Journal of Advance Computational Research
Vol. 3 No. 1 (January-June, 2018)

A Quality Metric for Wrapper Methods in Multiclass Subject Invariant Brain Computer Interfaces

Vikas Gottemukkula, Jesse Sherwood and Reza Derakhshani
School of Computing and Engineering, University of Missouri, Kansas City

RECEIVED: October 26,2017. Revised February 27, 2018

Abstract: Using error rate as the scalar metric for the evaluation of one versus rest (OVR) classifier is a major challenge in data-driven design of multi class brain computer interfaces. With unbalanced datasets, OVR classifiers require an accurate measure of the performance that considers multiple quality metrics, such as sensitivity and specificity, in addition to overall correct rates. With a 4-class BCI, the typical correct rates of 60-80% could be misleading as they are in the vicinity of 75% recognition rate of trivial OVR classifier. By devising a scalar quality factor (Q), calculated from correct classification rates, sensitivity, and specificity, we mitigated this degeneracy for a 4-class subject-independent brain computer interface implemented by four SVM or naïve Bayesian OVR classifiers and wrapper feature selection. Using the Q factor we fitted a single model to the motor cortex EEGs of 10 untrained subjects. The average cross-validation correct rate, sensitivity, and specificity of the resulting OVRs were as high as 83.9%, 78.5% and 80.8%, respectively. Within the confines of our experiment, we conclude that sensitivity and specificity-corrected accuracies, when used as the guide in wrapper methods, are able to avoid trivial classification in multiclass subject-independent BCIs.

Keywords: Brain Computer Interface, Human Movements, Support Vector Machines, k-Nearest Neighbor Classifiers, Bayes Classifiers, Feature Extraction, Pattern Recognition, Wrapper Methods, One vs. Rest Classifiers.

1. INTRODUCTION

Brain Computer Interface (BCI) is a rather nascent research topic dating back to the early 1970's [37]. BCI involves classification of various brain signals to communicate with external devices. BCI can serve as an alternative or augmentative communication pathway by bypassing possibly non-functional motor subsystems [19]. Brain activity can be monitored using a variety of methods such as electroencephalography (EEG), magneto-encephalography (MEG), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI). However, EEG has been the modality of choice for brain-state detection due its affordability, ease of usage, and portability [20]. EEG signals are mostly composed of small scalp-induced potentials from the electrical activity of the cortex. It is hypothesized that some movements and thoughts might have unique EEG signatures across different areas of the brain [7]. However, these signatures have high variance between different subjects [3]. Additionally, the classifier aspect of BCI faces challenges as the number of states increases, and thus the majority of studies attempt to discriminate between 2 or 3 classes.

Correspondence email address: **vgq57@mail.umkc.edu**

As any other pattern-recognition system, BCI requires an effective evaluation criterion that is easy to compute. When implementing multi-class BCIs with two-class classifiers, a system of one vs. one or one vs. the rest classifiers (OVR) may be used. In a one versus one method for an N-class problem, all possible pairs of classes are applied to dichotomizing classifiers. This requires N(N-1)/2 classifiers which grows rapidly with N, number of target mental states. On the other hand, one can solve for the same problem using N OVRs. However, simple error-based evaluation of OVR classification may lead to degeneracy in this case, where an N-class OVR classifier system with equal number of instances from each class may show an (N-1)/N correct rate while failing to properly detect a single instance (trivial classification). This problem becomes more pronounced as (a) the number of classes, N, increases and the probability of chance classification, 1/N for an evenly distributed dataset, becomes smaller, (b) when feature selection and classification are data-driven, where using the aforementioned simple correct rate as selection feedback signal may trap the system in the degenerate trivial solution, and (c) when the OVR classifiers have discrete outputs, and thus ROC-based metrics that compensate for class-label frequencies cannot be readily applied.

A summary of recent approaches to remedy this issue has been presented in the literature [2]. Some of the more promising methods such as Cohen's kappa coefficient [1], information transfer rate [6, 21] and utility metric [5] may yield desirable results. However, they have not been applied to the problem of degeneracy in classification quality feedback during wrapper processes, not to mention their shortcomings when dealing with the aforementioned issue [5]. Although many contributions were made to the field of multi-class BCI [8], very few attempts have been made to solve subject-invariant BCI (SIBCI) problems [19], let alone multi-class SIBCIs [13, 20, 30], where most of the offered solutions need re-calibration for each subject. Classification-guided feature selection (wrapper) methods have been reported in BCI designs [26, 31, 32]. However, they utilize simple error rate as classification feedback and the issue of imbalance between sensitivity and specificity during degenerate classification has not been systematically addressed. Ultimately, the objective of subject invariant BCI is to create a system that has minimal calibration or training time for a new user. This is accomplished by training a single system in advance with data from a diverse user population, in order to find the most common features which represent the overall group [14, 25]. Here we propose a new a scalar quality factor (section 2.3) which considers the combined effects of the confusion matrix and cost functions as a simple, efficient, and arguably better alternative to guide classification-guided feature selection and classification, also known as wrapper methods [16]. By optimizing multiple classification objectives, our quality factor will allow for a non-degenerate wrapper method solution regardless of the number of classes and frequency of their instances. As a challenging test bed which usually leads to the trivial classifiers with wrapper methods, we apply the proposed method to a subject-invariant BCI (SIBCI) problem, noting that the majority of current BCI systems offer subject-specific solutions given the complexity of SIBCI realization. As a proof of concept, we recruited 10 randomly selected subjects and adapted one BCI to the whole subject population with equal number of signal epochs per subject, as detailed below.

2. METHODS

In this study, a four-class BCI system was developed using EEG signals garnered from motor cortex areas of 10 untrained volunteers attempting four body movements [20, 21]. The target movements are left hand, right hand, left leg, and right leg; and designated as Movement 1, Movement 2, Movement 3 and Movement 4, respectively. Our choice of the four motor movements was based on the literature and our own previous study [12, 28, 39, 43]. The recorded signals thereafter undergo preprocessing, feature extraction, feature selection, and classification (Fig. 1); with the latter two being realized by an integrated supervised learning scheme (wrapper method). Furthermore, to increase the statistical power of the results and gauge the generalization power of each classifier, we performed a 5-fold cross validation with 10 Monte Carlo repetitions for each OVR classification.

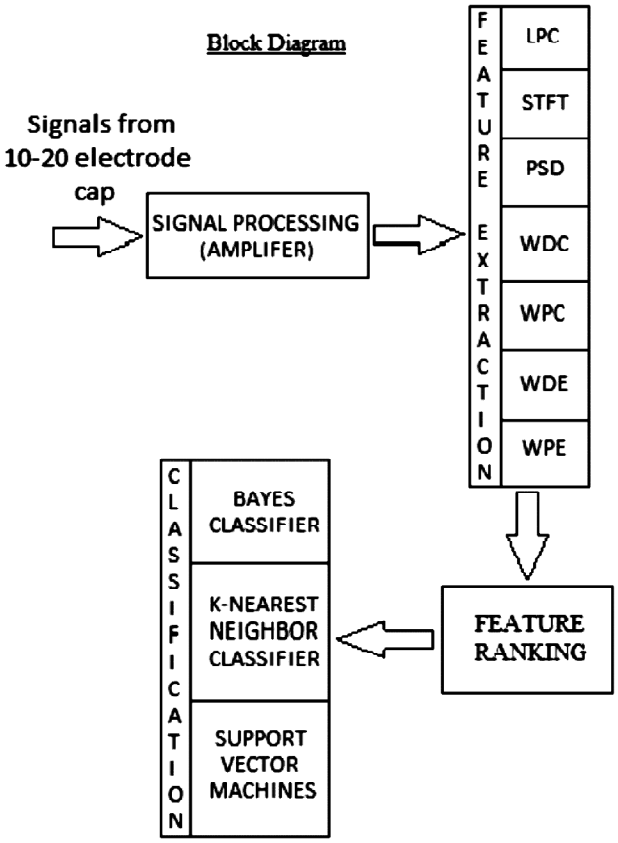

Figure 1: A Block Diagram of the Overall Method

2.1. Protocol

We recorded EEG signals from 10 untrained volunteers (UMKC IRB #090218) using standard 10-20 electrode placement with two ear references (Fig 2). The microvolt-level scalp-collected EEG signals are amplified and filtered to reject power line interference. This was accomplished with a NeuroPulse Systems MS-24R bioamplifier with 1.5 – 34 Hz bandpass filter and a sampling rate of 256 Hz.

The subjects were asked to perform four subtle movements. No preliminary training sessions were provided. Each of the four intended movements was repeated for 24 times, producing 960 eight-second epochs. Each subject attended two similar sessions conducted 4-6 weeks apart. During each session they were asked to perform 4 intended movements 12 times each over the course of 2 hours. All the instructions to the subjects were given using a computer screen prompt, while asking them not to blink during the task performance. They were asked to position their both feet on a foam pad, loosely gripped rubber balls with each hand, and apply slight pressure based on the displayed instruction (pseudo movements). A short audio tone preceded each displayed command prompt, which stayed on the screen for the eight second recording duration. Thereafter, subjects were given a 10 second break before the next pseudo movement task. Although each task was recorded for eight seconds, only the first two seconds was used for the analysis. C3 and C4 electrodes captured the motor cortex signal (Fig. 2). For better spatial resolution, large Laplacian filters were applied, where the reference potentials were derived from the difference of C3 and the average of F3, T3, P3, and Cz (left hemisphere). Similarly, C4 was referenced to the average of F4, T4, P4, and Cz (right hemisphere). Left and right hemisphere signals were filtered separately.

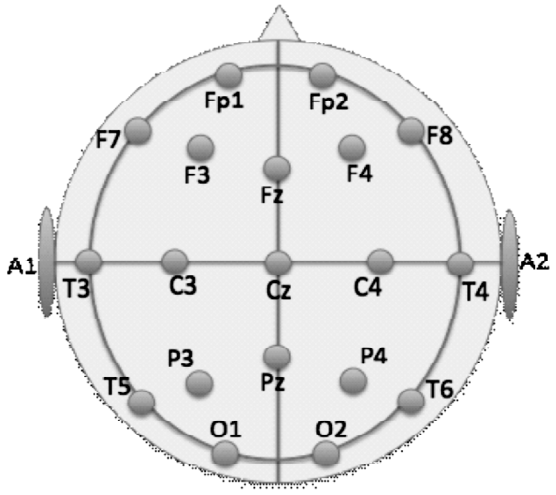

Figure 2: 10-20 EEG Electrode Placement System

2.2. Feature Extraction

After filtering and separating the data epochs by subject, movement and session, each two-second signal segment was processed to extract the following five feature modalities.

Linear Predictive Coefficients

Linear predictive coding (LPC) filter coefficients are popular in speech recognition [24], and have also been successfully applied to time domain feature extraction in BCI [10]. Our studies show that increasing LPC orders initially improves the classification, followed by a plateau and then increased cross validation error due to over-parameterization when filter order reaches the vicinity of 20 [4, 40]. Accordingly, after a three-fold down sampling, 18th order LPC features of each two-second EEG snippets were extracted.

Short Time Fourier Transforms

Given their non-stationarity, EEG signals maybe better characterized by their time-frequency features, and in this case by using the short time Fourier transforms (STFT) [41]. As the second feature modality, EEG STFTs with sliding window length of 1 second, 255 sample overlap, and a spectral range of 1-48 Hz were aggregated into four 12 Hz frequency bins and used as feature vectors.

Power Spectral Density

EEGs representing imagined motor tasks are attributed to event-related desynchronization of neuronal signals. This is reflected in power spectral energy (PSD) of the EEG signals, and thus their choice as the study's spectral features [29]. The Welch periodogram spectral estimation method with a Hamming window of length 33 and 97% overlap was used for the calculation of frequency bin energies. The 1-48 Hz PSD span was divided into twelve 4 Hz frequency bins, and then the spectral energy was calculated as the area under each non-overlapping bins. Settings were chosen by trial and error.

Wavelets

As a multi-resolution time-frequency signal transform, Wavelet decomposition coefficients from filter banks (WDC) were used as spectro-temporal EEG features. Wavelet packet decomposition (WPC) was used as another feature extraction method [36]. Though redundant for signal reconstruction, features from the Wavelet packets may be better for classification, as they retain subdivision of high frequency details into sub-bands which might be advantageous in extracting high frequency features [27, 45].

Based on their classification performance given our previous study [18], from biorthogonal spline,

Gaussian, Morlet, Daubechies, Meyer, reverse biorthogonal spline, Coiflet and symlet wavelet families and their variants, we considered reverse biorthogonal and symlet families for this modality. Wavelet features were constructed using two methods: as the filter banks outputs or aggregation of sub-band energies, where the latter foregoes temporal information in the interest of time shift-invariance by marginalizing over the shift parameter [36]. The wavelet decomposition filter bank output coefficients, or WDC, were calculated by reverse biorthogonal 3.7 wavelets. Wavelet packet filter banks output coefficients, or WPC, were calculated with symlet 15 wavelets. The energy of wavelet decompositions were next marginalized over shift (time) and their energies across different scales (frequency) were calculated as WDE feature sets using reverse biorthogonal 3.1 wavelets. Wavelet packets energy features, or WPE, were calculated in a similar fashion.

2.3. Classifier Quality Factor

As intimated during the introduction, the classification rate by itself may not successfully measure the performance of an unbalanced OVR model, especially when sensitivity and specificity rates are lopsided, leading to degeneracy and trivial classification despite seemingly high classification rates. This is an OVR classification caveat, especially in conjunction with wrapper methods (section 2.4) for data-driven co-evolution of feature selection and classification using challenging datasets. For instance, our experience shows that simple error rate guided wrapper methods, when attempted for a 4-class subject-independent BCI (or SIBCI [17], where one classifier is used to fit a group of subjects), tends to provide solutions with accuracy figures around 75%. Upon closer inspection, it is found that each classifier has a sensitivity or specificity of either almost zero or one (trivial classifier), indicating a premature convergence of the wrapper method to a degenerate solution, where each classifier keeps accepting or rejecting all its incoming data points.

More specifically, in designing classifiers for M-class BCI, we have investigated a type of the OVR classifier that uses M individual classifiers wherein each compares two groups. We place all of the samples from one class in group 1, and the rest from the remaining M-1 classes in group 2. Assuming an equal number of samples from each of the M classes, our model has M-1 times more samples in group 2 than in group 1. With this unequal frequency between the two groups, the accuracy calculation based on counting of true positives and true negatives is biased. Knowing that the theoretical degenerate accuracy rate for a random classifier is 1/M, yet we see that an accuracy calculation based only on true positives and true negatives will produce (M-1)/M accuracy rate as one of the two possible degenerate trivial classifier modes in this particular OVR implementation[1]. This effect becomes more critical as we implement multi-class SIBCI, as this trivial classifier solution asymptotically approaches 100% as the number of classes, M, increases. In data-driven integrated feature-classifier selection scenarios, such as wrapper methods, a poorly functioning classifier at an apparently biased (M-1)/M correct rate would be chosen over a working classifier with a lesser but entirely acceptable accuracy. As M increases, the trivial classifier would refuse a larger range of classifiers otherwise showing less than (M-1)/M performance. For the problem at hand, M=4, and the corresponding degenerate mode settles at 75%. This trivial classification rate is significant and critical when we realize that for a given information transfer rate the accuracy rate of a given BCI decreases for an increasing number of classes [8]. At a constant bit per trial rate, a two-class accuracy of 98% corresponds to 77% for 4 classes, *which occurs very close to the degenerate rate*. Similarly, 66% overall accuracy corresponds to six classes while the corresponding degenerate mode occurs at 83% correct rate. Either of the two degenerate modes can be detected by inspection of the number of false positives and false negatives, leading to the idea and definition of the later described Q-factor.

In view of this, a better characterization of the multi-class BCI would be through its confusion matrix, where all the four descriptors of classification are given. The main drawback of the confusion matrix for wrapper methods utilizing discrete-output classifiers is the need for a single scalar statistic as the classifier feedback to its feature search and selection routine, describing multiple objectives based on the main diagonal of the confusion matrix (true positives and negatives) as well as magnitudes and the ratios of the off-diagonal values (false positives and negatives).

Little has been done to directly resolve this multi-class BCI problem. In fact, most solutions try to avoid it by using other classifier arrangements. One method

to address the issue was demonstrated in the BCI 2005 Competition through the use of Cohen's Kappa coefficient [1, 2]. It is a scalar value derived from the confusion matrix and measures the correlation between predicted and actual classes. The Cohen k statistic falls in the category of chance-corrected agreement statistics. One of the deficiencies of the Cohen's method was demonstrated by Gwet [22]. While Cohen's method was devised for the chance correction problem, it is used to address the degenerate (M-1)/M trivial classification rate. Cohen's k statistic has been shown to be biased by the overall trait prevalence rate, that is the occurrence or deficiency of the true and false positives. Unfortunately, this trait prevalence is significant in the trivial classification mode mentioned above, and manifests itself as some varying degree of lopsidedness in the off-diagonal values of the confusion matrix through its direct influence on the false positive count while at the same time having no impact on the false negative value.

Our approach addresses the problem by providing a scalar metric derived from the confusion matrix. It reflects the accuracy count of the main diagonal and addresses the trivial (M-1)/M classification rate problem while mitigating the trait prevalence issue or any other causes of off-diagonal imbalances by penalizing the lopsidedness between the off-diagonal values, whether it occurs due to a higher value of false positives or false negatives. Additionally, the magnitudes of the off-diagonal terms are accounted for by using the sensitivity and specificity of the confusion matrix to calculate the penalty term, with a provision for emphasizing or diminishing the sensitivity of the metric to the lopsidedness of the sensitivity vs. specificity.

This improved classification quality metric is henceforth referred to as Quality factor, Q. More specifically, one may define Q as the ratio of correct rate to the sensitivity vs. specificity or its inverse, whichever greater:

$$Q = \frac{1 - overall\ error}{\left(\max\left(\frac{sensitivity}{specificity}, \frac{specificity}{sensitivity} \right) \right)^{p}}$$

where the optional parameter p adjusts the sensitivity of Q to asymmetry between sensitivity and specificity, set to 1 for this study. Q improves upon the simple classification accuracy by making adjustments based on the imbalance between true positive and true negative rates. In the case where the sensitivity and specificity are equal, Q defaults back to the overall accuracy rate. This factor is especially valuable when using classifiers with discrete output, such as support vector machines, where metrics such as receiver operating curves are not readily applicable. To further increase the saliency of Q factor, it was calculated over validation data using a 5-fold cross-validation, and averaged over ten Monte Carlo reshufflings of the dataset. This process adds to the statistical power of the results while incorporating performance of the classifier over unseen validation data (generalization) into Q and thus the wrapper method.

2.4. Classification-guided Feature Ranking and Selection

In data-driven pattern recognition, supervised multivariate feature construction is a subset search in the feature space to achieve the desired classification by maximizing a quality metric [4, 16], essentially an optimization process [4]. Ideally, given D feature components, 2^D-1 feature vectors need to be evaluated to rule out redundant, irrelevant, or otherwise corrupted or detrimentally correlated feature components. However, an exhaustive search is not practical choice for many problems with large D [4, 16], such as the problem at hand with the plethora of feature elements discussed in 2.2. Also known as wrapper methods [33], a group of non-exhaustive solutions are based on supervised searches in the feature space that are guided by the subsequent classification rates. Thus wrapper methods incorporate the embedded classifier's capabilities and biases into their feature selection, yielding superior performance by constructing features vectors that are matched to the utilized classifier [33, 38]. One such method is based on classification-guided feature ranking and concatenation. An incrementally augmenting wrapper method builds feature vectors by starting from the top of a ranked feature list, concatenating elements in order until the classification metric of choice is optimized over a predefined span of D, essentially a variant of the best-first wrapper [34, 38]. This semi-greedy wrapper method assumes independence between feature components in the interest of speed [38]. Considering each continuous-valued feature element as a univariate dichotomizing discriminant, one may obtain a plot of sensitivity versus 1-specificity (false positive rate) by varying the binary decision threshold. The resulting curve is known as

the Receiver Operating Characteristic (ROC) curve, which is an important tool in characterization of OVR classification whenever indicator functions with continuous outputs are available [42]. The area under ROC curve, or ROC AUC, is a scalar descriptor of any single feature's overall classification power across all the different decision thresholds. ROC AUC is especially important for dealing with unknown or multimodal class distributions, providing a distinct advantage over traditional methods such as t-test that require a priori knowledge of feature probability distribution. Accordingly, and in preparation for the above-mentioned wrapper method, the features within each modality were ranked using ROC AUC, where a higher area under the curve is indicative of better overall sensitivity and specificity [42]. Using a ranked list for each modality, features were aggregated from the top until the Q factor was maximized. Since the ranking was performed using data from all the 10 subjects, the results yield subject-independent feature saliencies within that group.

2.5. Classification Methods

k-Nearest Neighbor Classifier

The k-Nearest Neighbor classifier (kNN) classifies an unknown sample based on the majority vote of its k nearest neighbors according to their class labels [44]. Feature space distance was calculated by Euclidean metric. The value of k needs to be pre-determined for better classification. In this study, k was varied from 1 to 100. kNNs with k values between 16 and 47 provided relatively better results, though not as good as the next two classification methods.

Bayesian Classifier

Naïve Bayesian classifiers assume independence of variables and find the parameters for class distributions such as means and the covariance matrices from the training data through maximum likelihood method [35]. We used a Bayesian classifier with linear discriminant function assuming normal multivariate distributions with diagonal covariance estimate pooled across the classes.

Support Vector Machines

Support Vector Machines (SVM) separate classes in their kernel space using maximum margin hyper-planes [11]. Gaussian and Polynomial kernels were used in this analysis. Gaussian kernels are characterized by their spread, σ, whereas the polynomial kernels are defined by their order, n. The tested range was 1 to 25 for σ, and 2 to 5 for n. The box constraint (or C parameter) values were changed from 0.1 to 100 to control sensitivity of SVM boundaries to outliers using "soft" or "hard" margins, allowing a trade-off between the slack variables, misclassification penalty, and the discriminant rigidity [9].

3. RESULTS

Using the earlier described best-first wrapper method, ROC-AUC ranked features from each modality were provided to three classification methods, namely Naïve Bayesian, kNN, and SVM. For kNN (k=1-100) and SVMs (Polynomial with n = 2-5, and Gaussian with σ =1-25; and C=0.1-100 in both cases) all the different variations of each classifier were examined. The input vectors were constructed by incremental concatenation from the top of each ranked feature list until the Q factor was maximized. All results were calculated using 5-fold cross-validation with 10 Monte Carlo repetitions to better gauge the predictive power of the results (generalization).

3.1. Movement 1

Using the wrapper method, the Naïve Bayesian and OVRs were able to classify Movement 1 verses rest of the movements with a Q of approximately 72%, followed by SVM, using WPC feature modality (Table 1). The SVM box constraints (C) and σ values were chosen based on the best-attained Q factor. Generally speaking, regardless of σ, Q factor decreased with C, indicating a preference for soft margins that are less sensitive towards outliers. This is expected given the low signal to noise ratio of EEG-based SIBCI dataset. On the other hand, Q factor initially increased with ó, followed by an abrupt decline for >15. Note that although the third ranked classifier has a better correct rate than the first, its Q factor is lower (Table 1). Such Q-factor selection of the features and classifiers enables the wrapper algorithm to escape trivial classification and degenerate states related to large imbalance between sensitivity and specificity not reflected in correct-rate.

3.2. Movement 2

For movement 2, the Gaussian SVM OVRs outperformed other classifiers, followed by naïve Bayesians (Table 2). On the bottom of the list were

kNNs, with a maximum Q of 47% (WPC features, k value of 18). Again, smaller C values (0.1-0.2) provided better SVM results. This indicates that more slack was needed to allow for softer margins, meaning that outliers in the training datasets needed to be misclassified to yield a more sensible maximum margin boundary. Our best results for Movement 2 were all obtained from WPC modality (Table 2).

3.3. Movement 3

Similar to the first movement, naïve Bayesian was the best OVR classifier for Movement 3, followed by the Gaussian SVMs (Table 3). kNNs again failed to match these numbers with their best result yielding a Q factor of 52% (WDC features, k value of 31).

3.4. Movement 4

For this seemingly most challenging case, Gaussian SVMs with $\sigma = 15$ and $C = 0.1$ or 0.2 yielded the best results using WPC features (Table 4). The stronger performance of Gaussian SVMs in detection of this difficult class, as indicated by the Q figures, indicates a more nonlinear decision boundary. Corroborating the highly nonlinear nature of class boundaries for this movement, low-order polynomial SVMs were only able to reach quality factors in the vicinity of 50%.

Table 1
Best 5 Classifiers for Movement 1

Classifier	*Configuration*	*Selected Features (ROC AUC ranks)*	*Feature type*	*Sensitivity*	*Specificity*	*Correct rate*	*Q factor*
Naïve Bayes	Diagonal Linear	1-79	WPC	0.7	0.801	0.832	0.727
Naïve Bayes	Diagonal Linear	1-83	WPC	0.699	0.803	0.833	0.725
Naïve Bayes	Diagonal Linear	1-81	WPC	0.695	0.803	0.838	0.725
Naïve Bayes	Diagonal Linear	1-66	WPC	0.697	0.804	0.836	0.725
SVM	σ =15, C=2	1-17	WPC	0.689	0.734	0.749	0.703

Table 2
Best 5 Classifiers for Movement 2

Classifier	*Configuration*	*Selected Features (ROC AUC ranks)*	*Feature type*	*Sensitivity*	*Specificity*	*Correct rate*	*Q factor*
SVM	σ =15, C=0.2	1-70	WPC	0.72	0.804	0.827	0.741
SVM	σ =15, C=0.1	1-65	WPC	0.715	0.792	0.813	0.734
Naïve Bayes	Diagonal Linear	1-80	WPC	0.701	0.808	0.839	0.728
Naïve Bayes	Diagonal Linear	1-78	WPC	0.701	0.803	0.833	0.727
Naïve Bayes	Diagonal Linear	1-68	WPC	0.700	0.801	0.834	0.727

Table 3
Best 5 Classifiers for Movement 3

Classifier	*Configuration*	*Selected Features (ROC AUC ranks)*	*Feature type*	*Sensitivity*	*Specificity*	*Correct rate*	*Q factor*
Naïve Bayes	Diagonal Linear	1-90	WDC	0.702	0.757	0.772	0.716
Naïve Bayes	Diagonal Linear	1-88	WDC	0.703	0.755	0.769	0.716
Naïve Bayes	Diagonal Linear	1-89	WDC	0.698	0.751	0.769	0.715
Naïve Bayes	Diagonal Linear	1-86	WDC	0.700	0.752	0.768	0.715
SVM	σ =15, C=0.1	1-80	WPC	0.700	0.755	0.770	0.714

Table 4
Best 5 Classifiers for Movement 4

Classifier	*Configuration*	*Selected Features (ROC AUC ranks)*	*Feature type*	*Sensitivity*	*Specificity*	*Correct rate*	*Quality factor*
SVM	σ = 15, C=0.1	1-51	WPC	0.677	0.719	0.732	0.689
SVM	σ = 15, C=0.2	1-50	WPC	0.679	0.718	0.726	0.687
SVM	σ = 15, C=0.1	1-50	WPC	0.679	0.722	0.730	0.687
SVM	σ = 15, C=0.1	1-40	WPC	0.674	0.707	0.718	0.685
Naïve Bayes	Diagonal Linear	1-24	WDC	0.617	0.648	0.659	0.628

4. DISCUSSION

As more complex BCI applications require the discrimination between a larger number of brain states, the use of multi-objective metrics such as the Q factor will ease the required minimum number of individual classifiers by permitting greater use of OVR models. Used in conjunction with wrapper methods, we showed that this metric can be successfully applied to multi-class subject-invariant BCIs, a challenging problem known to lead to trivial OVR classification. More specifically, a simple best-first wrapper method (Fig. 3), applied to ROC AUC-ranked features with Q as the OVR classifier feedback, could successfully realize a 4-class subject independent BCI. Given the structure of Q, the resulting recognition rates were not only better in terms of correct classification rates, but also the overall sensitivity and specificity figures were more in balance. On the other hand, using the same dataset but without the Q factor, even more sophisticated wrapper methods such as sequential forward selection, sequential backward selection [38], and classification guided subset selection [23] failed to achieve any acceptable results, where the methods degenerated to the trivial classification using the same OVR models.

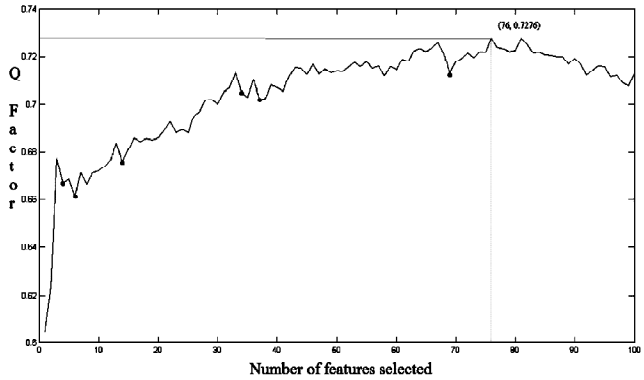

Figure 3: Quality Factor Verses Number of Features using a Best-first Wrapper with Naïve Bayesian OVR. Q Peaks at 0.728 using the first 76 Features from ROC-AUC Ranked WPC

The utilized best-first wrapper uses ROC AUC as a measure of classifiabilty to rank features within each modality. The input feature vector is then determined as the first D elements from the ranked list that together maximize the ensuing OVR's Q, compared to all other D values s from the same list. Figure 3 shows one such wrapper feature selection by plotting Q factor vs. the number of included features. As evident from the figure, the aforementioned wrapper method selected the first 76 ranked features, where Q factor peaked at 0.728 within the given 100 feature elements of the modality. The Q factor fluctuations are due to the presence of correlated elements in the concatenated feature vector, also known as the nesting problem [38]. In wrapper processes, it is possible to remove such features by backward passes after a sequential forward feature aggregation, eliminating features detrimentally correlated with subsequent selections during the evolution of the feature vector. As an example, again consider the plot of Q with respect to the length feature vector D for Movement 1 Bayesian OVR (Fig. 3), where D corresponds to number of selected features from the ROC AUC ranked WPCs. Q reaches a maximum of 0.728 at D=76, which is the conclusion point for the utilized first-best wrapper, with average cross validation sensitivity of 0.7, specificity of 0.801, and correct rate of 0.832. However, by removing five local minima at D = 4, 6, 34, 37, and 69 (marked with dots on the graph), Q factor was raised to 0.738. Though a small gain, this simple experiment confirms the existence of nesting problem. Thus, we expect that similar but more advanced sequential selection methods, such as SFFS [38], when guided by Q factor, lead to even better OVR solutions to multi-class SIBCI. This will be the subject to our future work.

As for the choice of classifier models, and given their interaction with the integrated feature selection in wrapper methods, different naïve Bayesian, SVM, and kNN classifiers were examined. Generally speaking, naïve Bayesian classifiers with diagonalized covariance matrices and normal distribution assumption, followed by Gaussian kernel SVMs, performed better than kNNs and polynomial kernel SVMs. The prevalence and success of linear discriminant naïve Bayesian classifiers can be ascribed to (a) their stability and robustness given the diagonalized covariance matrix and their ability to estimate their parameters from a relatively small number of data points, yielding better generalization and validation-based Q, and (b) the assumption of variable independence by the naïve Bayesian classifier, which is in line with that of the univariate feature ranking method used by our wrapper process.

Another interesting point is the superior performance of wavelet features, both as packets and regular decompositions, compared to other feature modalities including wavelet energies marginalized over shift, PSDs, and LPCs. This attests to the importance of temporal progressions in the given BCI

problem, as features modalities that did not appropriately incorporate such transients were left out by the Q-guided wrapper method.

5. CONCLUSION

While the subject of BCI performance evaluation has not received as much attention as the other aspects of the field, it becomes more important as BCI is applied to more complex situations such as subject invariant or multiclass systems. As a challenging real world application scenario, we designed a subject invariant 4-class BCI based on the EEG data of 10 different untrained subjects. This is a complex problem given the fact that the spatio-temporal characteristics of the EEGs differ with respect to subject and trial [8]. This inter- and intra- subject EEG variability, combined with larger number of target classes and untrained subjects, leads to degenerate multi-class BCI systems when simple classification rate is used as the metric for data driven OVR feature selection and classification. Here we introduced a simple but effective metric, Q factor, in conjunction with wrapper methods to avoid the trivial classification. Overall, best performing Q-guided classifiers employed features selected from either WPC or WDC modalities. Gaussian SVMs and naïve Bayesian classifiers with linear discriminant functions outperformed kNN and polynomial SVM classifiers, which is in line with other reports on the choice of classifiers for BCI [15, 29].

Among the candidate feature modalities, WPC and WDC emerged as the best using the Q-based wrapper methods, as opposed to non-temporal modalities such as PSD and wavelet scale energy features, except for STFT. This means that within our experiment settings, features that have proper shift (time) information are more salient, and that the time-locked EEG information benefits from multi-resolution, non-sinusoidal wavelet decompositions during Q-guided wrapper feature extraction and classification.

By using Q factor instead of simple error rate, we not only avoided degenerate convergence of our wrapper methods to trivial classification, but also garnered robust results for an otherwise challenging problem of fitting one multi-class BCI to a group of untrained subjects. The best obtained OVRs for each movement were as follows (Tables 1 through 4, cross-validation results): Movement 1: sensitivity 0.7, specificity 0.801, correct rate 0.832, and a Q factor of 0.727 (Naïve Bayesian classifier, 79 WPC features). Movement 2: sensitivity 0.72, specificity 0.804, correct rate 0.827, and a Q factor of 0.741 (SVM classifier, 70 WPC features). Movement 3: sensitivity 0.702, specificity 0.757, correct rate 0.772, and a Q factor of 0.716 (Naïve Bayesian classifier, 90 WDC features). Movement 4: sensitivity 0.677, specificity 0.719, correct rate 0.732, and a Q factor of 0.689 (SVM classifier, 15 WPC features). This compares favorably with other studies which assert that 70% classifier accuracy is adequate to control a two class BCI [9], notwithstanding that we worked on a more complex 4-class problem, and calibrated a single BCI to 10 untrained subjects. This OVR result also compares with our other cross validation results on the same data set but with a different classification scheme, using error correcting codes which replaces four OVRs with six balanced non-OVR classifiers combined by majority vote [17]. As a part of our future work, we intend to use other wrapper methods with Q factor and larger datasets, especially that the latter will allow for further blind tests beyond the current cross validations. We also wish to use Q factor in conjunction with data-driven feature extraction and classification to help with other challenging aspects of BCI system such as long-term invariance, and introduction of loss functions into Q to optimize OVRs according to costs of different error types.

Acknowledgements

This work was supported in part by University of Missouri Research Board and University of Missouri – Kansas City Faculty Research Grant.

Note

1. A second degenerate mode tends to 1/M which is also the random rate.

References

[1] A. Schlögl, F. Lee, H. Bischof, G. Pfurtscheller, Characterization of Four-Class Motor Imagery EEG Data for the BCI-Competition, *Journal of Neural Engineering* (2005), 2 4.

[2] A. Schlögl, J. Kronegg, J.E. Huggins and S.G. Mason, Evaluation Criteria for BCI Research, Towards Brain-computer Interfacing, MIT press (2007), pp. 327-342.

[3] A. J. Rowan and E. Tolunsky, A Primer of EEG, Elseiver, Philadelphia, PA (2003).

[4] A.K. Jain, R. P. W. Duin and J. Mao, Statistical Pattern Recognition: A Review, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2000), 22 1.

[5] B. Dal Seno, M. Matteucci and L.T. Mainardi, The Utility Metric: A Novel Method to Assess the Overall Performance of Discrete Brain-computer Interfaces, Neural Systems and Rehabilitation Engineering (2010), 18 1.

[6] B. Obermaier, C. Neuper, C. Guger and G. Pfurtscheller, Information Transfer Rate in a Five-classes Brain-Computer Interface, *IEEE Trans. Neural System Rehabilitation Engineering* (2001), 9 3.

[7] B.Z. Allison, E.W. Wolpaw and J.R. Wolpaw, Brain Computer Interface Systems: Progress and Prospects, Expert Review of Medical Devices (2007), 4 4.

[8] C. Christoforou, R. M. Haralick, P. Sajda and L.C. Parra, The Bilinear Brain: Towards Subject-Invariant Analysis, International Symposium on Communication, Control and Signal Processing (2010).

[9] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, New York (2006).

[10] C. W. Anderson, E. A. Stolz, and S. Shamsunder, Discriminating Mental Tasks using EEG Represented by AR Models, Engineering in Medical Biology Society Annual Conferences (1995), 2.

[11] D. J. Sebald and J. A. Bucklew, Support Vector Machine Techniques for Nonlinear Equalization, *IEEE Transactions on Signal Processing* (2000), 48 11.

[12] E. A. Curran and M. J. Stokes, Learning to Control Brain Activity: A Review of the Production and Control of EEG Components for Driving Brain-computer Interface (BCI) Systems, *Brain and Cognition* (2003), 51 3.

[13] F. Galan, P. W. Ferrez, F. Oliva, J. Guardia and J. R. Millan, Feature Extraction for Multi-class BCI using Canonical Variates Analysis, *IEEE International Symposium on Intelligent Signal Processing* (2007), 1-6.

[14] F. Lotte and C. Guan, Learning from other Subjects Helps Reducing Brain-computer Interface Calibration Time, *IEEE International Conference on Acoustics Speech and Signal Processing* (2010), 614-617.

[15] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche and B. Arnaldi, A Review of Classification Algorithms for EEG-based Brain–computer Interfaces, *Journal of Neural Engineering* (2007), 4 2.

[16] I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research* (2003), 3.

[17] J. Sherwood and R. Derakhshani, An Ensemble Method for Quasi Movement, Subject-Independent Brain Computer Interfaces, *IST Transactions on Biomedical Sciences and Engineering* (2010).

[18] J. Sherwood and R. Derakhshani, On Classifiability of Wavelet Features for EEG-based Brain-computer Interfaces, *International Joint Conference on Neural Networks*, Atlanta, Georgia (2009).

[19] J. R. Wolpaw and D. J. McFarland, Control of a Two Dimensional Movement Signal by a Non-invasive Brain Computer Interface in Humans, Published by Proceedings of the National Academy of Sciences of the United States of America (2004), 101 51.

[20] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller and T. M. Vaughan, Brain–computer Interfaces for Communication and Control, *Clinical Neurophysiology* (2002), 113 6.

[21] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson and T. M. Vaughan, Brain–Computer Interface Technology: A Review of the First International Meeting, *IEEE Transactions on Rehabilitation Engineering* (2000), 8 2.

[22] K. Gwet, Inter-rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity, *Statistical Methods for Inter-Rater Reliability Assessment* (2002), 2.

[23] L. Li, D.M. Umbach, P. Terry and J.A. Taylor, Application of the GA/KNN Method to SELDI Proteomics Data, *Bioinformatics* (2004), 20 10.

[24] L. B. Jackson, Digital Filters and Signal Processing, Kluwer Academic Publishers, Second Edition (1989), 255-257.

[25] M. Alamgir, M. Grosse-Wentrup and Y. Altun, Multitask Learning for Brain-computer Interfaces, AISTATS Thirteenth International Conference on Artificial Intelligence and Statistics (2010), 17-24.

[26] M. Schroder, M. Bogdan, T. Hinterberger and N. Birbaumer, Automated EEG Feature Selection for Brain Computer Interfaces, *First International IEEE Conference on Neural Engineering* (2003), 626-629.

[27] M. Vetterli and C. Herley, Wavelets and Filter Banks: Theory and Design. *IEEE Transactions on Signal Processing* (1992), 40 9.

[28] P. Doynov, J. Sherwood and R. Derakhshani, Classification of Imagined Motor Tasks, IEEE Region 5 Conference at Kansas City (2008).

[29] P. Herman, G. Prasad, T.M. McGinnity and D. Coyle, Comparative Analysis of Spectral Approaches to Feature Extraction for EEG-Based Motor Imagery Classification, *IEEE Transactions on Neural Systems and Rehabilitation* (2008), 16 4.

[30] P. Shenoy, M. Krauledat, B. Blankertz, R.P.N. Rao and K. R. Muller, Towards Adaptive Classification of BCI, *Journal of Neural Engineering* (2006), 3.

[31] Q. Novi, C. Guan, D.H. Tran and X. Ping, Sub-band Common Spatial Pattern for Brain-computer Interface, 3rd *International IEEE/EMBS Conference on Neural Engineering* (2007), 204-207.

[32] Q. Wei, X. Gao and S. Gao, Feature Extraction and Subset Selection for Classifying Single-trial ECoG during Motor Imagery, 28th Annual International

Conference of the IEEE on Engineering in Medicine and Biology Society (2006), 1589-1592.

[33] R. Kohavi and D. Sommerfield, Feature Subset Selection using the Wrapper Model: Overfilling and Dynamic Search Space Topology, *Proceedings 1st International Conference on Knowledge Discovery and Data Mining* (1995).

[34] R. Ruiz, J.C. Riquelme and J.S. Aguilar-Ruiz, Incremental Wrapper-based Gene Selection from Microarray Data for Cancer Classification, *Pattern Recognition* (2006), 39 12.

[35] R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification, Wiley, 2nd edition, New York (2001).

[36] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, San Diego (1999).

[37] S. Ortiz Jr, Brain-Computer Interfaces: Where Human and Machine meet, *IEEE Trans. Computer Technology News* (2007).

[38] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Academic Press, Burlington, MA (2009).

[39] S. G. Mason, A. Bashashati, M. Fatourechi, K.F. Navarro and G.E. Brich, A Comprehensive Survey of Brain Interface Technology Designs, Annals of Biomedical Engineering (2007), 35 2.

[40] S. J. Roberts and W. D. Penny, Real-time Brain-computer Interfacing: A Preliminary Study using Bayesian Learning, *Medical and Biological Engineering and Computing* (2006), 38 1.

[41] S. K. Mitra, Digital Signal Processing: A Computer Based Approach, Mc Graw Hill, New York (2006).

[42] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters* (2006), 27 8.

[43] T. E. Doyle, Z. Kucerovsky and A. Ieta, Affective State Control for Neuroprostheses, *Engineering in Medicine Biology Society Annual Conference EMBS* (2006), 1248-1251.

[44] T. M. Cover and P. Hart, Nearest Neighbor Pattern Classification, *IEEE Transactions on Information Theory* (1967), 13 1.

[45] W. Ting, Y. Guo-zheng, Y. Bang-hua and S. Hong, EEG Feature Extraction based on Wavelet Packet Decomposition for Brain Computer Interface, *Measurements* (2008), 41 6.

[46] Y. Yang, An Evaluation of Statistical Approaches to Text Categorization, *Information Retrieval* (1999), 1 1.