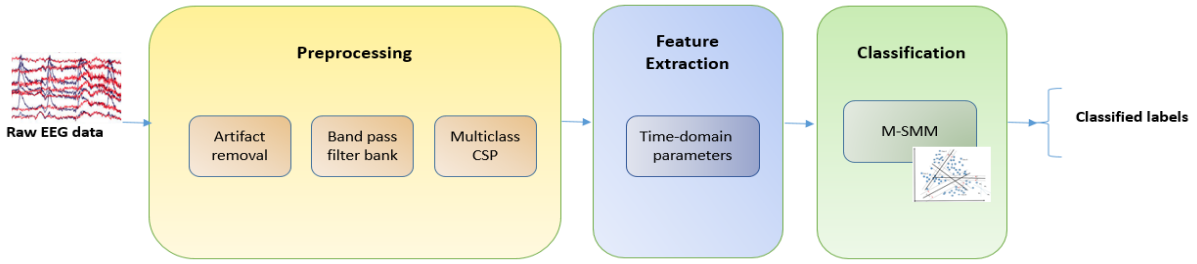

Fig. 2. Illustration of proposed framework equipped with M-SSM for EEG signal classification

The other two terms are linear functions, hence, they are also convex. In conclusion, the objective function in Eq.10 is convex but non-differentiable and non-smooth. In convex optimization setting, sub-gradient of the nuclear norm function cannot be used in standard descent approaches, as a result solving it directly is difficult. Thus alternative approach is required to update W. As we know, the dependency of matrix W can be revealed by its $rank(W)$, so we can impose rank on W. To conclude, rank matrix minimization is non-convex and NP-hard and can be solved as

$$arg \min_{W^{d\times c}, b^c} \frac{1}{2}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c} ||W_j - W_k||_F^2 + \tau \sum_{j=1}^{c} ||W_j - W_k||_* + \frac{1}{2}\sum_{j=1}^{c} b_j^2 + C\sum_{j=1}^{c}\sum_{k=j+1}^{c}\sum_{y_i \in j,k} [1 - \bar{y}_i^{jk} f_{jk}(x_i)]_+ \quad (11)$$

whereas as $W \in \mathbb{R}^{d\times c}$, $b \in \mathbb{R}^c$ and decision function $f_jk(x_i) = (W_j - W_k)^T x_i + (b_j - b_k).y_n^{jk}$ is the resultant class that is classified for unlabeled data.

We select the training sample randomly in each iteration. The objective function in Eq. 11 can be rewritten as

$$arg \min_{W^{d\times c}, b^c} \frac{1}{2}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c} ||W_j - W_k||_F^2 + \tau \sum_{j=1}^{c} ||W_j - W_k||_* + \frac{1}{2}\sum_{j=1}^{c} b_j^2 + C\left(\sum_{j=\bar{y}_i+1}^{c} [1 - \bar{y}_i^{jk} f_{\bar{y}_i j}(x_i)]_+ + \sum_{j=1}^{\bar{y}_i - 1} [1 + \bar{y}_i^{jk} f_{j\bar{y}_i}(x_i)]_+\right) \quad (12)$$

As all the terms in objective function in Eq. 12 are non-smooth and non-differential, thus, stochastic gradient descent and Nesterov methods can not be applied. Since the objective function is convex in all four terms, we have employed widely used framework ADMM for convex optimization problem, by breaking the objective function into sub-problems that are easier to optimize.

The problem in Eq.12 can be equivalently written as,

$$arg \min_{W,b} P(W) + Q(S)$$

$$s.t \quad S - W = 0$$

Where $S \in \mathbb{R}^{P\times Q\times k}$ is an additional decision variable to split the primal problem into two sub problems.

TABLE I
ALGORITHMIC PROCEDURE OF SPARSE SUPPORT MATRIX MACHINE

Input: : Labeled Training dataset: $[X_i, y_i]$ where $X_j \in \mathbb{R}^{m\times n}$ for $j = 1, ..., N$, Lagrangian multiplier$\mathcal{L}$, learning rate $\eta \in \{0, 1\}$, $p > 0$,$t = 1$, C, τ
Output: Matrix W

Step-I: Initialize the $W, S, \mathcal{L} = 0$
While not converge **do**
Step-II Minimize S with respect to W

$$\min_S L_S = G(S) + \langle \mathcal{L}, S\rangle + \frac{p}{2}||W - S||_F^2$$

Step-III Minimize W with respect to S

$$\min_S L_W = H(W)_+ + \langle -\mathcal{L}, W\rangle + \frac{p}{2}||S - W||_F^2$$

for $i = 1$ to c do
Step-IV: **if** $1 - f_{jk}(x_i) \leq 0$

$$\nabla W = \frac{1}{p+1}(\Lambda + pS - Cx_i)$$

Step-V: **if** $1 - f_{jk}(x_i) \leq 0$

$$\nabla W = \frac{1}{p+1}(\Lambda + pS + Cx_i)$$

end for
for $i = 1$ to n **do**
Pick $i_t \in 1, 2, ...n$ randomly and update parameter
for $i = 1$ to c **do**

$$W = W - \eta \nabla W$$

$$S = \frac{1}{p}\mathbb{D}_\tau(pW - \Lambda)$$

$$\mathcal{L} = \mathcal{L}^t + p(S^{t+1} - W^{t+1})$$

$$p^{t+1} = \beta p^t$$

end while

$$P(W,b) = \frac{1}{2}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c}||W_j - W_k||_F^2 + \frac{1}{2}\sum_{j=1}^{c} b_j^2 + C\left(\sum_{j=\bar{y}_i+1}^{c}[1-\bar{y}_i^{jk} f_{\bar{y}_i j}(x_i)]_+ + \sum_{j=1}^{\bar{y}_i-1}[1+\bar{y}_i^{jk} f_{j\bar{y}_i}(x_i)]_+\right) \tag{13}$$

and

$$Q(S) = ||W_j - W_k||_* \tag{14}$$

where $P(W)$ is the hinge loss function obtained from negative likelihood, $Q(S)$ is an additional penalty function defined on singular value of matrix. For simplicity, we used term W instead of $W_j - W_k$ and W_i. To solve the Eq. 12, we applied augmented Lagrangian method and obtained

$$L(W,b,S,) = P(W)+G(S)+\frac{p}{2}||S-W||_F^2+\langle\mathcal{L},(S-W)\rangle \tag{15}$$

where $p > 0$ is the hyperparameter and $\mathcal{L}$ is the Lagrange multiplier.

We have divided the optimization problem in Eq. 12 into two sub-problems W and S. Solving it iteratively, we first needed to minimize S and W followed by updating the Lagrangian multiplier accordingly as,

$$S^{t+1} = arg\min_S L(S, W^t, \mathcal{L}^t) \tag{16}$$

$$W^{t+1} = arg\min_S L(S^{t+1}, W, \mathcal{L}^t) \tag{17}$$

$$\mathcal{L} = \mathcal{L}^t + p(S^{t+1} - W^{t+1}) \tag{18}$$

Where t and t+1 are the t^{th} and $(t+1)^{th}$ iterations respectively.

Minimizing the objective function in Eq. 16 with respect to S by fixing W, is to minimize the sum of all terms S term. Assuming W is fixed, we get

$$\min_S L_S = G(S) + \langle\mathcal{L},S\rangle + \frac{p}{2}||W-S||_F^2 \tag{19}$$

To update S, Eq. 19 can be solved by minimizing L_s. As L_s is non-differential but convex, the sub-gradient of L_s is computed as (see the proof in theorem 1)

$$S^{t+1}\frac{1}{p}\mathbb{D}_\tau(pW - \mathcal{L}) = \frac{1}{p}U_0(\Sigma_0 - \tau I)V_0^T \tag{20}$$

Where $\mathbb{D}$ is the singular value threshold operator.

Theorem 1. For $\tau \geq 0$, one optimal solution for the following problem

$$\min_S L_S = G(S) + \langle\mathcal{L},S\rangle + \frac{p}{2}||W-S||_F^2$$

is

$$S_c^{t+1} = \frac{i}{1+p}\mathbb{D}_\tau(pW_c - V_c)$$

Where $\mathbb{D}_\tau$ is the singular value thresholding operator (defined in section III).

Similarly, fixing S and minimizing the objective function with respect to W

$$\min_S L_W = H(W)_+ + \langle -\mathcal{L}, W\rangle + \frac{p}{2}||S-W||_F^2 \tag{21}$$

The Eq. 21 is the non-negative convex sum of term $H(W)$ (combination of hinge loss, Frobenius norm and penalty term) and linear and square functions.

Here, we have two different cases i.e. $j = y_i$ and $j \neq y_i$. Considering $j = y_i$ first, we have

$$W^{t+1} = \frac{1}{p+1}(\mathcal{L} + pS + \begin{cases} -Cx_i & if \quad 1 - f_{jk}(x_i) > 0 \\ 0 & if \quad 1 - f_{jk}(x_i) \leq 0 \end{cases} \tag{22}$$

Similarly, when $j = y_i$, W is updated as

$$W^{t+1} = \frac{1}{p+1}(\mathcal{L} + pS + \begin{cases} Cx_i & if \quad 1 - f_{jk}(x_i) > 0 \\ 0 & if \quad 1 - f_{jk}(x_i) \leq 0 \end{cases} \tag{23}$$

Finally the Lagrangian multiplier can be updated as

$$\mathcal{L} = \mathcal{L}^t + p(S^{t+1} - W^{t+1}) \tag{24}$$

$$p^{t+1} = \beta p^t \tag{25}$$

Theorem 2. Optimal solution of S^{t+1} such that

$$\min_S L_S = G(S) + \langle\mathcal{L},S\rangle + \frac{p}{2}||W-S||_F^2$$

satisfies $0 = \partial L_s(S_c^{t+1})$, now we are required to find one S_c subject to

$$0 \in S_c + \tau\partial||S_c||_* + \mathcal{L} + p(S_c - W_c)$$

Let $U_c\Sigma_c V_C^T$ denotes the singular decomposition of an arbitrary matrix S_c.

Sub gradient of nuclear norm (defined in section III) $\partial||S_c||_*$ is

$$\partial||S_C||_* = U_C V_c^T + Z : Z \in \mathbb{R}^{l_1 \times l_2}, U_c^T$$

The above equation can be rewritten as

$$\partial||S_C||_* = 0, V_c$$

Which can be simplified as

$$\partial||S_C||_* = 0, ||Z||_F < 1$$

Let Y denotes $PW_c - \mathcal{L}_c$ and decompose it as $Y = U_1\Sigma_1 V_1^T + U_2\Sigma_2 V_2^T$ where U and V are the singular vectors associated with singular values greater than

τ (smaller than or equal).

If $S_c = \frac{U_1(\Sigma_1 - \tau I)V_1^T}{1+p}$; according to $0 \in S_c + \tau\partial||S_c||_* + \mathcal{L} + p(S_c - W_c)$, we have the following relation

$$\partial||S_c||_* = \frac{1}{\tau}[Y - (1+p)S_c]$$

The above Eq. can be simplified as

$$\partial||S_c||_* = U_1V_1^T + \frac{1}{\tau}U_2\Sigma_2V_2$$

Consider $Z = U_2\Sigma_2V_2$, $U_c = U_1$ and $V_c = V_1$, we have $0 \in \partial L_s$ when $S_c^* = S_c$

C. Theoretical Justification

In this section, we theoretically analyze and illustrate how M-SMM possesses some elegant features as compared to conventional SVM, conventional elastic net SMM [27] and MSMM [26]. As discussed earlier, data is unbiased in real world, thus one versus rest class problem will not work and will affect the performance. Similarly, one versus one strategy has high space and time complexity, especially in case of matrix data, since it requires training of $\frac{c(c-1)}{2}$ SVM classifiers. M-SMM works same as one vs one fashion and does not to use voting strategy and compute decision function for each class. We build a classifier for every two classes, however, different from one versus one strategy, it use C matrices to simulate all these binary classifiers, thus it does not need to use vote strategy c(c1)/ 2 times. It just need to compute decision function c times . This results in reduction of space complexity to same level as one-vs-rest strategy and find the largest value.

We now show that S-SMM is the generalization of multi-class SMM. Considering the hinge loss of proposed objective function as shown in Eq.10.

$$C\sum_{j=1}^{c-1}\sum_{k=j+1}^{c}\sum_{y_i \in j,k}\xi_i^{jk}$$

The above equation can be written as

$$C\sum_{j=1}^{c-1}\sum_{k=j+1}^{c}\sum_{y_i \in j,k}[1 - \bar{y}_i^{jk}f_{jk}(X_i)]_+$$

$$= C\sum_{j=1}^{c-1}\sum_{k=j+1}^{c}\left(\sum_{y_i \in j}[1 - f_{jk}(x_i)]_+ + \sum_{y_i \in k}[1 - f_{kj}(X_i)]_+]\right)$$

$$= C\sum_{j=1}^{c-1}\sum_{y_i \in j}\sum_{k=j+1}^{c}[1 - ((W_j - W_k)^T X_i + (b_j - b_k)))]_+$$

$$= C\sum_{i=1}^{n}\sum_{k \neq y_i}[1 - ((W_{y_i} - W_k)^T X_i + (b_{y_i} - b_k)))]_+$$

The objective function in Eq. 12 can be written as below, which is MSMM [26].

$$arg\min_{W^{d\times c}, b^c} \frac{1}{2}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c}||W_j - W_k||_F^2 + \sum_{j=1}^{c}||W_j - W_k||_* + \frac{1}{2}\sum_{j=1}^{c}b_j^2 + C\sum_{j=1}^{c}\sum_{i=1}^{n}[1 - \bar{y}_i^{jk}f_{jk}(X_i)]_+ \quad (26)$$

The proposed objective function degenerate to SMM.

VI. Experimental Evaluation

In this section, we described the experimental setup and evaluation of the proposed approach. To validate the effectiveness of proposed classifier, we extensively evaluated the proposed M-SMM and compared it with MSMM [26], SMM [27], BSMM [35], MSVM [36], KNN [37] and SCSSP[25] as well as winners of BCI competitions on benchmark EEG datasets (IIIa and IIa) using four different evaluation metrics (recall, prevision, F-measure and kappa coefficient).

A. Dataset

In this experiment, we have used two publicly available benchmark data-sets namely IIIa (BCI competition III)[1] and IIa[2] (BCI competition IV). IIIa consisted of 60 channel single trial EEG signal obtained from three subjects(k3b, k6b and l1b) while performing four classes of motor imagery (left-hand, right-hand, foot and tongue labeled as class 1, 2, 3 and 4 respectively). IIIa consisted of 45, 30, 30 trials per class for subject k3b, k6b and l1b respectively. Similarly, IIa data-set collected in two sessions from nine subjects performing four classes of motor imagery (left-hand, right-hand, foot and tongue). IIIa consisted of 288 in total (72 trails per motor imagery). It consisted of 22 EEG channels and 3 monopoloar EOG channels. IIIa and IIa are sampled with 250 Hz and band-pass filtered between 0.5 Hz and 100 Hz. In this experiment, we have considered two subjects (k6b and l1b) for IIIa data-set and EEG channel for IIa data-set.

We have conducted k-fold ($k = 5$) cross validation to analyze the generalization of the results to an independent dataset. The reason behind k-fold cross validation is that, it guarantees that each sample eventually become the part of training as well as testing sets. For this purpose, we have divide the trials of each subjects into 5 sets. We have repeated the experiments five times and each time different test set is selected while others four sets are considered as training dataset.

B. Evaluation Metrics

In order to evaluate the performance of proposed classifier, we employed different evaluation metrics such as kappa coefficient, precision, recall and F-measure. Furthermore, we have also compared the training time with state of the art approaches. Kappa measure provide evaluation comparison as

[1]http://www.bbci.de/competition/iii/download
[2]http://www.bbci.de/competition/iv/dataset2a

TABLE II
KAPPA/ERROR RATE %: CLASSIFICATION PERFORMANCE OF DIFFERENT ALGORITHMS ON DATA-SET IIIA

Subject	BCI Com.	KNN	MSVM	SCSSP	SMM	BSMM	MSMM	M-SMM
k3b	0.83/18.6	0.81/14	0.89/8.3	0.71/22.3	0.852/11.1	0.94/4.4	0.948/3.9	0.961/3.6
l1b	0.74/22.1	0.49/38	0.68/24.2	0.69/36.2	0.71/21.7	0.8/15	0.811/14.2	0.85/13.2
Avg	**0.78/19.8**	**0.65/26**	**0.78/16.3**	**0.64/23.6**	**0.78/16.4**	**0.87/9.7**	**0.88/9.0**	**0.89/9.2**

TABLE III
KAPPA/ERROR RATE%: CLASSIFICATION PERFORMANCE OF DIFFERENT ALGORITHMS ON DATASET IIA

Sub	BCI Comp.	KNN	MSVM	SCSSP	BSMM	SMM	MSMM	M-SMM
S1	0.68/24	0.71/22	0.72/21	0.62/26	0.73/21	0.69/0.23	0.73/20	0.76/18
S2	0.42/44	0.4/45	0.37/47	0.28/54	0.4/45	0.23/0.58	0.43/43	0.44/39
S3	0.75/19	0.77/17	0.76/17	0.6/26	0.75/19	0.69/0.24	0.84/11	0 0.84/8.4
S4	0.48/39	0.45/41	0.36/48	0.33/51	0.51/37	0.54/0.35	0.59/31	0.64/28
S5	0.4/45	0.38/47	0.42/43	0.15/64	0.39/46	0.32/0.51	0.5/38	0.55/41
S6	0.27/55	0.24/57	0.19/61	0.25/56	0.32/51	0.15/0.63	0.41/44	0.45/39
S7	0.77/17	0.69/23	0.66/25	0.41/44	0.81/14	0.72/0.21	0.85/12	0.88/11
S8	0.76/18	0.62/29	0.45/41	0.6/31	0.71/22	0.71/0.22	0.77/17	0.81/13.7
S9	0.61//26	0.48//39	0.56/33	0.66/25	0.62/29	0.63/0.27	0.72/21	0.77/14
avg	**0.57/32**	**0.53/36**	**0.5/37**	**0.44/42**	**0.58/31**	**0.52/0.36**	**0.65/26**	**0.74/16**

it consider the accuracy occurring by chance better. Higher the value of k means gain is classification performance and $k > 0$ shows the gain is better than random guess. It is defined as $k = \frac{accuracy - p_o}{1 - p_o}$. Here, p_o is the random guess i.e. for a k-class dataset with balanced sample sizes among different classes, we have $p_o = \frac{1}{k}$. The other evaluation measures we have used are precision, recall and F measure. Precision also referred as positive predictive value (PPV) is the true positive relevant measure and is calculated as $P = \frac{tp}{tp+fp}$. Recall is referred to as the true positive rate or sensitivity, is the ratio of correctly predicted positive observations to the all observations in actual class. Recall is calculated as $R = \frac{t_p}{t_p+f_n}$. F1 score takes both false positives and false negatives into account, is the weighted average of precision and recall. It is needed when we are seeking a balance between precision and recall. It is calculated as $F_1 = 2\frac{R \times P}{R+P}$.

C. EEG Preprocessing and Feature Extraction

Motor imagery-based BCI, which translates the mental imagination of movement to commands, is the huge inter-subject variability with respect to the characteristics of the brain signals [38]. Furthermore, poor characteristics of EEG data such as measurement artifacts, outliers and non-standard noises make it challenging task. In order to reduce the variations, spatial filtering has prevent itself as an effective method for extraction of features has been used as a preprocessing technique to explore the discriminative spatial patterns and eliminate uncorrelated information. In this paper, we have used Filter Bank Common Spatial Pattern (FBCSP) algorithm [38] to filter out the artifacts and unrelated sensorimotor rhythms by performing autonomous selection of discriminative subject-specific frequency range for band-pass filtering of the EEG measurements. To select dominant channels for each motor imagery task, we have applied CSP [20] followed by Time domain parameters for feature selection [39] due to its robust performance [3], [26], [40]. We have fed the time domain parameters to mulit-class support matrix machines for classifications.

D. Results

The main goal of this work is to elucidate the best comparable performance as compared to state of the art approaches followed by computational complexity. In this experiment, we have used four evaluation measures to compare the performance of proposed approach with seven state of the art approaches on two publicly EEG data-sets. As our contribution is on classification of matrix data, thus, to compare the proposed classifier for fair comparison, we employed the same preprocessing and feature extraction approach for other approaches. The evaluation results on data-set IIIa shown in table II and IV obtained highest score in the validation procedure. We have transformed the matrix into vectors followed by PCA for dimensioanlity reduction for vector based methods such as BCI competition winner, MSVM, KNN, and SCSSP. To compare the performance on multiclass problem, we have extended the approaches using OvR strategy except MSMM.

We have also computed the error rate in Kappa measure for better comparison. The evaluation results on data-set IIa are shown in table III, and V. From results of both data-sets, we observed that classifier based on maximizing the inter-class hyperplane margin for matrix data provided better results as compared to those methods based on vectors. It further validated that leveraging the structural information of data is greatly beneficial to the improvement of the classification performance.

Notice that, the objective function consist of $\tau \sum_{j=1}^{c} ||W_j - W_k||_*$. Here τ manages the penalty by controlling the number of low rank of the regression parameter. It determine the structural information. Large value of τ impose heavy penalty that set most of the singular values in the regression parameter to zero which results in losing most structural information embedded in data. Figure 3 shows the convergence process

of M-SMM on subjects k3b and l1b of IIIa dataset. We have used ADMM for convex optimization problem, by breaking the objective function into sub-problems that are easier to optimize. Notice that M-SMM converges to the global optimum in only few iterations. Similar trends also occur IIa dataset.

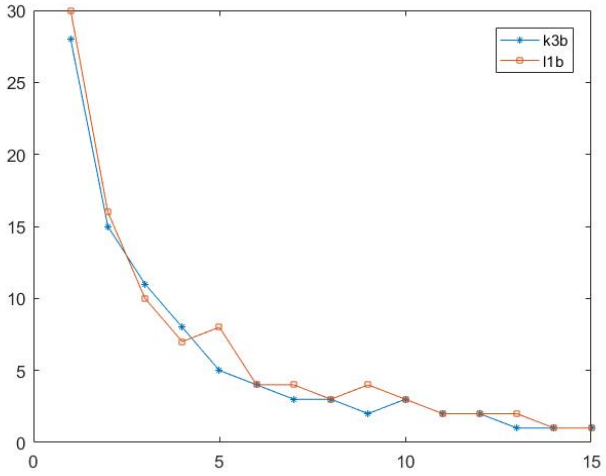

Fig. 3. Convergence process of M-SMM on subject k3b and l1b of IIIa dataset

E. Parameter Setting

The objective function that is combination of Frobenius norm, nuclear norm and hinge loss function, thus there are several parameters τ, p, learning rate η, t and C, are required to be adjusted, in order to compute the objective function. τ is a penalty added on the nuclear norm that captures the correlation of data matrix. Thus, it determine how much structural information is involved in the classication. We notice that magnitude of τ manages the penalty on nuclear norm by controlling the number of singular value (rank) of the regression parameter. Large value of τ results powerful penalty on the structure information as a results most of the singular values in the regression parameter are set to zero which results in losing most structural information embedded in data. We observe that the the proposed model degenerates to the problem [34] for vector data, when $\tau = 0$. Figure 4 validate the aforementioned claim. Notice that results are same as of MSMM when $\tau = 0$, similarly, the results starts to degrade when τ is larger. Thus, we concluded that the proposed model is a generalization of SVM and possess sparse and low-rank properties. As a result, it considers correlation among matrices and performs feature selection simultaneously. In this experiment, we have set the learning rate $\eta = 0.21$, $\tau = 2.6$ and $p = 3$ for IIIa dataset and learning rate $\eta = 0.23$, $\tau = 3$ and $p = 3$ for IIa dataset.

F. Computational Complexity

One of the major objectives of proposed approach was computational efficiency. As discussed in earlier sections, the existing methods required $\frac{c(c-1)}{2}$ support matrix machines, that is computational complex In this work, we have used same strategy as of OvsO, however rather than computation of $\frac{c(c-1)}{2}$ support matrix machine, we simulated the OvsO strategy using using c support vectors. To investigate the computational efficiency studies of the classification model. We have compared the run time of the algorithms based on matrix data. The experiments was conducted on Intel Xeon E5-1620,3.7GHz, 16GB RAM, Window 7. We compared the average training and testing time on both data-sets between different methods. We have only selected classifiers (i.e. SMM, BSMM and MSMM) based on matrix data. The average training and testing time on both data-sets are shown in table VI. It can be depicted that M-SMM training time is comparable with other approaches however, in comparison of the testing time, it is much faster. The reason behind more training time and better testing time is that we have more

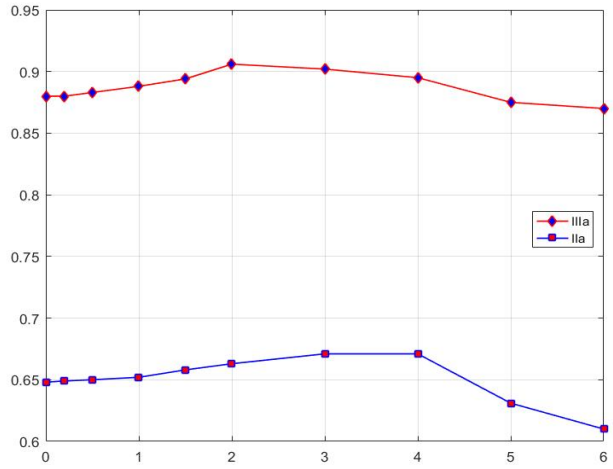

Fig. 4. Behaviour of τ on on the classication performance for IIa and IIIa datasets

TABLE IV
COMPARATIVE EVALUATION OF CLASSIFICATION PERFORMANCE OF DIFFERENT ALGORITHMS ON IIIA DATA-SET

Method	Kappa	Precision	Recall	F_1 Score
KNN	0.732	0.768	0.799	0.804
MSVM	0.784	0.85	0.838	0.844
BSMM	0.871	0.91	0.903	0.906
SMM	0.782	0.847	0.836	0.841
MSMM	0.880	0.916	0.91	0.913
M-SMM	**0.916**	**0.927**	**0.918**	**0.922**

TABLE V
COMPARATIVE EVALUATION OF CLASSIFICATION PERFORMANCE OF DIFFERENT ALGORITHMS ON IIA DATA-SET

Method	Kappa	Precision	Recall	F 1 Score
KNN	0.527	0.684	0.645	0.663
MSVM	0.499	0.689	0.624	0.653
BSMM	0.581	0.715	0.686	0.7
SMM	0.519	0.674	0.64	0.656
MSMM	0.648	0.751	0.736	0.744
M-SMM	**0.671**	**0.793**	**0.766**	**0.761**

TABLE VI
COMPARISON OF AVERAGE TRAINING AND TESTING TIME (IN SECONDS) ON IIIA AND IIA DATA-SETS

Classifier	IIIa		IIa	
	Training	Testing	Training	Testing
SMM	18.995	0.0594	47.198	0.243
BSMM	20.381	0.0636	47.198	0.243
MSMM	22.257	0.0541	65.528	0.230
M-SMM	**24.366**	**0.0414**	**67.261**	**0.161**

number of parameters in training whereas we require C vector for testing respectively.

G. Discussion

In this section, we provide the comprehensive analysis of the proposed approach. Notice that, the M-SMM achieved better performance as compared to the state of the art methods. Results shows that proposed approach is able to finds the representative features from high-dimensional space that are used for classification. Nuclear norm promotes the structural sparsity and shares similar sparsity patterns across multiple predictors. τ determines the level of structural information involved in the classication by controlling the number of singular value (rank) of the regression parameter. This means greater the value of τ could account more structural information encoded in the matrix results in improving the classication accuracy. M-SMM reveals the geometric structure embedded in the data due to the fact that it select the features by maintaining the spatial structural information of the matrix.

Comparing with aforementioned experimental evaluation, we have the following interesting observations

(I) M-SMM degenerates to the problem [34] for vector data, when $\tau = 0$. Thus, it is a generalization of SVM and possess sparse and low-rank properties.

(II) Larger value of τ results powerful penalty on the structure information. However, too large value of τ results in decreasing the performance due to the fact that high value of τ results in setting the singular values in the regression parameter to zero which discard the structural information embedded in matrix.

In this work, we have presented a multiclass support matrix machines with the perspective of maximizing the intra-class margins. As a case study, we solved one of the important problem of EEG classification to show the performance of the proposed approach. Results showing considerable improvement in accuracy as well as the computational complexity is also attractive. Although in this experiment, we have applied EEG dataset for validation, however proposed approach is general machine learning classifier and could be applied to any high dimensional data involving multiclass problem.

VII. CONCLUSION

In this work, we presented an novel classifier name Multiclass Support Matrix Machine (M-SMM) from the perspective of maximizing the intra-class margins (maximizing the distance between training point and hyper-plane) for multiclass classification of high dimensional data such as EEG classification. We combined the hinge loss, nuclear and Frobenius norm and followed the idea of maximizing the margin between two-class problem and use c support matrices to simulate all binary classifier rather than computing support vector between every two classes. The objective function not only maximized the inter-class margins but was spectral extension of conventional elastic net that combines the property of low rank and joint sparsity together to deal with complex high dimensional noisy data. Hence resulted in an improved classification performance supported by the experimental evaluation. The M-SMM has achieved 0.916 k value for IIIa in comparison to 0.88 and 0.782 for MSMM and SSM respectively. Similarly, 0.671 k value for IIa in comparison to 0.648 and 0.519 for MSMM and SSM respectively. In conclusion, the numerical results suggest that our method is superior to previous approaches and demonstrates the promise of M-SMM for real-world applications.

REFERENCES

[1] Y. Li, J. Pan, J. Long, T. Yu, F. Wang, Z. Yu, and W. Wu, "Multimodal bcis: target detection, multidimensional control, and awareness evaluation in patients with disorder of consciousness," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 332–352, 2016.

[2] J. Jin, E. W. Sellers, S. Zhou, Y. Zhang, X. Wang, and A. Cichocki, "A p300 brain–computer interface based on a modification of the mismatch negativity paradigm," *International journal of neural systems*, vol. 25, no. 03, p. 1550011, 2015.

[3] Q. Zheng, F. Zhu, J. Qin, B. Chen, and P.-A. Heng, "Sparse support matrix machine," *Pattern Recognition*, vol. 76, pp. 715–726, 2018.

[4] R. H. Grabner, A. C. Neubauer, and E. Stern, "Superior performance and neural efficiency: The impact of intelligence and expertise," *Brain research bulletin*, vol. 69, no. 4, pp. 422–439, 2006.

[5] M. I. Razzak, M. Imran, and G. Xu, "Big data analytics for preventive medicine," *Neural Computing Applications*, pp. 1–35, 2019.

[6] M. I. Razzak, R. A. Saris, M. Blumenstein, and G. Xu, "Robust 2d joint sparse principal component analysis with f-norm minimization for sparse modelling: 2d-rjspca," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[7] H. Zhou and L. Li, "Regularized matrix regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 463–483, 2014.

[8] F. Qi, Y. Li, and W. Wu, "Rstfc: A novel algorithm for spatio-temporal filtering and classification of single-trial eeg," *a, a*, vol. 1, p. 1, 2015.

[9] A. Subasi and M. I. Gursoy, "Eeg signal classification using pca, ica, lda and support vector machines," *Expert systems with applications*, vol. 37, no. 12, pp. 8659–8666, 2010.

[10] Y. Zhang, Y. Wang, J. Jin, and X. Wang, "Sparse bayesian learning for obtaining sparsity of eeg frequency bands based feature vectors in motor imagery classification," *International journal of neural systems*, vol. 27, no. 02, p. 1650032, 2017.

[11] Y. Zhang, Y. Wang, G. Zhou, J. Jin, B. Wang, X. Wang, and A. Cichocki, "Multi-kernel extreme learning machine for eeg classification in brain-computer interfaces," *Expert Systems with Applications*, vol. 96, pp. 302–310, 2018.

[12] W.-C. Hsu, L.-F. Lin, C.-W. Chou, Y.-T. Hsiao, and Y.-H. Liu, "Eeg classification of imaginary lower limb stepping movements based on fuzzy support vector machine with kernel-induced membership function," *International Journal of Fuzzy Systems*, vol. 19, no. 2, pp. 566–579, 2017.

[13] D. S. de Lucena, S. R. Moreno, V. C. Mariani, and L. dos Santos Coelho, "Support vector machine optimized by artificial bee colony applied to eeg pattern recognition," in *Brain Function Assessment in Learning: First International Conference, BFAL 2017, Patras, Greece, September 24-25, 2017, Proceedings*, vol. 10512. Springer, 2017, p. 213.

[14] A. Datta and R. Chatterjee, "Comparative study of different ensemble compositions in eeg signal classification problem," in *Emerging Technologies in Data Mining and Information Security*. Springer, 2019, pp. 145–154.

[15] M. Li, H. Xu, X. Liu, and S. Lu, "Emotion recognition from multichannel eeg signals using k-nearest neighbor classification," *Technology and Health Care*, no. Preprint, pp. 1–11, 2018.
[16] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals," *Computers in biology and medicine*, vol. 100, pp. 270–278, 2018.
[17] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated eeg-based screening of depression using deep convolutional neural network," *Computer methods and programs in biomedicine*, vol. 161, pp. 103–113, 2018.
[18] M. Li, W. Chen, and T. Zhang, "Classification of epilepsy eeg signals using dwt-based envelope analysis and neural network ensemble," *Biomedical Signal Processing and Control*, vol. 31, pp. 357–365, 2017.
[19] H. Dose, J. S. Møller, S. Puthusserypady, and H. K. Iversen, "A deep learning mi-eeg classification model for bcis," in *2018 26th European Signal Processing Conference*. IEEE, 2018.
[20] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller, "Spatio-spectral filters for improving the classification of single trial eeg," *IEEE transactions on biomedical engineering*, vol. 52, no. 9, pp. 1541–1548, 2005.
[21] R. Tomioka, G. Dornhege, G. Nolte, B. Blankertz, K. Aihara, and K.-R. Müller, "Spectrally weighted common spatial pattern algorithm for single trial eeg classification," *Dept. Math. Eng., Univ. Tokyo, Tokyo, Japan, Tech. Rep*, vol. 40, 2006.
[22] X. Liao, D. Yao, D. Wu, and C. Li, "Combining spatial filters for the classification of single-trial eeg in a finger movement task," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 821–831, 2007.
[23] H. Zhang, H. Yang, and C. Guan, "Bayesian learning for spatial filtering in an eeg-based brain–computer interface," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 7, pp. 1049–1060, 2013.
[24] T.-E. Kam, H.-I. Suk, and S.-W. Lee, "Non-homogeneous spatial filter optimization for electroencephalogram (eeg)-based motor imagery classification," *Neurocomputing*, vol. 108, pp. 58–68, 2013.
[25] A. S. Aghaei, M. S. Mahanta, and K. N. Plataniotis, "Separable common spatio-spectral patterns for motor imagery bci systems," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 1, pp. 15–29, 2016.
[26] Q. Zheng, F. Zhu, J. Qin, and P.-A. Heng, "Multiclass support matrix machine for single trial eeg classification," *Neurocomputing*, vol. 275, pp. 869–880, 2018.
[27] L. Luo, Y. Xie, Z. Zhang, and W.-J. Li, "Support matrix machines," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, pp. 938–947. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045219
[28] L. Wolf, H. Jhuang, and T. Hazan, "Modeling appearances with low-rank svm," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–6.
[29] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Bilinear classifiers for visual recognition," in *Advances in neural information processing systems*, 2009, pp. 1482–1490.
[30] M. Dyrholm, C. Christoforou, and L. C. Parra, "Bilinear discriminant component analysis," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 1097–1111, 2007.
[31] K. Zhang, L. Lan, Z. Wang, and F. Moerchen, "Scaling up kernel svm on limited resources: A low-rank linearization approach," in *Artificial Intelligence and Statistics*, 2012, pp. 1425–1434.
[32] A. F. Martins, N. A. Smith, P. M. Aguiar, and M. A. Figueiredo, "Structured sparsity in structured prediction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1500–1511.
[33] Y. Guermeur, "Combining discriminant models with new multi-class svms," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 168–179, 2002.
[34] J. Xu, X. Liu, Z. Huo, C. Deng, F. Nie, and H. Huang, "Multi-class support vector machine via maximizing multi-class margins," in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 3154–3160.
[35] T. Kobayashi and N. Otsu, "Efficient optimization for low-rank integrated bilinear classifiers," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 474–487.
[36] T. Joachims, T. Finley, and C.-N. J. Yu, "Cutting-plane training of structural svms," *Machine Learning*, vol. 77, no. 1, pp. 27–59, 2009.
[37] S. Bhattacharyya, A. Khasnobish, S. Chatterjee, A. Konar, and D. Tibarewala, "Performance analysis of lda, qda and knn algorithms in left-right limb movement classification from eeg data," in *Systems in Medicine and Biology (ICSMB), 2010 International Conference on*. IEEE, 2010, pp. 126–131.
[38] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b," *Frontiers in Neuroscience*, vol. 6, p. 39, 2012. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2012.00039
[39] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer, "Eeg-based discrimination between imagination of right and left hand movement," *Electroencephalography and clinical Neurophysiology*, vol. 103, no. 6, pp. 642–651, 1997.
[40] C. Vidaurre, N. Krämer, B. Blankertz, and A. Schlögl, "Time domain parameters as a feature for eeg-based brain–computer interfaces," *Neural Networks*, vol. 22, no. 9, pp. 1313–1319, 2009.