Cooperative Evolution Multiclass Support Matrix Machines

Imran Razzak
School of Information Technology
Deakin University, Geelong, Australia
imran.razzak@ieee.org

***Abstract*—Support Matrix Machines are one the efficient learning approach for the classification of complex nature data. However, either it can only deal with binary class problem or can deal with multi-class classification problem by breaking the problem into number of binary class problem and solving them individually or through solving larger optimization. Aiming to improve performance of support matrix machines, in this paper, we present Multi-class Support Matrix Machine based on evolutionary optimization (MSMM-CE) by breaking down the original multi-class problem of support matrix into sub-problems in cooperative fashion. The proposed objective function is a combination of binary hinge loss function for specific class, Frobenius and nuclear norms as a penalty that promote low rank and sparsity as well as an additional penalty term to penalize the multiclass classification error. The additional penalty term allow us to decompose the problem into sub-problems and solving them in simultaneously in cooperative fashion. The proposed objective function learns for each class and consider the information from other classes, that results in solving the problem in parallel. A comprehensive experimental study on publicly available benchmark EEG dataset is carried out to investigate the proposed approach that confirms the superiority of MSMM-CE for accurate classification of EEG signal associated with motor imagery in BCI applications. MSMM-CE provides a generalized solution to investigate the complex and nonlinear high dimensional data for various real-world applications.**

***Index Terms*—SVM, matrix classification, cooperative evolution, SMM, support matrix machine, multiclass classification**

I. Introduction

Recent advancement in data acquisition devices are generating massive size of high dimensional data that has increased the scope of classifying it directly without losing information during vectorization. Thus, researchers are focusing on development of efficient methods for classifying data directly from matrix without converting it into vectors, which exploits the correlation between the columns or rows of matrix. Rank-k SVM, models the regression matrix as a sum of k rank-one orthogonal matrix [1]. Luo et. al. combined hinge loss, nuclear and Forbenius norm that captured the correlation within each matrix [2]. Zheng et al. presented Sparse Support Matrix Machine (SSMM) which is a combination of hinge loss, nuclear norm and ℓ_1 norm and can simultaneously capture the intrinsic structure of each matrix and select useful features as well. [3]. 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, the methods based on matrix expect MSMM are originally developed for binary classication problem [4]. Although, these methods can be used as multi-class classifier by breaking down the multiclass problem into several two class problems such as One-vs-Rest (OvR) or One-vs-One (OvO) approaches (e.g. In OvsR strategy, we solve the mutli-class problem by splitting it into n binary problems, whereas OvsO strategy can be solved by splitting the problem into $\frac{c(c-1)}{2}$ binary class problem but are computationally expensive and may results in unbalanced distribution of input samples. Nonetheless, these methods suffer for complex optimization.

Evolutionary algorithms solve the complex optimization problem by decomposing the problem into single optimization and has recently has been applied to many machine learning problems [5], [6]. One of its integration is with with support vector machines [4]–[11]. Most of these methods are either binary class optimization or deal hyper parameter optimization. In order to address the aforementioned limitations, we present a novel classification approach named Cooperative Evolution Mulitclass Sparse Matrix Machine (MSMM-CE) through solving the complex optimization problem in single objective optimization by taking the advantage of problem decomposition of multiclass problem into single the regression matrix that is not only low-rank, but sparse. MSMM-CE is a combination of hinge loss for model fitting, elastic net penalty as a regularization on the regression matrix and an additional penalty term (cooperative penalty term) that consider the information from other classes. The regularization term is a linear combination of Frobenius norm and nuclear norm to control the low rank properties whereas the cooperative penalty term penalizes the errors occurs in classification. Thus, MSMM-CE takes full advantage of low rank plus sparsity and decomposition of classification problem into sub problem in a fashion that each sub problem learns the support matrix of specific class but also consider information from other classes.

Nonetheless, these methods are either are too complex to optimize or mainly for binary class problem. Recent success of evolutionary algorithms integration in to machine learning algorithms explicitly SVM showed its advantage to solve complex optimization problem. Co-evolutionary algorithms are able to optimize more population simultaneously by discomposing the problem into different sub problems and assign to different population which then solve the sub-problem individually in cooperative fashion. To evaluate the performance of proposed approach, we applied MSMM-CE to

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

challenging problem where the correlation between rows and columns plays important roles such as as EEG. MSMM-CE provided superior performance compared to the state of the art methods such as SMM, SSMM and MSMM that shows the effectiveness and the strong empirical value of utilizing cooperative evolution in MSMM-CE for real-world applications.

Compared to the state-of-art featured selection methods, the **key contributions** of this paper as follows:

- We present a multiclass SMM based on cooperative evolution (MSMM-CE) that takes advantage of cooperative evolution to discompose the problem into sub problems.
- MSMM-CE works by effectively combining hinge loss for model fitting, and additional penalty terms (cooperative and nuclear norm) resulting not only low rank plus sparse but also considering the information form other classes in cooperative fashion.
- We show that from a dimensionality perspective, MSMM-CE exhibit distinctive learning properties.
- Unlike OvO strategy, we have used cooperative evolution to discompose the problem into sub problems that are optimized simultaneously in cooperative fashion.

II. Notations and Preliminaries

In this section, we presents the notations that have been used through this paper. Vector, scalar, and matrix are represented by the lowercase bold letter (e.g. **x**), lowercase letter (e.g. x) and uppercase letter (e.g. X). Consider the matrix $I_p \in \mathbb{R}^{p\times p}$. The singular value decomposition (SVD) of a matrix $X($ such that $X \in \mathbb{R}^{p\times q}$, is denoted as

$$X = U\Sigma V^T$$

where $\Sigma = diag(\sigma_1, \sigma_2 ... \sigma_r, 0...0)$ is the rectangular diagonal matrix, U is the unitary matrix, and V^T is the conjugate transpose of matrix U.

We can represent the nuclear norm of X is

$$||X||_* = \sum_{i=1} r\sigma_i$$

and the Forbenius norm of X as

$$||X|| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q} x_{pq}^2}$$

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 [10], thus we consider the sub-differential of nuclear norm of matrix ($||X||_*$) which can be denoted by $\partial||A||_*$ that is a set of subgradients. For a matrix X of rank r, we can write

$$\partial||A||_* = \{U_X V_X^T + Z : Z \in \mathbb{R}^{p\times q}, \quad U_X^T Z = 0, ZV_X = 0, ||Z||_2 \leq 1 \quad (1)$$

III. The Problem Formulation

In order to make it easier to understand the proposed method, we first provide brief description and formalization of matrix classification problem followed by multiclass classification.

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 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 multi-class 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. The above equation is unbalanced problem due to one-vs-all approach even though each class consist of same number of samples which effects the end results. In order to overcome aforementioned challenge, one-vs-one can be used and voting strategy can be used, thus it requires $\frac{c(c-1)}{2}$ models.

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

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 multi-class classification [12] which can be written as

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

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 multi-class binary SVM and proposed c vectors to simulate one-vs-one binary classifiers [13]

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

such that

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

$$arg \min \frac{1}{2} tr(w^T w) + C \sum \xi_i^j \quad (6)$$

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

As mentioned above, we needed to reshape the matrix into vector which results in losing the correlation among columns or rows in the matrix, however results proved the effectiveness of multi-class SVM. To be benefited from rich structural information hidden in the data, recently support matrix machine has been proposed. By directly transforming the equation 6 for matrix, we get

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

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 reshaping process. 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 equation 6 can be rewritten for matrix classification as

$$argmin \frac{1}{2} tr(W^T W) + C \sum 1 - y_i [tr(W^T X_i) + b]_+ \quad (8)$$

The hinge loss enjoys the large margin principle and also embodies sparseness and robustness, major desirable properties for a good classifier. Motivated by this, Luo et. al. presented sparse matrix machine shown in Eq. 9 [2]. The objective function in Eq. 9 consists of hinge loss plus nuclear norm and Frobenius norm as reguarlizer.

$$argmin \frac{1}{2} tr(W^T W) + \tau ||W||_* + C \sum 1 - y_i [tr(W^T X_i) + b]_+ \quad (9)$$

Recently, Zheng et al. presented a multiclass classifier (objective function shown in Eq. 10) by reducing the slack variables. The objective function consists of hing loss as well as regularization terms that help to extend the margin rescaling loss to support matrix-form data. [14]. It aimed to minimize the regularized loss which maximizes the margins between different categories.

$$argmin \frac{1}{2} tr(W^T W) + \tau ||W||_* + \frac{C}{N}\sum_{i=1}^{n} \xi_i \quad (10)$$

IV. Mulitclass SMM-Cooperative Evolution

Nonetheless, these methods are either are too complex to optimize or mainly for binary class problem. Recent success of evolutionary algorithms integration in to machine learning algorithms explicitly SVM showed its advantage to solve complex optimization problem. Coevolutionary algorithms are able to optimize more population simultaneously by discomposing the problem into different sub problems and assign to different population which then solve the sub-problem individually in cooperative fashion.

In the following text, we introduce the multi-class support matrix machine with the aim of find the support vectors in single step parallelly. The proposed approach is able to, maximize the multi-class the margins in single step, reduce the data the redundancy, and consider the strong correlation between columns and rows in a matrix data. MSMM-EC works in a cooperative fashion by breaking down the problem into sub problems which are then optimized simultaneously with the aim to learn support matrices for each class. The objective function in Eq. 11 is combination of low rank matrices as well as sparse properties aiming to capture the correlations efficiently with each data matrix. The additional cooperative evolution penalty term penalizes the multi-class classification error. Thus, MSMM-CE is able to consider the information from other class and penalizes the support matrices that do not work with each other that results in solving the problem in parallel.

A. Objective Function

Given a k-class ($k \geq 2$) 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 $y_i \in \{1, 2, ...k\}$ is the corresponding class labels. The traditional support matrix classifier ($arg \min \frac{1}{2}||W|| + C\sum_{i=1}^{n}\xi$) are only binary class classification, thus, these methods are infact can not deal with multi-class problems. We devised a novel objective function that adopt the cooperative evolution approach with the aim to solving the the class specific each sub problem in single optimization simultaneously by finding set of support vector of each class in cooperative fashion. In other words, each sub problem manage the optimization of each specific class with additional information from other classes.

To this end, we have the following the objective function corresponding to penalized margin maximization for cth class

$$arg \min_{W_r, \xi^r} \underline{\frac{1}{2} tr(W_k^T W_k) + \tau ||W_k||_* + \frac{C}{N} \sum_{i=1}^{n_k} \xi_i^k} \\ \underline{\underline{+ \frac{1}{n_c} \sum_{i=1}^{n_c} \varrho(X_i^k)}} \quad (11)$$

such that

$$y_i f(X_i) \geq 1 - \xi_i^r, \\ \xi_i^r \geq 0$$

Where $W \in R^{pq}$ represents the regression parameter in the form of matrix. $\varrho(X_i^k)$ is the penalty term imposed on the classes that does not fit together. C is the non-negative parameters to balance regularization as well as loss term, n_k are the number of samples in class k, ξ is the slack variable for cth class.

Notice that there are four terms (Nuclear norm, Forbenius norm, cooperative penalty term and hinge loss) in the objective function in Eq.11. We know that the nuclear norm and Forbenius norm, both satisfy the triangle homogeneity properties, whereas the rest of the terms are linear functions, thus we can say all four terms in objective function are convex, however, non-differentiable and non-smooth. Thus, we are not able to use subgradient of nuclear norm in standard descent approaches as a result, driving the solution is complex. An alternative approach is required to approximate the matrix W, which can be solved by imposing the rank on W. To make derivation easier, we first consider the objective function without cooperative penalty term and added it at later stage. To conclude, rank matrix minimization is NP hard and can be solved as

$$arg \min_{W_r, \xi^r} \frac{1}{2} ||W||_F \quad + \quad \tau ||W_c||_* \quad + \quad \frac{C}{N} \sum_{i=1}^{n_c} \xi_i^k \quad (12)$$

whereas as $W \in \mathbb{R}^{d \times c}$

We can not apply the Nesterov methods and stochastic gradient descent as all terms are non-differential and non-smooth. As we know that the the objective function is convex thus we can break the objective function into sub-problems to optimize. We can rewrite the problem in Eq.12 as

$$arg \min_{W,b} P(W) + Q(S) \quad (13)$$

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

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

$$P(W, b) = ||W||_F^2 + \frac{C}{N} \sum_{i=1}^{n_k} \xi_i^k \quad (14)$$

and

$$Q(S) = ||W||_* \quad (15)$$

where $Q(S)$ is additional penalty function defined on singular value of matrix and $P(W)$ is the hinge loss function obtained from negative likelihood. We can solve the Eq. 13 as

$$L(W, \mathcal{L}, S,) = P(W) + G(S) + \frac{p}{2} ||S - W||_F^2 + \langle \mathcal{L}, (S - W) \rangle \quad (16)$$

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

The objective function in Eq. 10 results in optimal solutions for single class label. However, our target is to consider the information from other classes in cooperative fashion. Thus, we imposed an additional penalty term (cooperative penalty term) into the objective function that takes advantage of cooperative evolution of every other class to consider them in learning of specific class. The cooperative penalty term penalizes the error occur in multiclass classification by punishing those classes that does not fit with it.

$$L(W, S, \mathcal{L}) = P(W) + G(S) + \frac{p}{2} ||S - W||_F^2 + \langle \mathcal{L}, (S - W) \rangle + \\ \frac{q}{n_k} \sum_{i=1}^{n_k} \varrho(X_i^k) \quad (17)$$

where

$$\varrho(X_i^k) \;=\; \begin{cases} Z, & \text{if } Z > 0 \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

and

$$Z = 2 + \sum_{i=1}^{n_K} \mathcal{L}_i^K \langle X_i, X \rangle - \sum_{i=1}^{n_k} \mathcal{L}_i^K \langle X_i, X \rangle \quad (19)$$

where $k \in 1, 2, ...K$ is the the index class with largest activation value and q controls the strength penalty.

B. Theoretical Justification

In this section, we theoretically analyze and illustrate how MSMM-CE possesses some elegant features as compared to conventional support vector machines, conventional elastic net SMM [2] and SSMM [3]. Conventional methods required data to be reshaped into vectors, which results increase in dimensionality as well as loss of structural information exist in EEG signals. MSMM-CE is a combination of nuclear norm, Frobenius norm, class oriented hinge loss and cooperative evolution penalty for information share among classes. We solved the objective function in cooperative fashion which discompose the problem into sub problems. Each problem

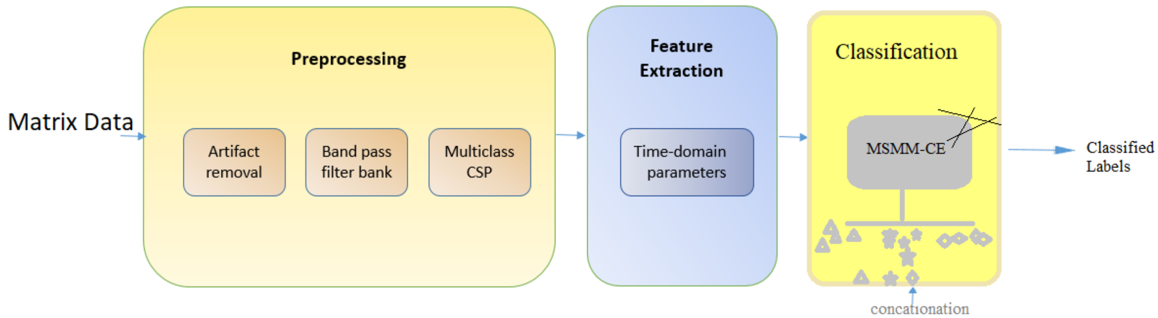

Fig. 1. Illustration of proposed framework equipped with MSMM-CE

is optimized simultaneously to solve multiclass classification problem by penalizing the classes that does not fit with the certain class. Thus, MSMM-CE solve the problem in single optimization by considering the information from the other classes during the learning of support vectors. Each sub problem share the same samples representation. The number of variables for each sub problem depends upon the number of samples in that class, thus number of samples does not increase during optimization unlike other SVM methods.

MSMM-CE is a combination of nuclear norm and Frobenius norm, which enjoys the property of grouping effect (i.e. columns and rows strong correlation). The Frobenius norm helps to prevent the model from overfiting and nuclear norm leveraged to capture the global structure of the matrix.

Notice that, the cooperative penalty term penalizes the objective function for the classes that do not fit together ($Z > 0$). This phenomena helps to maximize the inter class margin for multiclass problem. Each sub problem learn the support vector for specific class based on its own samples as well as considers the information form other classes. This results in parallel optimization for each class. The objective function degenerates to classical support vector machine if $\tau = 0$ and $Z = 0$. This shows that the proposed objective function is a special case of traditional support vector machines.

V. Experimental Evaluation

In this section, we present the detail experiment and evaluate the proposed multiclass support matrix machine. In oder to validate validate the robustness, we have performed extensiv evaluation of proposed MSMM-CE and compared it performance with state of the art methods such as SSMM [3], SMM [2], BSMM [15], MSVM [16], KNN [17] and SCSSP [18] as well as winners of BCI competitions on benchmark EEG datasets (IIIa and IIa). We have performed k(5)-cross validation by randoming partitioning the data to observe the generalization of the results.

A. Results

Our target of this this work is to improve the performance as compared to state of the art matrix based methods followed by an improvement in computational complexity. To show the gain in performance, we have used four evaluation measures and compared the performance of MSMM-CE with state of the art methods on two publicly available EEG datasets. Table I and table III show the evaluation results. For vector based methods, we first transformed the matrix into vectors followed by dimensioanlity reduction using PCA and to form multiclass problem form binary class problem, we have used OvR strategy except MSMM. Furthermore, for better comparison, we have evaluated the performance using error rate in Kappa measure. The evaluation results on data-set IIa are shown in table II, and table IV. Notice that, matrix based classifier achieved better results as compared to those methods based on vectors. This shows that importance of structural information for the classification for EEG classification. In comparison to matrix based methods, MSMM-CE achieved better performance.

We can observe that the proposed approach (objective function in 11) consist of term τ that controls the the number of low-rank regression parameter and manages the penalty. We can observe that the larger the value of τ results in much heavy penalty on the regression coefficient thus setting mostly singular values to zero which in turn result in losing most structural information embedded in data. It shows that dividing the problem into sub problem is easier to optimize. by breaking the objective function into sub-problems that are easier to optimize. We can observe that MSMM-CE converges to the global optimum in only few iterations.

B. Parameter Setting

There are four terms in objective function that are nuclear norm, Forbenius norm, cooperative penalty term (introduce later) and hinge loss function. We can see that there are several parameters (learning rate η, τ,q, p, t and C) are required to be optimized. q determine the penalty on the classess that does to fit together. It helps to maximize the inter-class margin for multiclass classification problem. Smaller the value of q results in much smaller penalty and vise verse. In case of few classes, the cooperative penalty on classes is simple and effect, however, the problem get complicated and affect the results as the number of classes increase as the decision boundary get complex. τ controls the correlation of data matrix. It is

TABLE I

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

Subject	BCI Competation	KNN	MSVM	SCSSP	SMM	BSMM	MSMM	MSMM-CE
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	0.952/4.4
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	0.831/11.9
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	0.884/10.1

TABLE II

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

Sub	BCI Competation	KNN	MSVM	SCSSP	BSMM	SMM	MSMM	MSMM-CE
S1	0.68/24	0.71/22	0.72/21	0.62/26	0.73/21	0.69/0.23	0.73/20	0.75/21
S2	0.42/44	0.4/45	0.37/47	0.28/54	0.4/45	0.23/0.58	0.43/43	0.43/41
S3	0.75/19	0.77/17	0.76/17	0.6/26	0.75/19	0.69/0.24	0.84/11	0 0.84/10.6
S4	0.48/39	0.45/41	0.36/48	0.33/51	0.51/37	0.54/0.35	0.59/31	0.61/30
S5	0.4/45	0.38/47	0.42/43	0.15/64	0.39/46	0.32/0.51	0.5/38	0.52/40
S6	0.27/55	0.24/57	0.19/61	0.25/56	0.32/51	0.15/0.63	0.41/44	0.43/39
S7	0.77/17	0.69/23	0.66/25	0.41/44	0.81/14	0.72/0.21	0.85/12	0.86/11
S8	0.76/18	0.62/29	0.45/41	0.6/31	0.71/22	0.71/0.22	0.77/17	0.78/13.7
S9	0.61//26	0.48//39	0.56/33	0.66/25	0.62/29	0.63/0.27	0.72/21	0.75/14
avg	0.57/32	0.53/36	0.5/37	0.44/42	0.58/31	0.52/0.36	0.65/26	0.69/16

difficult to determine the level of structural information that can provide optimal results. We observe that the larger value of τ impose heavy penalty on the structure information thus setting most of the singular values to zero as a result, we can lose most of the structural information. We can notice that MSMM-CE degenerates to the problem [13] for vector data for the value of $\tau = 0$.

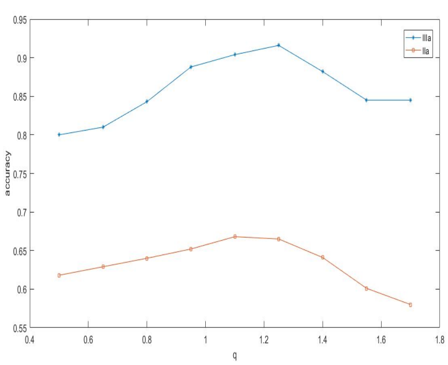

Fig. 2. Behaviour of cooperative penalty term q on on the classication performance for IIa and IIIa datasets

C. Computational Complexity

AS our major objectives in this work is to overcome the computational complexity. The existing multiclass support matrix approaches required $\frac{c(c-1)}{2}$ support matrix thus are computationally expensive. To overcome this complexity, we have decomposed the problem into smaller sub problem and solved them individually.This simultaneous optimization of multiclass problem makes is much faster as compared to OvO and OvA methods. To observe the gain in computational efficiency, we have compared the run time of the algorithms that are based on matrix data only. We conducted the experiments Intel 3.7GHz, 16GB RAM, Window 7. Table V describe the average testing and training time. Notice that MSMM-CE is much faster in both training and testing.

D. Discussion

Notice that, the proposed MSMM-CE showed better results in comparison to state-of-the-art. We can notice from the results that MSMM-CE is better able to find best representative and discriminant patters from high-dimensional data. We can observe that Nuclear norm promotes the structural sparsity as well as shares similar sparsity patterns across multiple predictors. τ controls the structural information in the classcation i.e. it controls 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. MSMM-CE 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. q determine the penalty on the classess that does to fit together. It helps to maximize the inter-class margin for multiclass classification problem. Smaller the value of q results in much smaller penalty and vise verse. In case of few classes, the cooperative penalty on classes is simple and effect, however, the problem get complicated and affect the results as the number of classes increase as the decision boundary get complex.

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

(I) Larger value of q results powerful penalty on the multiclass classification error (the classes that does not fit together). However, too large value of q results affect the the performance due to the fact that high value of q results in penalizing the other classes that leads to biased problem, this could be solved using variable q based on error.

TABLE III
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
MSMM-CE	0.907	0.927	0.918	0.922

TABLE IV
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
MSMM-CE	0.656	0.793	0.766	0.761

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

Classifier	IIIa		IIa	
	Training	Testing	Training	Testing
SMM	21.995	0.0594	64.198	0.243
BSMM	20.381	0.0636	65.198	0.243
MSMM	22.257	0.0541	65.528	0.230
MSMM-CE	19.42	0.0496	62.261	0.20

(III) We noticed that MSMM-CE performed slightly better for imbalance classification problem and in the presence of outliers.

(IV) MSMM-CE similar to MSMM, SMM in term of support vectors and learn the simpler function with better and complex decision boundaries (unlike OvO and OvA, the decision boundaries boundries does not overlap as optimized simultaneously by considering the information from other classes.) thus MSMM-CE is able to learn the support vectors that can be use to measure the complexity of the model.

VI. CONCLUSION

In this work, we presented an novel classifier name Multi-class Support Matrix Machine (MSMM-CE) by decomposing the complex multiclass classification problem into sub problem and solve them individually in cooperative fashion. Unlike the decomposition based multiclass classification approaches, MSMM-CE optimizes the objective function in single model leading to simple decision function as compared to OvO and OVA SVM. We combined the hinge loss, nuclear and Forbenius norm and followed the idea of cooperative evolution in natural fashion by penalizing the classes that does not fit together. Hence, resulted in an improved classification performance supported by the experimental evaluation. Furthermore, it not only leveraged the structural information and avoided the inevitable upper bound for the number of selected features but have the property of low rank. Results showed considerable gain in performance as compared to state-of-the-art classifiers. In conclusion, the numerical results suggested that MSMM-CE is advantageous to previous approaches. It shows the promise of MSMM-CE on real-world applications.

REFERENCES

[1] 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.

[2] 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

[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] I. Razzak, I. A. Hameed, and G. Xu, "Robust sparse representation and multiclass support matrix machines for the classification of motor imagery eeg signals," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, pp. 1–8, 2019.

[5] A. Rosales-Perez, S. Garcia, H. Terashima-Marin, C. A. C. Coello, and F. Herrera, "Mc2esvm: Multiclass classification based on cooperative evolution of support vector machines," *IEEE Computational Intelligence Magazine*, vol. 13, no. 2, pp. 18–29, 2018.

[6] A. Rosales-Pérez, A. E. Gutierrez-Rodríguez, S. García, H. Terashima-Marín, C. A. C. Coello, and F. Herrera, "Cooperative multi-objective evolutionary support vector machines for multiclass problems," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2018, pp. 513–520.

[7] J. Wang, L. Li, D. Niu, and Z. Tan, "An annual load forecasting model based on support vector regression with differential evolution algorithm," *Applied Energy*, vol. 94, pp. 65–70, 2012.

[8] G. Acampora, F. Herrera, G. Tortora, and A. Vitiello, "A multi-objective evolutionary approach to training set selection for support vector machine," *Knowledge-Based Systems*, vol. 147, pp. 94–108, 2018.

[9] F. E. Zulvia and R. Kuo, "Gradient evolution-based support vector machine algorithm for classification," in *IOP Conference Series: Materials Science and Engineering*, vol. 319, no. 1. IOP Publishing, 2018, p. 012062.

[10] M. I. Razzak, M. Blumenstein, and G. Xu, "Multi-class support matrix machines by maximizing the inter-class margin for single trial eeg classification," *IEEE Computational Intelligence Magazine*.

[11] I. Razzak and T. M. Khan, "One-class support tensor machines with boundedhinge loss function for anomaly detection," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[12] Y. Guermeur, "Combining discriminant models with new multi-class svms," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 168–179, 2002.

[13] 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.

[14] 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.

[15] T. Kobayashi and N. Otsu, "Efficient optimization for low-rank integrated bilinear classifiers," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 474–487.

[16] 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.

[17] 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.

[18] 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.