Window Gaussian Kernel Support Vector Machines for Motor Imagery Classification

Dr. Aparna Chaparala[1], Dr. M. Sreelatha [2]
Associate Professor[1], Professor[2]: Dept. of CSE
RVR&JC College of Engineering
Guntur, India

Dr. B. Raveendra Babu
Professor: Dept. of CSE
VNR Vignana Jyothi Institute of Engineering &Technology
Hyderabad, India

Abstract — **Support vector machines (SVMs) are supervised learning methods used in classification and regression analysis, which uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fitting of the data. A Brain Computer Interface (BCI) enables people with muscular disability to directly control machines using their thought process. Electro Encephalography (EEG) is being used widely as input for motor imagery classification by BCI systems. As BCI features are unstable over time, a low Variance may be a key to low BCI classification error. SVM decision rule is a simple linear function in kernel space making SVM stable and with low Variance. Also, SVM's robustness with respect to dimensionality makes it a suitable classier for BCI system whose input is high dimensional. To avoid the over-fitting and under-fitting of the SVM Kernel, the Kernel's width must adapt to the feature space distribution. In this work a method to find the ideal width is proposed.**

Keywords—BCI; SVM; Kernel; Mahalanobis;

I. BCI SYSTEMS

Brain-Computer Interface (BCI) is based on two adaptive controllers, the subject's brain producing activity encoding the thoughts, reflecting the brain's function, and the system that translates this activity into control signals/device commands [3]. Many factors determine BCI system performance and they include brain signals measured, signal processing methods extracting signal features, algorithms which translate features into device commands, output devices executing commands, feedback to user, and user's characteristics.

Most BCI systems are based on a machine learning algorithm, which learns from training data and discriminates varied brain activity patterns. It adapts BCI system to a specific subject's brain, thereby decreasing the subject's learning load. Machine learning algorithms are made up of feature extraction and classification modules. Feature extraction transforms raw brain signals into a representation ensuring easy classification. The goal of feature extraction is detection and removal of noise and other unnecessary information from input signals, while simultaneously retaining important information to differentiate signal classes. Signal processing methods are used to extract feature vectors from brain signals. Neurophysiologic a priori knowledge helps decide which brain signal feature will have most discriminative information for a selected paradigm. Machine learning algorithms translate such features into a control signal. BCI tools/techniques like signal acquisition, feature extraction, signal processing, classification techniques and machine learning algorithms have a part in development/improvement of BCI technology.

Machine learning methods' role is in discriminating EEG patterns representing various brain activity types. Machine learning depends on features extracted and classification algorithms used. Classification is guided by 2 general approaches. The first follows concept of "simple methods first" by using linear classifiers alone. Various studies show the linear classification methods never performed worse than non-linear classifiers in BCI systems [12]. The second approach extends machine learning algorithms functionality by regularizing and combining multiple classifiers. Bayesian Linear Discriminant Analysis (BLDA) is an extension of Fisher Linear Discriminant Analysis (FLDA) [5] and a combination of multiple linear SVM classifiers [4] which have a regularization parameter selection. SVM was used in BCI researches as it is a powerful pattern recognition approach especially for high dimensional problems [4]. EEG Signals are high dimensional with low signal-to-noise ratio.

II. LEARNING AND GENERALIZATION

Initial machine learning algorithms are designed to learn simple function representations. So, the learning goal was to output a hypothesis that correctly classified training data [17]. The ability of hypothesis to classify data correctly other than the training set is known as generalization. SVM performs better as over generalization is avoided whereas neural networks might over generalize easily [18].

Many linear classifiers (hyper planes) separate the data and of which only one achieves maximum separation. The reason why a hyper plane is not used to classify, is it might be closer to specific data points compared to others and this not being the choice, the concept of maximum margin classifier is found to be an apparent solution. Maximum margin is given as

$$\text{margin} \equiv \arg\min_{\mathbf{x}\in D} d(\mathbf{x}) = \arg\min_{\mathbf{x}\in D} \frac{|\mathbf{x}\cdot\mathbf{w}+b|}{\sqrt{\sum_{i=1}^{d} w_i^2}}$$

The maximum margin provides better empirical performance as even for a small error made in the location of the boundary and this gives less chance for causing a misclassification. Main advantage would be circumvented for local minima and better classification.

For calculating the SVM to correctly classify all the data, following mathematical calculations are used:

(a) If $Y_i = +1; \quad wx_i + b \geq 1$

(b) If $Y_i = -1; \quad wx_i + b \leq 1$

(c) For all i; $y_i\left(w_i + b\right) \geq 1$

In the equation, x is a point in the vector space and w is a weight vector. So to separate data (a) should be greater than zero. SVM selects one hyperplane among all possible hyper planes which has the maximum distance. If training data is good and test vectors are located in radius r from training vector then a chosen hyper plane is located at farthest possible location from data [13]. This hyper plane which maximizes margin also bisects on convex hull of two datasets.

The distance from the closest point on the hyperplane to the origin can be got by maximizing x. Also, for the other side points the scenario is same.

$$MaximumM\arg in = M = 2/\|w\|$$

Now maximizing the margin is same as minimum. It is a quadratic optimization problem and need to be solved for w and b. To solve this, the quadratic function needs to be optimized with linear constraints. A solution can be constructed based on a dual problem with a Langlier's multiplier αi being associated. w and b are to be found such that

$$\Phi\left(w\right) = ½\ |w'||w|$$

is minimized; And for all

$$\left\{\left(x_i,\ y_i\right)\right\} : y_i\left(w\ *\ x_i + b\right) \geq 1$$

The solution arrived at is

$$w = \Sigma\alpha_{i\,*}x; b = y_k - w\ *\ x_k$$

for any x_k such that $\alpha_k \neq 0$

The classifying function will have the following form:

$$f\left(x\right)\ = \Sigma\alpha_i y_i x_i\ * x\ + b$$

First the problem with optimization is converted to the dual form in which w is removed, and a Lagrangian is only a function of λ_i. To solve the problem the L_D should be maximized with respect to λ_i. Dual form simplifies optimization and a major achievement is dot product obtained from this.

The kernel trick just chooses a suitable function corresponding to a dot product of nonlinear mapping. A particular kernel is chosen by trial and error on test set; choosing a problem based right kernel or application enhances SVM's performance.

III. SVM Classifiers for BCI

SVM gave really good results in several synchronous experiments [6], should it be in its linear [16] or nonlinear form [10], [14], in binary [15] or multiclass BCI [14]. RFLDA and SVM share similar properties like being a linear and regularized classifier; also, their training algorithms are similar. Consequently, it also gives very interesting results in some experiments [16], [8]. The first reason for this success may be regularization. BCI features are noisy and contain outliers [8]. Regularization overcomes this and increases classifier's generalization capabilities. Hence, regularized classifiers, particularly linear SVM, have outperformed un-regularized classifiers of the same kind, i.e., LDA, during several BCI studies [8]. Similarly a nonlinear SVM outperformed an un-regularized nonlinear classifier, MLP, in another BCI study [14].

The next reason is SVM's simplicity. Indeed, SVM decision rule is a simple linear function in kernel space making SVM stable and with low Variance. As BCI features are unstable over time, a low Variance may be a key to low BCI classification error. The last reason is SVM's robustness with respect to dimensionality. This enables SVM to get very good results with even high dimensional feature vectors and small training set [4]. But, SVMs are not drawback free for BCI as they are slower than other classifiers, but fast enough for real-time BCI, e.g., [7].

IV. Proposed Methodology

The following flowchart in figure 1 depicts the proposed methodology.

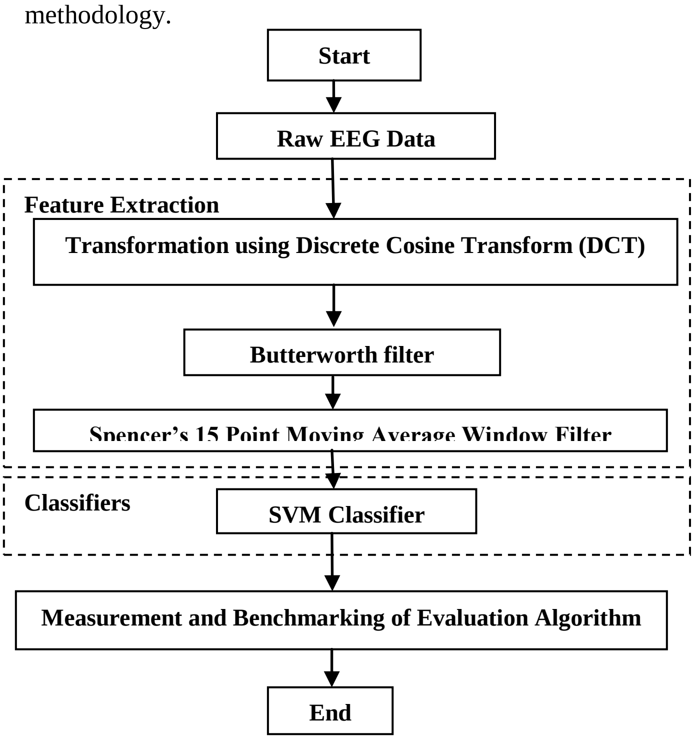
Start
Raw EEG Data
Feature Extraction
Transformation using Discrete Cosine Transform (DCT)
Butterworth filter
Spencer's 15 Point Moving Average Window Filter
Classifiers
SVM Classifier
Measurement and Benchmarking of Evaluation Algorithm
End

Fig 1. Flowchart for Proposed Methodology

A. Feature Extraction

For correct classification of a given BCI system, the features extracted from the signal are crucial. The accuracy depends on the properties of the features and how they are used. Amplitude values of EEG signals [10], Band Powers (BP) [19], and Power Spectral Density (PSD) values [9], [11] are used as features. EEG data is transformed into the frequency domain using Discrete Cosine Transform. Butterworth filter is used to remove unwanted frequencies and noise. The efficiency of the BCI depends upon the methods used to process the brain signals and classify various patterns of brain signal accurately to perform different tasks. In our earlier investigations, pre-processing of the EEG signals for efficient feature extraction [2] and classification of using the Semi Partial Recurrent Neural Network are studied [1]

B. Classifier

SVMs use inner product as a metric to measure the similarity or distance between patterns. The dependent relation between the pattern's attribute is mapped as

$$\langle \Phi(x), \Phi(y) \rangle$$

for the pattern x and y or this is represented as the kernel function:

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle$$

Gaussian RBF kernel is very commonly used and is formulated as

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2} \right)$$

To avoid the over-fitting and under-fitting of the SVM Kernel, the Kernel's width must adapt to the feature space distribution. In this work a method to find the ideal width is proposed. In dense areas, the width is narrowed and the weight assigned is less than 1 whereas in sparse areas, the width is increased and the weight assigned is more than 1. The relationships are given as follows:

a. The relation between σ and λ

$$k(x,y) = \exp\left(-\lambda \|x-y\|^2 \right)$$

b. Relation between similarity and distance

$$d^2 = \|\Phi(x) - \Phi(y)\|^2 = 2 - 2k(x,y)$$

c. Relation between dense vs. sparse in feature space - pattern x drops in dense area, the x's closest members are calculated using Mahalanobis distance formula and the values are obtained:

$$sim\_Mab(x) = \frac{1}{k}\sum_i k(x,x_i), \quad \mathrm{x}_i \in k - Mab$$

And also *sim_Mab(x)* represents the index of density of x's neighborhood.

Training of SVMs is done by finding α_i, expressed as minimizing a dual quadratic form:

$$\min_\alpha \psi_{(\alpha)} = \min_\alpha \tfrac{1}{2} \sum_i \sum_j y_i y_j K(x_i, x_j) \alpha_i \alpha_j - \sum_i \alpha_i$$

$0 \le \alpha_i \le C$ and linear equality constraint $\sum_i y_i \alpha_i = 0$.

The α_i are the Lagrange multipliers.

V. Results and Discussion

In this work Data set provided by University of Tübingen, Germany, Dept. of Computer Engineering and Institute of Medical Psychology and Behavioral Neurobiology, and Max-Planck-Institute for Biological Cybernetics, Tübingen, Germany, and Universität Bonn, Germany, Dept. of Epileptology is used as the first data set. A subject imagines movements of either the left small finger or the tongue during the experiments and 8x8 electrodes attached to the contralateral motor cortex that record the signals. All readings were taken on a sampling rate of 1000Hz and after amplification were stored as microvolt values. Each trail was recorded for 3 second duration. The recordings were started only after 0.5 seconds from the visual cue end to avoid visually evoked potentials. 168 instances of a single patient are used to validate the proposed algorithm. 80% of the data is used for training and the remaining for testing. After extracting the features, the obtained features are used as classification attributes for Naïve Bayes, IBL and SVM. The confusion matrices for different classiers of the investigation are given in tables 1 – 4. The classification accuracy obtained for various classifiers considered in the study is tabulated in Table 5 and shown in Figure 2. classification and misclassification percentages are shown in Figures 3 and 4 respectively.

Table 1 – Confusion Matrix for Naïve Bayesian Classifier

Naïve Bayesian	**Predicted**		
Actual		***a***	***b***
	a	84	15
	b	11	58

Table 2 – Confusion Matrix for IBL Classifier

	Predicted		
Actual		***a***	***b***
	a	89	8
	b	11	60

Table 3 – Confusion Matrix for SVM Classifier

	Predicted		
Actual		a	b
	a	87	8
	b	10	63

Table 4 – Confusion Matrix for WGKSVM Classifier

	Predicted		
Actual		***a***	***b***
	a	88	6
	b	10	64

Table 5 - Classification accuracy

Technique	Classification Accuracy
Naïve Bayes	84.06
IBL	88.41
SVM	89.21
Proposed WGK-SVM	90.48

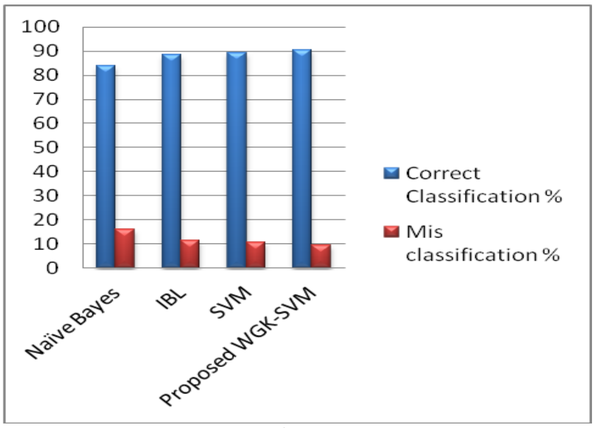

Fig 2. Classification accuracy

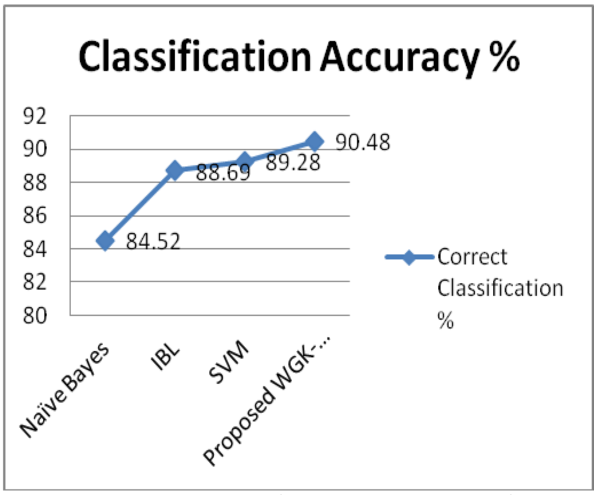

Fig 3. Graph representing classification accuracy obtained for various classifiers of the study

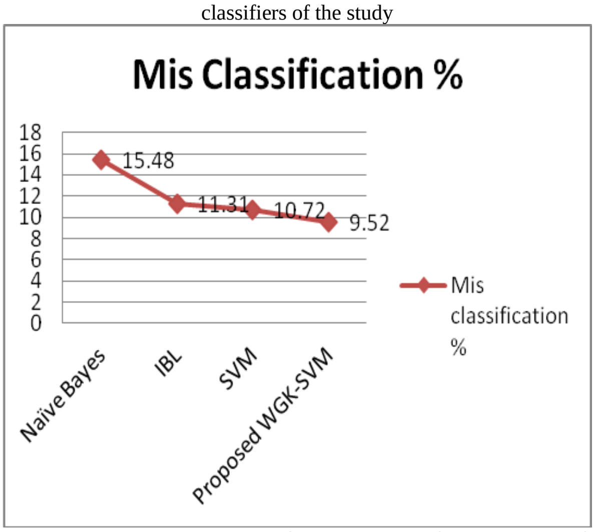

Fig 4. Graph representing misclassification obtained for various classifiers

From Figure 2, it is observed that the classification accuracy of the proposed system improves by 2.07%. The SVM are more tolerant to irrelevant attributes, redundant attributes leading to better classification of the EEG instances. For large real-world data sets, the SVM yields statistically significantly better results than Naïve Bayes and IBL.

Though the results are satisfactory, further investigation is required to improve the classification accuracy and efficiency.

VI. Conclusion

To avoid the over-fitting and under-fitting of the SVM Kernel, the Kernel's width must adapt to the feature space distribution. In this work, Window Gaussian Kernel Support Vector Machines (WGK-SVM), a method to find the ideal width is proposed. In dense areas, the width is narrowed and the weight assigned is less than 1 whereas in sparse areas, the width is increased and the weight assigned is more than 1. EEG data in time domain is transformed to frequency domain using discrete cosine transform. The frequency of interest is extracted using Butterworth band pass filter. Artifacts are removed using Spencer's 15 point window. Experimental results showed that the proposed WGK-SVM achieves satisfactory results. Further investigation needs to be carried out to improve classification accuracy.

References

[1] Aparna, C.,Murthy, J. V. R., Babu, B. R., & Rao, M. C. S., "A Novel Neural Network Classifier for Brain Computer Interface", Computer Engineering and Intelligent Systems, 3(3), 10-16, 2012

[2] Aparna, C., Murthy, J. V. R., & Babu, B. R., "Energy Computation For BCI Using DCT and Moving Average Window for Noise Smoothening", International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.1, February 2012, PP:15-21.

[3] Selim, M. Abdel Wahed and Y. Kadah, "Machine Learning Methodologies in Brain-Computer Interface Systems," CIBEC 2008 Biomed. Eng., pp. 1-5, Dec 2008.

[4] Rakotomamonjy and V. Guigue. "BCI Competition III: Dataset II- Ensemble of SVMs for BCI P300 speller," IEEE Trans. Biomed. Eng., vol. 55, no. 3, pp. 1147-1154, March 2008.(Pubitemid351301245)

[5] Hoffmann U, Vesin JM, Ebrahimi T, Diserens K, "An efficient P300-based brain-computer interface for disabled subjects". J Neurosci Methods 167:115–125, 2007

[6] Lotte Fabien, Lécuyer Anatole, Lamarche Fabrice, and Arnaldi Bruno : "Studying the Use of Fuzzy Inference Systems for Motor Imagery Classification", IEEE Transactions On Neural Systems And Rehabilitation Engineering, Vol. 15, NO. 2, JUNE 2007.

[7] Thulasidas, M., Guan, C., Wu, J., 2006. "Robust classification of EEG signal for brain-computer interface"., IEEE Trans. Neural Syst. Rehab. Eng.14 (1), 24-29.

[8] Muller K R , Dornhege G, Blankertz B, and Curio G: "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms", IEEE Trans. Biomed. Eng. 51, 2004, pp :993–1002.

[9] S. Chiappa and S. Bengio. : "Hmm and iohmm modeling of eeg rhythms for asynchronous bci systems", In European Symposium on Artificial Neural Networks ESANN, 2004.

[10] Kaper M., P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter : "BCI competition 2003- data set iib: support vector machines for the p300 speller paradigm", IEEE Transactions on Biomedical Engeneering, 2004, 51(6):1073-1076.

[11] Millan J R : "The need for on-line learning in brain–computer interfaces" Proc. Annual Int. Joint Conf. on Neural Networks (Budapest, Hungary), 2004.

[12] V. Franc, V. Hlavac. Statistical Pattern Recognition Toolbox for Matlab, User's guide. June 24, 2004,

[13] J.P.Lewis, Tutorial on SVM, CGIT Lab, USC, 2004.

[14] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut, "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification," IEEE Trans. Neural. Syst. Eng. , vol. 11, pp. 141–144, June, 2003

[15] Garcia GN, Ebrahimi T, Vesin JM. "Support vector EEG classification in the Fourier and time–frequency correlation domains", In: Proc. first international IEEE EMBS conference on neural engineering; 2003. p.p591–4.

[16] Blankertz B, Curio G, and Muller K R : "Classifying single trial EEG: Towards brain–computer interfacing", Advances in Neural Information Processing Systems, 2001, pp 157–64.

[17] Cristianini N. and Shawe-Taylor J. 2000. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK.

[18] Mitchell, T.M. Machine Learning 1997, New York, NY, USA McGraw-Hill

[19] Pfurtscheller G. and Aranibar A. : "Event-related cortical desynchronization detected by power measurements of scalp EEG", Electroencephalography and Clinical Neurophysiology, Vol. 42, Iss. 6, 1997, pp. 817-826.