EURASIP Journal on Applied Signal Processing 2005:19, 3128–3140

Feature Selection and Blind Source Separation in an EEG-Based Brain-Computer Interface

David A. Peterson
Department of Computer Science, Center for Biomedical Research in Music, Molecular, Cellular, and Integrative Neurosciences Program, and Department of Psychology, Colorado State University, Fort Collins, CO 80523, USA
Email: petersod@cs.colostate.edu

James N. Knight
Department of Computer Science, Colorado State University, Fort Collins, CO 80523, USA
Email: nate@cs.colostate.edu

Michael J. Kirby
Department of Mathematics, Colorado State University, Fort Collins, CO 80523, USA
Email: kirby@math.colostate.edu

Charles W. Anderson
Department of Computer Science and Molecular, Cellular, and Integrative Neurosciences Program, Colorado State University, Fort Collins, CO 80523, USA
Email: anderson@cs.colostate.edu

Michael H. Thaut
Center for Biomedical Research in Music and Molecular, Cellular, and Integrative Neurosciences Program, Colorado State University, Fort Collins, CO 80523, USA
Email: michael.thaut@colostate.edu

Received 1 February 2004; Revised 14 March 2005

Most EEG-based BCI systems make use of well-studied patterns of brain activity. However, those systems involve tasks that indirectly map to simple binary commands such as "yes" or "no" or require many weeks of biofeedback training. We hypothesized that signal processing and machine learning methods can be used to discriminate EEG in a direct "yes"/"no" BCI from a single session. Blind source separation (BSS) and spectral transformations of the EEG produced a 180-dimensional feature space. We used a modified genetic algorithm (GA) wrapped around a support vector machine (SVM) classifier to search the space of feature subsets. The GA-based search found feature subsets that outperform full feature sets and random feature subsets. Also, BSS transformations of the EEG outperformed the original time series, particularly in conjunction with a subset search of both spaces. The results suggest that BSS and feature selection can be used to improve the performance of even a "direct," single-session BCI.

Keywords and phrases: electroencephalogram, brain-computer interface, feature selection, independent components analysis, support vector machine, genetic algorithm.

1. INTRODUCTION

1.1. EEG-based brain-computer interfaces

There is a fast-growing research and development effort underway to implement brain-computer interfaces (BCI) using the electroencephalogram (EEG) [52]. The overall goal is to provide people with a new channel for communication with the external environment. This is particularly important for patients who are in a "locked-in" state in which conventional motor output channels are compromised.

One simple, desirable BCI function would allow individuals without motor function to respond to questions with simple "yes" or "no" responses [35]. Yet most BCI research has used experiments that require an indirect mapping between what the subject does and the effect on an external

system. For example, subjects may be required to imagine left- or right-hand movement in order to use the BCI [3, 37, 39]. If they want to use the BCI to respond yes/no to questions, they have to remember that left-hand imagined movement corresponds to "yes," and right-hand imagined movement corresponds to "no." Other BCI research requires extensive subject biofeedback training in order for the subject to gain some degree of voluntary influence over EEG features such as slow cortical potentials [5] or 8–12 Hz rhythms [53]. For both the imagined movement and biofeedback scenarios, the mapping between what the subject does and the effect on the BCI is indirect. In the latter case, a single session is insufficient and the subject must undergo many weeks or months of training sessions.

A more direct approach would simply have the subject imagine "yes" or "no" and would not require extensive biofeedback training. While imagined movement and bidirectional influence over time- and frequency-domain amplitude can be readily detected and used as control signals in a BCI, the EEG activity associated with complex cognitive tasks such as imagining different words is much more poorly understood. Can advances in signal processing and pattern recognition methods enable us to distinguish whether a subject is imagining "yes" or "no" by the simultaneously recorded EEG? Furthermore, can that distinction be learned in a single recording session?

1.2. The EEG feature space

The EEG measures the scalp-projected electrical activity of the brain with millisecond resolution at up to over 200 electrode locations. Although most EEG-based BCI research uses far fewer electrodes, research into the role of the specific topographic distribution of the electrodes [54] suggests that dense electrode arrays may standardize and enhance the system's performance. Furthermore, advances in electrode and cap technology have made the time required to apply over 200 electrodes reasonable even for BCI patients. EEG analyses, including much of the EEG-based BCI research, make extensive use of the signals' corresponding frequency spectrum. The spectrum is usually divided into five canonical frequency bands. Thus, if one considers the power in each of these bands for each of 200 electrodes, each trial is described by 1000 "features." If interelectrode features such as cross-correlation or coherence are considered, this number grows combinatorially. As in many such problems, a subset of features will often lead to better dissociation between trial types than the full set of features. However, the number of unique feature subsets for N features is 2^N, a space that cannot be exhaustively explored for N greater than about 25. This is but one reason why most EEG research uses only a very small number of features. A significant number of features are discarded, including features that might significantly improve the accuracy with which the signals can be classified.

1.3. Blind source separation of EEG

Given a set of observations, in our case a set of time series, blind source separation (BSS) methods such as independent component analysis (ICA) [22] attempt to find a (usually linear) transformation of the observations that results in a set of independent observations. Infomax [4] is an implementation of ICA that searches for a transformation that maximizes the information between the observations and the transformed signals. Bell and Sejnowski showed that a transformation maximizing the information is, in many cases, a good approximation to the transformation resulting in independent signals. ICA has been used extensively in analyses of brain imaging data, including EEG [26, 34], magnetoencephalogram (MEG) [47, 49], and functional magnetic resonance imaging (FMRI) [26]. Assumptions about how independent brain sources are mixed and map to the recorded scalp electrodes, and the corresponding relevance for BSS methods, are discussed extensively in [27].

Maximum noise fraction (MNF) is an alternative BSS approach for transforming the raw EEG data. It was initially introduced in the context of denoising multispectral satellite data [14]. Subsequently it has been extended to the denoising of time-series [1] and it has been compared to principal components analysis and canonical correlation analysis in a BCI [2]. The basis of the MNF subspace approach is to construct a set of basis vectors that optimize the amount of noise (or, equivalently, signal) captured. Specifically, the maximum noise fraction basis maximizes the noise-to-signal (as well as the signal-to-noise) ratio of the transformed signal. Thus, the optimization criterion is based on the ratio of second-order statistical quantities. Furthermore, unlike ICA, the basis vectors have a natural ordering based on the signal-to-noise ratio. MNF is similar to the second-order blind identification (SOBI) algorithm and requires that the signals have different autocovariance structures. The requirement exists because of the second-order nature of the algorithm.

The relationship of MNF to ICA is a consequence of the fact that they both provide methods for solving the BSS problem [1, 21]. Initial results for the application of MNF to the analysis of EEG time-series demonstrated MNF was simultaneously effective at eliminating noise and extracting what appeared to be observable phenomenon such as eye blinks and line noise [28, 29]. It is interesting that ICA and MNF perform similarly given their disparate formulations. This suggests that under appropriate assumptions (see [1, 21, 28]) the mutual information criterion and the signal-to-noise ratio can be related quantities. However, in the instance that signals of interest are mixed such that they share the same subspace, the MNF approach provides a representation for the mixed and unmixed subspaces.

1.4. Classification and the feature selection problem

The support vector machine (SVM) classifier [45, 48] learns a hyperplane that provides a maximal soft margin between the data classes in a higher-dimensional transform space determined by a choice of kernel function. Although SVMs can fail in problems with many nuisance features [19], they have demonstrated competitive classification performance in difficult domains as diverse as DNA microarray data [8], text categorization [25], and image classification [40]. They have

also been successfully employed in EEG-based BCI research [6, 12, 32, 56]. In contrast to competing nonlinear classifiers such as multilayer perceptrons, SVMs often exhibit higher classification accuracy, are not susceptible to local optima, and can be trained much faster. Because we seek feature subsets that maximize classification accuracy, the feature subset search needs to be driven by how well the data can be classified using the corresponding feature subsets, the so-called "wrapper" approach to feature selection [30]. Thus the speed characteristic of SVMs is particularly important because we will train and test the classifiers for every feature subset we evaluate.

Our prior research with EEG datasets from a cognitive BCI [2] and movement prediction BCI [12] demonstrated the benefit of feature selection for small and large feature spaces, respectively. There are many ways to implement the feature selection search [7, 16, 42]. One logical choice is a genetic algorithm (GA) [13, 20]. GAs provide a stochastic global search of the feature subset space, evaluating many points in the space in parallel. A population of feature subsets is evolved using crossover and mutation operations akin to natural selection. The evolution is guided by how well feature subsets can classify the trials. GAs have been successfully employed for feature selection in a wide variety of applications [15, 51, 55] including EEG-based BCI research [12, 56]. GAs often exhibit superior performance in domains with many features [46], do not get trapped in local optima as with gradient techniques, and make no assumptions about feature interactions or the lack thereof.

In summary, this paper evaluates a feature selection system for classifying trials in a novel, challenging BCI using spectral features from the original, and two BSS transformations of, scalp recorded EEG. We hypothesized (1) that classification accuracy would be higher for the feature subsets found by the GA than for full feature sets and random feature subsets and (2) that the power spectra of the BSS transformations would provide feature subsets with higher classification accuracy than the power spectra of the original signals.

2. METHODS

2.1. Subjects

The subjects were 34 healthy, right-handed fully informed consenting volunteers with no history of neurological or psychiatric conditions. The present paper is based on data from eight of the subjects who met certain criteria for behavioral measures and details of the EEG recording procedure. Specifically, we selected eight subjects that wore caps with physically linked mastoids for the reference. Other subjects wore a cap with mastoids digitally linked for the reference. Although the difference between physically and digitally linked mastoid reference is minor, it can be nontrivial depending on the relative impedances at the two mastoid electrodes [36]. Thus, to eliminate the possibility that the slight difference in caps could influence the questions at hand, we elected to consider only those subjects wearing the cap with physically linked mastoids. We also considered only those subjects that exhibited reasonable inter-response intervals and a reasonably even distribution of "yes"/"no" responses in a separate, voluntarily decided premotor visualization version of the task (described in a separate forthcoming manuscript). The subjects were selected on these criteria only, before their EEG data was reviewed. The eight subjects were 19 + / − 1 years of age and included five females.

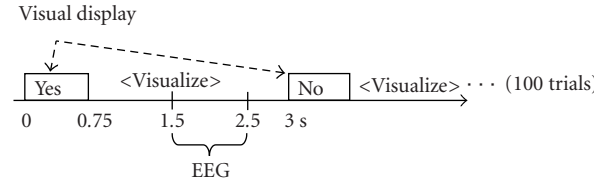

Figure 1: *BCI task timeline.* Subjects were asked to visualize the most recently presented word until the next word is displayed. The period of simultaneously recorded EEG used for subsequent analysis was 1000 milliseconds long beginning 750 milliseconds after display offset and 500 milliseconds before the next display onset.

2.2. BCI experiment procedure

On each of 100 trials subjects were shown one of the words "yes" or "no" on a computer display for 750 milliseconds and were instructed to visualize the word until the next word is displayed (see Figure 1). There were 50 "yes" trials and 50 "no" trials presented in random order with a maximum of three of the same stimulus in a row. Because in subsequent analyses we planned to ignore the first two trials due to experiment start-up transients, the first two trials were required to include exactly one of each type.

2.3. EEG recording and feature composition

The EEG was continuously recorded with a 32-electrode cap (QuikCap, Neuroscan, Inc.), pass band of 1–100 Hz, and sampled at 1 kHz. Although much higher than the 200 Hz required by Nyquist, we typically sample at 1 kHz for the mere convenience that in subsequent time-domain analyses and plots, samples are equivalent to milliseconds. Electrodes FC4 and FCZ were excluded because of sporadic technical problems with the corresponding channels in the amplifier. The remaining 30 electrodes used in subsequent analysis included bipolar VEOG and HEOG electrodes commonly used to monitor blinks and eye movement artifacts. All other electrodes were referenced to physically linked mastoids. We did not employ any artifact removal or mitigation in the present study, as we sought to measure performance without the added help or complexity of artifact mitigation techniques.

The BSS methods were applied to the continuously recorded EEG data from the beginning of the first epoch to the end of the last. The majority of the continuous record represented task-related activity because the intertrial period was only approximately 30 milliseconds. We used the Matlab implementation of Infomax available as part of the EEGLAB

software[1] [10]. The EEGLAB software first spheres the data, which decorrelates the channels. This simplifies the ICA procedure to finding a rotation matrix which has fewer degrees of freedom [23]. Except for the convergence criteria, all of the default parameter values for EEGLAB's Infomax algorithm were used. Initially, extended Infomax, which allows for sub-Gaussian as well as super-Gaussian source distributions, was used. No sub-Gaussian sources were extracted on the first two subjects so the standard Infomax approach was used on all of the subject data. An initial transformation matrix was found with a tolerance of 0.1. The algorithm was then rerun with this transformation matrix and a tolerance of 0.001.

To investigate whether comparing Infomax ICA and the MNF method would be of empirical value, a simple test was performed on the data set for several subjects. Both transforms were applied to each subject's data and the resulting components were compared. The cross-correlation for all Infomax-MNF component pairs was computed, and the optimal matching was found. This matching paired the components so that the maximal cross-correlation was achieved. Had the components produced been the same, the cross-correlation measure would have been 100%. Cross correlations of 60–70% were found in the tests performed, and so we decided the two transforms were sufficiently dissimilar to warrant the evaluation of both in the study.

Each of the original, Infomax, and MNF-transformed data were "epoched" such that the one-second period beginning 750 milliseconds after stimulus offset was used for subsequent analysis. Because iconic memory is generally thought to last about 500 milliseconds, this choice of temporal window should minimize the influence of iconic memory and place relatively more weight on active visualization processes. We then computed spectral power for each channel (component) and each trial (epoch) using Welch's periodogram method that uses the average spectra from overlapping windows of the epoch. We computed averaged spectral power in the delta (2–4), theta (4–8), lower alpha (8–10), upper alpha (10–12), beta (12–35), and gamma (35–50 Hz) frequency bands. Thus, the full feature set contains 30 electrodes $\times$ 6 spectral bands each for a total of 180 features. The first and second trials were excluded to reduce the transient effects of the start of the task. Thus, all subsequent analyses use 49 trials of each type ("yes," "no") for each subject. All reported results are for individual subjects.

2.4. Classification

In the present report, we sought subsets from a very large feature set that would maximize our ability to distinguish "yes" from "no" trials. The distinction was tested with a support vector machine (SVM) classifier and an oversampled variant of 10-fold cross-validation.

As discussed in the introduction, we chose a support vector machine (SVM) classifier because of its record of very good classification performance in challenging problem domains and its speed of training. We used a soft margin SVM[2] with a radial basis function (RBF) kernel with $\gamma = 0.1$. The SVM was trained with regularization parameter $\upsilon = 0.8$, which places an upper bound on the fraction of error examples and lower bound on the fraction of support vectors [44]. Given m training examples $X\{x_1, \ldots, x_m\} \subseteq R^N$ and their corresponding class labels $Y = \{y_1, \ldots, y_m\} \subseteq \{-1, 1\}$, the SVM training produces nonnegative Lagrange multipliers α_i that form a linear decision boundary:

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i k(x, x_i) \tag{1}$$

in the feature space[3] defined by the Gaussian kernel (of width inversely proportional to γ):

$$k(x, x_i) = \exp\left(-\gamma ||x - x_i||^2 \right). \tag{2}$$

On each feature subset evaluation, we trained and tested the SVM on one full run of stratified 10-fold cross-validation, randomly selecting with replacement 10% of the trials on each fold for testing.

2.5. Feature selection

We used a genetic algorithm (GA) to search the space of feature subsets in a "wrapper" fashion (see Figure 2). Individuals in the GA were simply bit strings of length 180, with a 1 indicating the feature was included in the subset and 0 indicating it was not. Our Matlab GA implementation was based on Goldberg's original simple GA [13], using roulette-wheel selection and 1-point crossover. We used conventional values for the probability of crossover (0.6) and that of mutation ($1/(4 * D)$, where D = number of features, or 0.0014). We evolved a population of 200 individuals over 50 generations. Each individual's "fitness" measure was determined by the corresponding subset's mean classification accuracy.

We instrumented the GA with a mechanism for maintaining information about the cumulative population, that is, all individuals evaluated thus far. Thus, individuals that were evaluated more than once develop a list of evaluation measures (classification accuracies). This took advantage of the inherent "resampling" that occurs in the GA because relatively "fit" individuals are more likely to live on and be reevaluated in later generations than "unfit" individuals. Such resampling, with different partitions of the trials into training/test sets on each new evaluation, reduces the risk of overfitting due to selection bias. The empirical effect of this oversampled variant of cross-validation and its role in feature selection search is illustrated in the first part of Section 3. All

[1] Available from the Swartz Center for Computational Neuroscience, University of California, San Diego, http://www.sccn.ucsd.edu/eeglab/index.html.

[2] The SVM was implemented with version 3.00 of the OSU SVM Toolbox for Matlab [33], which is based on version 2.33 of Dr. Chih-Jen Lin's LIBSVM.

[3] Here "feature space" refers to the space induced by the RBF kernel, not to be confused with the feature space, and implicit space of feature subsets, referred to elsewhere in the manuscript.

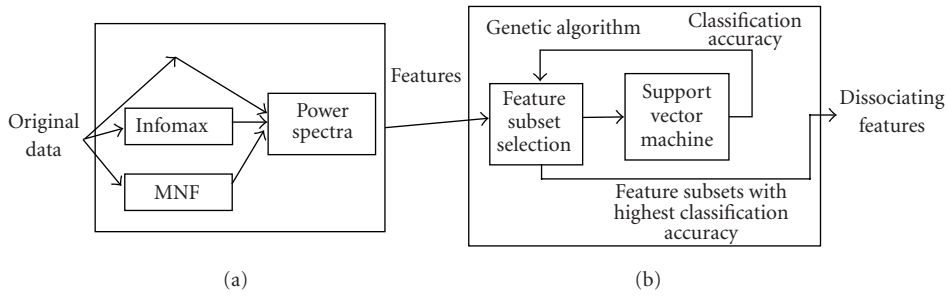

Figure 2: *Feature selection system architecture.* Three feature "families" were composed with parallel and/or series execution of signal transformations. Feature subsets are then evaluated with a support vector machine (SVM) classifier and the space of possible feature subsets searched by a genetic algorithm (GA) guided by the classification accuracy of the feature subsets. (a) Feature composition. (b) Feature selection. (Adapted from [12, Figure 1].)

subsequent reports of classification accuracy use the mean of the 10 best feature subsets that were subjected to at least five "sample evaluations" each.

3. RESULTS

3.1. Fitness evolution and overfitting at the feature selection level

Figure 3 shows how the fitness of feature subsets evolves over generations of the GA. In these and subsequent figures, the "chance" level of classification accuracy (50%) is shown with a dotted line. Note that even at the first generation of randomly selected feature subsets, the average performance of the population is slightly above chance at 54%. This suggests that, on average, randomly chosen feature subsets provide some small discriminatory information to the classifier. The approximately 70% accuracy maximum mean fitness in the first generation of the GA represents a single "sampling" of the 10-fold cross-validation. Thus, there exists a set of 10 randomly chosen training/test trial partitions for which one of the 200 initial, randomly chosen feature subsets gave 70% classification accuracy. However, such results need to be assessed with caution, as illustrated in the right panel of Figure 3. Further "sampling" for a given feature subset (i.e., repetitions of a full 10-fold cross-validation) gives a more accurate picture of that feature subset's ability to dissociate the "yes" and "no" trials.

3.2. The benefit of feature selection

Figure 4 shows how classification accuracy is improved when comparing feature subsets selected by the GA with full feature sets. For every BSS transformation (original, Infomax, and MNF) every subject's "yes"/"no" visualizations are better distinguished with feature subsets than with the whole feature set.

3.3. The benefit of BSS transformations

Figure 5 shows for each subject how the classification accuracies compare for the original signals and the two BSS transformations. For every subject, at least one of the BSS transformations leads to better classification accuracy than the original signals. Spectra of Infomax and MNF transformations performed statistically significantly better than the spectra of the original signals for every subject except subject 1 and MNF for subject 5 (Wilcoxon rank-sum test, alpha $= 0.05$). The relative performance of the three transformations does not appear to be an artifact of random processes in the GA because it holds across two entirely separate runs of the GA.

3.4. Intersubject variability in good feature subsets

Figure 6 shows the features selected for the feature subsets that provided the highest classification accuracy. For both subjects, the features include a diverse mix of electrodes and frequency bands. Although spatial trends emerge (e.g., the full power spectrum was included for electrodes FC3 and CZ), no single frequency band was included across all electrodes. Also, there appears to be some consistency between subjects in terms of the selected features. Subject 1's best feature subset included 106 features and subject 6's best feature subset included 91 features. The two subjects' best subsets had 57 features in common, including broadband features from central and left frontocentral scalp regions.

3.5. Feature values corresponding to the "yes" and "no" trials

Figure 7 shows the median values of the features across the 49 trials of each type for subject 6. Although a spatiospectral pattern of differences is shown in the lower part of the figure, none of the *individual* features exhibited significant differences between the two conditions. A few were significant at the $p < 0.05$ level (0.02-0.03), but certainly not after adjusting for multiple comparisons. Some of the features with notable differences between "yes" and "no" were included in subject 6's best feature subset (e.g., multiple bands from CZ, FZ, and FC3). However, a number of such features were not included in subject 6's best feature subset (e.g., delta band power in P3, F7, FP2, and F8—see Figure 6a and Figure 7c).

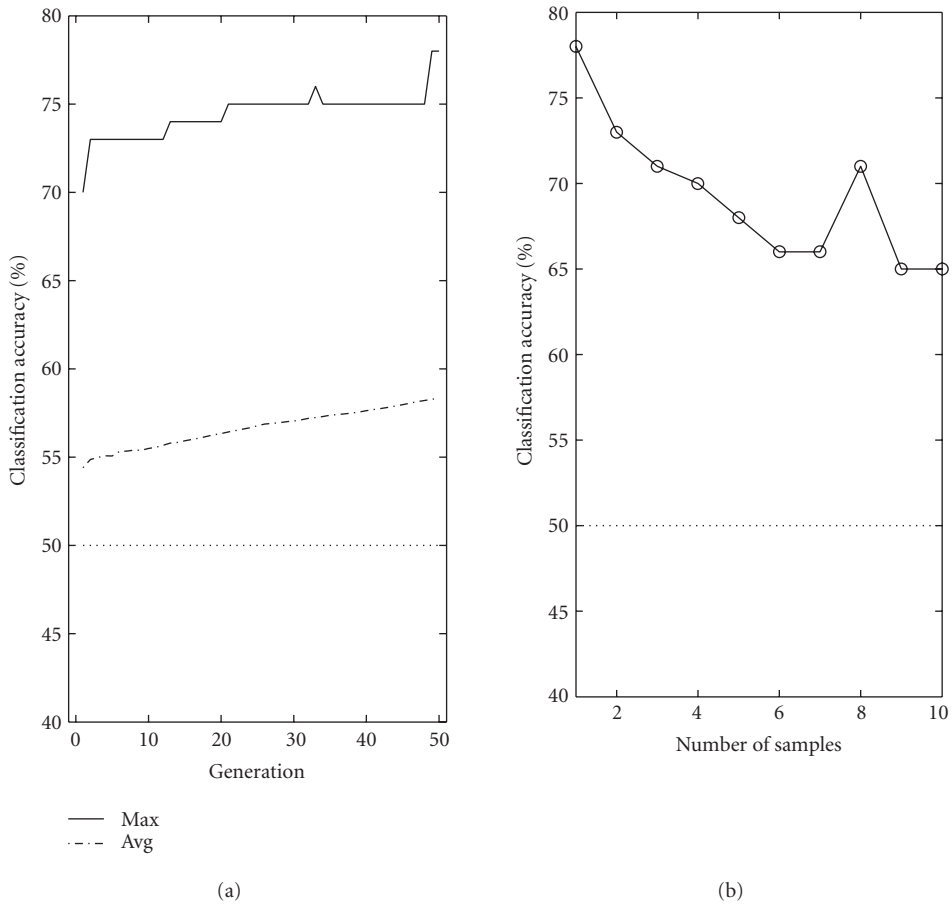

FIGURE 3: *Feature subset evolution and overfitting.* (a) Mean fitness of all individuals in the cumulative population as of that generation; "avg" is the average and "max" the maximum mean fitness. Data shown is for subject 6, Infomax transformation. Note that the maximum mean fitness in the cumulative population does not monotonically increase because repeated sampling of a particularly fit individual may reduce that individual's mean fitness value (see (b)). (b) Mean fitness of the best individual in the population for each of several different "sampling" values. Each "sample" is the mean classification accuracy from a full 10-fold cross-validation run, which uses 10 randomly selected train/test partitions of the trials for that subject. The generally decreasing function reflects overfitting at the feature selection level, whereby so many feature subset evaluations occur that the system finds train/test partitions of the trials that lead to higher-than-average fitness for a specific feature subset. Additional sampling of how well that feature subset classifies the data increases confidence that the oversampled result is not simply due to 10 fortuitous partitions of the trials.

4. DISCUSSION

4.1. Feature selection in the EEG-based BCI

We implemented a feature selection system for optimizing classification in a novel, "direct" EEG-based BCI. For all three representations of the signals (original, Infomax, and MNF) and for all subjects, the GA-based search of the feature subset space leads to higher classification rates than both the full feature sets and randomly selected subsets. This indicates that choosing feature subsets can improve corresponding classification in an EEG-based BCI. This also indicates that it is not simply smaller feature sets that lead to improved classification, but the selection of specific "good" feature subsets. Also, classification accuracy improves over generations of the GA's feature subset search, indicating that the GA's iterative search process leads to improved solutions. We ran the GA for over 700 generations for one subject's Infomax data, and the resultant feature subsets demonstrated more than a 14% increase in classification accuracy over that obtained after just 50 generations. Although this suggests an extensive search of the feature subset space may be beneficial, the roughly one week of additional computational time may be inappropriate for some BCI research settings.

Note that, as mentioned in the introduction, there are many ways to conduct the feature subset search and the GA is only one family of such search methods. Sequential forward (or backward) search (SFS) methods add features one at a time but can suffer from nesting wherein optimal subsets are missed because previously "good" features are no longer jointly "good" with other newer features and cannot be removed. The same limitation applies to backward versions of SFS that subtract single features from a full feature set. Floating versions of these methods, sequential forward floating search (SFFS), and sequential backward floating search

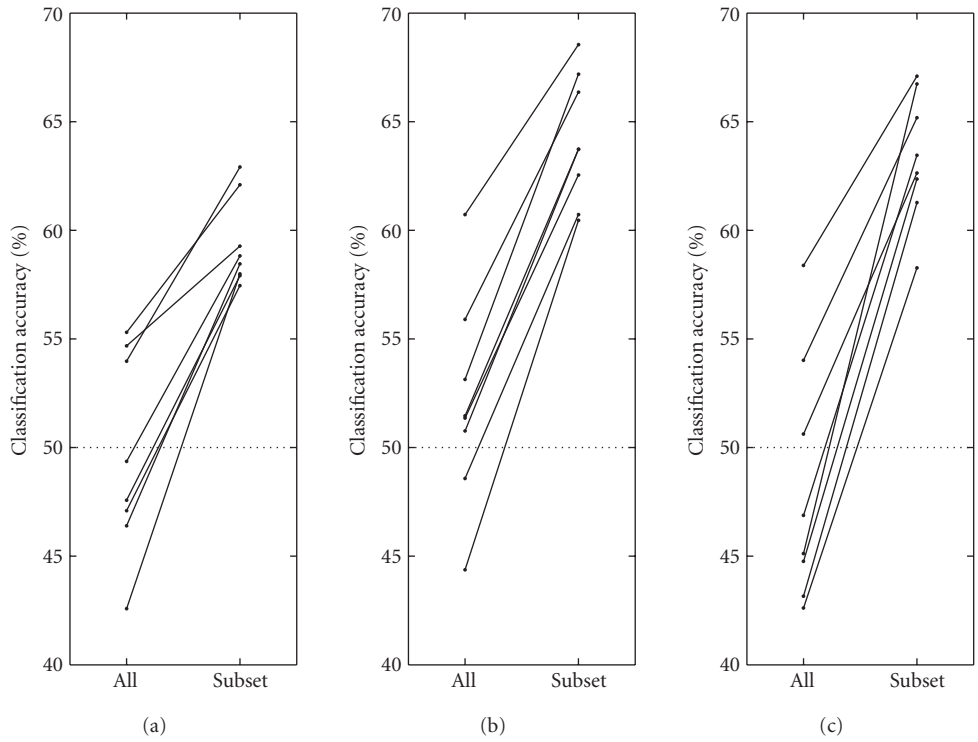

Figure 4: *Feature subsets outperform the whole feature set across feature classes and subjects.* "All" refers to the full set of all features, and "subset" refers to the feature subsets found by the GA. Each line connects the mean classification accuracies for both cases for a single subject for each of the (a) "original," (b) "Infomax," and (c) "MNF" transformations.

(SBFS) [41], mitigate the nesting problem by variably adding and taking away previously added features. In principle, both GAs and the floating methods allow for complex feature-feature interactions. However, their migration thru the subset space can differ substantially. Depending on how they are implemented, sequential methods can implicitly assume a certain ordering to the features, whereas GAs do not make that assumption. Similarly, SFFS/SBFS are not as "global" in their search as a GA. The floating search methods cannot "jump" from one subset to a very different subset in a single step as is inherent in typical GA implementations. Whether or to what extent these differences affect the efficacy of the search methods depends on the problem domain and needs to be evaluated empirically. A few investigators have compared the floating search methods SFFS/SBFS to GAs for feature selection [11, 24, 31]. Kudo and Sklansky have demonstrated that GAs outperform SFFS and SBFS when the number of features is greater than about 50 [31]. Another class of feature selection methods is known as "embedded" methods. In the embedded approach, the process of selecting features is embedded in the use of the classifier. One example is recursive feature elimination (RFE) [17, 50], which has recently been used in an EEG-based BCI [32]. RFE takes advantage of the feature ranking inherent in using a linear SVM. However, as with other embedded approaches to feature selection, it lacks the flexibility of wrapper methods because, by definition, the feature subset search cannot be separated from the choice of classifier. Feature selection research has only recently begun with EEG and a comparison of feature selection methods with EEG needs to be conducted.

We also demonstrated and addressed the issue of overfitting at the level of feature selection. The sensitivity of any single feature subset's performance to the specific set of 10 train/test trial partitions is a testament to the well-known but often overlooked trial-to-trial variability of the EEG. It is also an empirical illustration of overfitting resulting from extensive search of the feature subset space, also known as "selection bias" [43]. Our feature subset search conducts many feature subset evaluations (e.g., 200 individuals over 50 generations = 10, 000 evaluations) and there are many ways to randomly choose a partition of training/test trials. Thus, there exist 10 random training/test partitions of the trials for which specific feature subsets will do much better than average if evaluated over other sets of 10 random train/test partitions. Fundamentally, as more points in the feature subset space are tested, the risk of finding fortuitous sets of train/test partitions increases, so greater partition sampling is required. In the case of a GA-based feature selection algorithm, we could make the partition sampling dynamic by, for example, increasing the amount of resampling as the GA progresses thru generations of evolution. However, increasing the data partition sampling over the course of the feature subset search would of course slow down the system as the search progresses. Nevertheless, the GA's inherent resampling and the

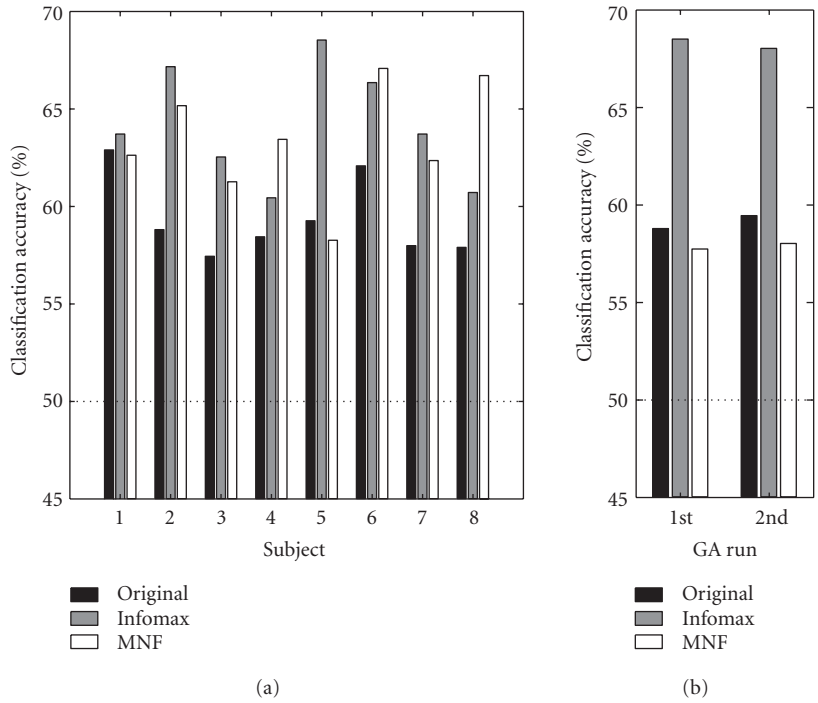

Figure 5: *The benefit of the BSS transformations and the replicability of their relative value between GA runs.* (a) Mean classification accuracy of the 10 best feature subsets with at least 5 "sample evaluations." (b) The performance results for the three transformations for subject 5 over two separate runs of the GA.

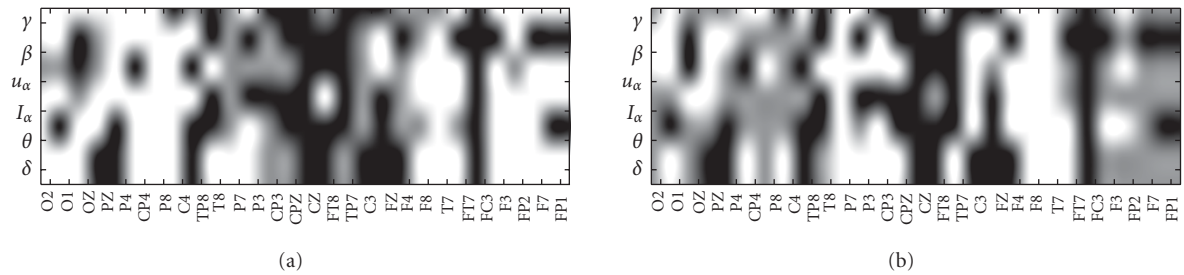

Figure 6: *Features selected in a "good" subset of the original spectral features and their overlap between two subjects.* (a) Subject 6, (b) subject 1. White indicates the feature was not selected, grey indicates that the feature was selected for that subject only, and black indicates the feature was selected for both subjects.

ease with which such resampling could be implemented in a GA provide yet another reason to use a GA for the feature subset search in extremely noisy domains such as EEG.

How best to address the overfitting issue remains an active line of research. There are numerous data partitioning and resampling methods such as leave-one-out or the bootstrap. Although we partially mitigated the issue by using an oversampled variant of cross-validation, a more principled approach needs to be developed for highly noisy, underdetermined problem domains. Although one should use as test data trials unseen during the feature subset search [43], this further exacerbates the problem of having so few trials as is typically the case with single-session EEG experiments. The current experiment had roughly 50 trials per condition per subject. Although experimental sessions with many more trials per condition raise concerns about habituation and arousal, the benefits for evaluating classifiers and associated feature selection may outweigh the disadvantages. In cases such as the present study with a limited number of trials, oversampling methods such as the bootstrap or the resampling GA variant we used may provide a reasonable alternative to the full, nested cross-validation implied by separate classifier model selection and feature subset search.

4.2. The classifier and subset search parameter space

We used only nonlinear SVMs in this study. A theoretical advantage over linear SVMs is that they can capture nonlinear relationships between features and the classes. However,

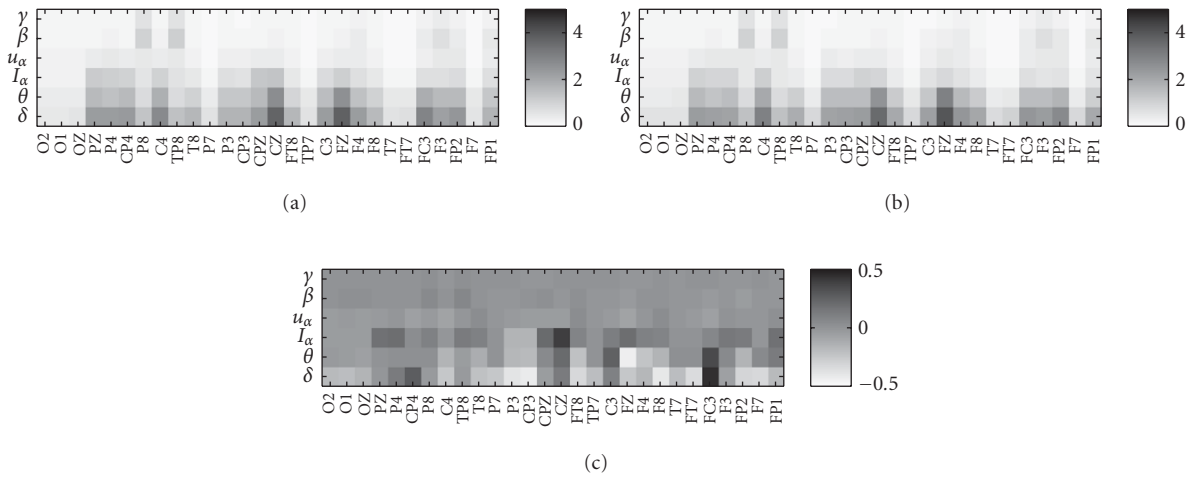

FIGURE 7: *Median feature values for the two kinds of trials.* (a) "Yes", (b) "no", and (c) difference values for subject 6, original spectra features. Bars on right show normalized spectral power (or power difference, for "yes"−"no").

nonlinear classifiers have the disadvantage that the classifier's weights do not provide a simple proxy measure of the input feature's importance, as is the case with the linear SVM formulation. We also used only one setting of SVM parameters in this study. The optimal width of the Gaussian SVM kernel, γ, in particular is known to be sensitive to the classifier's input dimensionality (number of features). Although we could have varied γ as a function of the subset size, we explicitly chose not to. If we had varied γ in a principled way (e.g., larger for fewer features), the exact formulation would be arbitrary. If we would have conducted SVM model selection and optimized γ empirically, it would have introduced another loop of cross-validation in addition to that used to train and test the SVM for every subset evaluation. This would not only be substantially more computationally demanding, but also exacerbate the risk of overfitting or reduce the amount of trials available for training/testing. In either case, allowing γ to vary would introduce another variable and we would not know whether differences in performance between feature subsets should be attributed to the subsets themselves or their correspondingly tuned classifier parameters. Although the relative performance of the full versus partial feature subsets is sensitive to γ, we expect that the relationship found in the present study would remain because feature selection usually improves classification accuracy in EEG-based BCIs. Note also that the relative performance of feature selection using the original versus BSS-based features was based on a consistent application of γ and the subsets contained roughly equivalent numbers of features.

We also used only one setting of GA parameters in this study. In general, one would expect that classification accuracy and the feature selection process are sensitive to the parameters used in the SVM and GA. In fact, especially in wrapper approaches to feature selection, the classifier's optimal parameters and optimal feature selection search algorithm parameters will not be independent. In other words, the optimal SVM model parameters will be sensitive to the specific feature subset, and vice versa. Thus, it may be suboptimal to conduct the model selection separately from the feature selection. Instead, the SVM model selection process and the feature subset search should be conducted simultaneously rather than sequentially. We have recently demonstrated this empirically with DNA microarray data [38], a domain with noise characteristics and input dimensionality not unlike that of EEG features. Although the SVM parameters could be encoded into a bit string and optimized with a GA in conjunction with the feature subset, the two optimization problems are qualitatively different and should probably be conducted with separate mechanisms. This remains a question for further research.

4.3. BSS in EEG-based BCI

Our results showed that the power spectra of the BSS transformations provided feature subsets with higher classification accuracy than the power spectra of the original EEG signals. This improvement held for seven out of eight subjects and was consistent across independent runs of the GA. The results suggest that BSS transformations of EEG signals provide features with stronger dissociating power than features based on spectral power of the original EEG signals. Infomax and MNF differed only slightly, but both provided a marked improvement in classification accuracy over spectral transformations of the original signals. This suggests that use of a BSS method may be more important than the choice of specific BSS method, although further tests with other BSS methods and other datasets would be required to substantiate that interpretation.

In some EEG research using ICA, the investigator evaluates independent components manually. This can be considered a manual form of feature selection. However as with the "filter" approach to feature selection, the features are

not selected based on their impact on the accuracy of the final classifier in which they are used. Rather, they are selected based on characteristics such as their scalp topography, the morphology of their time course, or the variance of the original signal for which they account. In some cases, the decision about which features to keep is subjective. In the present study we explicitly chose not to take this approach. Instead, we used the wrapper approach to search the full feature set based exclusively upon the components' contribution to classification. Of course, this does not preclude the possibility that preceding automated feature selection with a manual filter approach to feature selection would improve overall performance. Many domains benefit from the joint application of manual and automated approaches, including methods that do and do not leverage domain-specific knowledge.

4.3.1. "Good" feature subsets

Subjects' best feature subsets included many features from the full feature set. We believe that this may be at least partially the result of crossover in the GA, whereby new individuals will tend toward having approximately half of the features selected. The fitness function used by the GA to search the space of feature subsets used only those subsets' classification accuracy. We did not use selective pressure to reduce the number of features in selected subsets. However, this could be easily implemented by simply biasing the fitness function with a term that weights the cardinality of the subsets under consideration. If there exist many feature subsets of low cardinality that perform roughly as well as subsets with higher cardinality, then one would generally prefer the low-cardinality solutions because subsets with fewer features would, in general, be easier to analyze and interpret.

Good feature subsets included a disproportionately high representation of left frontocentral electrodes. This topography is consistent with a role for language production, including subvocal verbal rehearsal. It suggests that the cortical networks involved in rehearsing words may exhibit dissociable patterns of activity for different words. The spatial information in the EEG scalp topography is insufficient to determine whether the networks used for rehearsing the two words had differentiable anatomical substrates. However, such differences may be detectable with dipole analysis of high-density EEG and/or functional neuroimaging.

We compared subjects' good subsets of spectral power based on original EEG signals. Of the two subjects whose best feature subsets we analyzed, approximately 60% of the included features were common to both subjects. The common features included several spectral bands in left frontocentral electrodes. We did not compare subjects' good subsets using BSS-transformed EEG. One disadvantage of the BSS methods is that, because they are usually used to transform full continuous EEG recordings on a per-subject basis, there is no immediately apparent way to match one subject's components with another subject's components. Although this can be attempted manually, the process can be subjective and problematic. Often only some of the components have similar topographies and/or time courses between subjects, and the degree of similarity can be quite variable. Thus it may be difficult to compare selected features among different subjects when the features are based on BSS transformations of the original EEG signals.

The pattern of actual feature values was very similar for the "yes" and "no" trials. Because both conditions involved the same type of task, it is reasonable to assume that the associated brain activity would be similar at the level of scalp-recorded EEG. None of the individual features differed significantly between the two conditions. Although some of the features with highest amplitude differences between "yes" and "no" were included in the best (most dissociating) feature subsets, other such features were not. At the current point in this research, we cannot conclude whether this is because certain features were not considered in the GA-based search, or because the interactions of certain features do better than those single features. Evidence for or against the former interpretation could be excluded by adding a simple per-feature test to the GA's search of the feature subset space. Note that single features can have identical means (indeed, even identical distributions) for "yes" and "no" trials, yet contribute to a feature subset's ability to dissociate the two trial types because of class-conditional interdependencies between the features. Per-feature statistical tests, and some feature selection methods, for that matter, assume the features are independent, ignoring any interactions among the features. Such assumptions are generally too limiting for complex, high-dimensional domains such as EEG. Besides, even when the features are independent, there are cases when the d best features are not the same as the best d features [9, 18].

4.4. BCI application relevance

Our BCI task design provides a native interface for a patient without any motor control to directly respond "yes" or "no" to questions [35]. The paradigm provides a good model for a BCI setting in which the caregiver initiates dialog with the patient. Furthermore, it avoids the indirect mappings and extensive biofeedback training required in other BCI designs. However, this "direct" task design has some clear limitations. First, we do not have any control over what the subject is doing when they are supposed to be visualizing the word. The subjects could have been daydreaming on some trials or, perhaps even worse, still visualizing the word from an earlier trial. Of course this would degrade classification accuracy and may be a more severe problem for neurologically impaired patients compared to the healthy, albeit perhaps less motivated, volunteers we used. Second, even if subjects are performing the task as instructed, different subjects may use different cognitive processes with correspondingly different neural substrates. For example, subjects that maintain visualizations close to the original percept will recruit relatively more early visual system activity (e.g., in occipital/temporal areas), whereas subjects that maintain the word in a form of working memory will probably recruit the front-temporal components of the phonological loop. These two systems involve cortico-cortical and thalamocortical loops producing different changes in oscillatory electrophysiology usually manifest as changes in gamma and theta/alpha bands,

respectively. Thus, the spectral and topographic features that best distinguish the yes/no responses will most likely vary per subject. Indeed, this is one of the biggest motivations for taking a feature selection approach to EEG-based BCIs and conducting the feature selection search on a strictly per-subject basis as we did in the present study. Third, and perhaps most notably, the classification accuracy is far below that obtained in studies using "indirect" approaches. Nothing about our approach precludes having more than one session and therefore many more trials with which to learn good feature subsets and improve classification accuracy. Also, although indirect approaches will probably continue to provide high classification accuracy (and therefore a generally higher bit rate) for the near future, advances in basic cognitive psychology and cognitive neuroscience may provide more clues about what might be good EEG features to use to distinguish direct commands such as visualizing or imagining yes/no or on/off responses. In the meantime, BSS transformations and feature selection may provide moderate classification performance in "direct" BCIs and even help inform basic scientists about the EEG features on which to focus their research.

Our approach to feature selection is amenable to the development of on-line BCI applications. One could use the full system, including the GA, to learn off-line the best feature subset for a given subject and task, then use the trained SVM with that feature subset and without the GA in an on-line setting. Dynamic adjustments to the optimal feature subset can be continuously identified off-line and reincorporated into the on-line system. Also, as suggested in the results, the best feature subset may include features from only a small subset of electrodes. The potentially much smaller number of electrodes could be applied to the subject, reducing application time and the risk of problematic electrodes for easier on-line use of the BCI. Although we intentionally used a design without biofeedback, one could supplement this design with feedback. Other groups have found that incorporation of feedback can be used to increase classification accuracy. Feature selection could provide guidance on which features are most significant for dissociating classes of EEG trials, and therefore one source of guidance for choice of information to use in the feedback signals provided to the subject.

5. CONCLUSION

Signal processing and machine learning can be used to enhance classification accuracy in BCIs where a priori information about dissociable brain activity patterns does not exist. In particular, blind source separation of the EEG signals prior to their spectral power transformation leads to increased classification accuracy. Also, even sophisticated classifiers like a support vector machine can benefit from the use of specific feature subsets rather than the full set of possible features. Although the search for feature subsets exacerbates the risk that the classifier will overfit the trials used to train the BCI, a variety of methods exist for mitigating that risk and can be assessed over the course of feature subset search. Feature selection is a particularly promising line of investigation for signal processing in BCIs because it can be used off-line to find the subject-specific features that can be used for optimal on-line performance.

ACKNOWLEDGMENTS

The authors thank three anonymous reviewers for many helpful comments on the original manuscript, Dr. Carol Seger for use of Psychology Department EEG Laboratory resources, and Darcie Moore for assistance with data collection. Partial support provided by Colorado Commission on Higher Education Center of Excellence Grant to Michael H. Thaut and National Science Foundation Grant 0208958 to Charles W. Anderson and Michael J. Kirby.

REFERENCES

[1] M. G. Anderle and M. J. Kirby, "An application of the maximum noise fraction method to filtering noisy time-series," in *Proc. 5th International Conference on Mathematics in Signal Processing*, University of Warwick, Coventry, UK, 2001.

[2] C. W. Anderson and M. J. Kirby, "EEG subspace representations and feature selection for brain-computer interfaces," in *Proc. 1st IEEE Conference on Computer Vision and Pattern Recognition Workshop for Human Computer Interaction (CVPRHCI '03)*, vol. 5, Madison, Wis, USA, June 2003.

[3] F. Babiloni, F. Cincotti, L. Lazzarini, et al., "Linear classification of low-resolution EEG patterns produced by imagined hand movements," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 186–188, 2000.

[4] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[5] N. Birbaumer, A. Kubler, N. Ghanayim, et al., "The thought translation device (TTD) for completely paralyzed patients," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 190–193, 2000.

[6] B. Blankertz, G. Curio, and K.-R. Muller, "Classifying single trial EEG: towards brain computer interfacing," in *Neural Information Processing Systems (NIPS '01)*, T. G. Diettrich, S. Becker, and Z. Ghahramani, Eds., vol. 14, Vancouver, BC, Canada, pp. 157–164, December 2001.

[7] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[8] M. P. S. Brown, W. N. Grundy, D. Lin, et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 1, pp. 262–267, 2000.

[9] T. M. Cover, "The best two independent measurements are not the two best," *IEEE Trans. Syst., Man, Cybern.*, vol. 4, no. 1, pp. 116–117, 1974.

[10] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[11] F. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of niques for large scale feature selection," in *Pattern Recognition in Practice IV: Multiple Paradigms, Comparative Studies, and Hybrid Systems*, E. S. Gelsema and L. N. Kanal, Eds., pp. 403–413, Vlieland, The Netherlands, June 1994.

[12] 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 Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 141–144, 2003.

[13] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison Wesley, Reading, Mass, USA, 1989.

[14] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal ," *IEEE Trans. Geosci. Remote Sensing*, vol. 26, no. 1, pp. 65–74, 1988.

[15] C. Guerra-Salcedo and D. Whitley, "Genetic approach to feature selection for ensemble creation," in *Proc. Genetic and Evolutionary Computation Conference (GECCO '99)*, pp. 236–243, Orlando, Fla, USA, July 1999.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1157–1182, 2003.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[18] D. J. Hand, *Discrimination and Classification*, John Wiley & Sons, New York, NY, USA, 1981.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY, USA, 2001.

[20] J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, Mich, USA, 1975.

[21] D. R. Hundley, M. J. Kirby, and M. Anderle, "Blind source separation using the maximum signal fraction approach," *Signal Processing*, vol. 82, no. 10, pp. 1505–1508, 2002.

[22] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, New York, NY, USA, 2001.

[23] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[24] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 2, pp. 153–158, 1997.

[25] T. Joachims, "Text categorization with support vector machines," in *Proc. 10th European Conference on Machine Learning (ECML '98)*, pp. 137–142, Chemnitz, Germany, April 1998.

[26] T.-P. Jung, S. Makeig, M. J. McKeown, A. J. Bell, T.-W. Lee, and T. J. Sejnowski, "Imaging brain dynamics using independent component analysis," *Proc. IEEE*, vol. 89, no. 7, pp. 1107–1122, 2001.

[27] T. P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T. J. Sejnowski, "Analysis and visualization of single-trial event-related potentials," *Human Brain Mapping*, vol. 14, no. 3, pp. 166–185, 2001.

[28] M. J. Kirby and C. W. Anderson, "Geometric analysis for the characterization of nonstationary time-series," in *Perspectives and Problems in Nonlinear Science: A Celebratory Volume in Honor of Larry Sirovich*, E. Kaplan, J. Marsden, and K. R. Sreenivasan, Eds., chapter 8, Springer Applied Mathematical Sciences Series, Springer, New York, NY, USA, pp. 263–292, March 2003.

[29] J. N. Knight, *Signal Fraction Analysis and Artifact Removal in EEG*, Department of Computer Science, Colorado State University, Fort Collins, Colo, USA, 2003.

[30] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[31] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognition*, vol. 33, no. 1, pp. 25–41, 2000.

[32] T. N. Lal, M. Schroder, T. Hinterberger, et al., "Support vector channel selection in BCI," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1003–1010, 2004.

[33] J. Ma, Y. Zhao, and S. Ahalt, *OSU SVM Classifier Matlab Toolbox*, Ohio State University, Columbus, Ohio, USA, 2002.

[34] S. Makeig, M. Westerfield, T.-P. Jung, et al., "Functionally independent components of the late positive event-related potential during visual spatial attention," *The Journal of Neuroscience*, vol. 19, no. 7, pp. 2665–2680, 1999.

[35] L. A. Miner, D. J. McFarland, and J. R. Wolpaw, "Answering questions with an electroencephalogram-based brain-computer interface," *Archives of Physical Medicine and Rehabilitation*, vol. 79, no. 9, pp. 1029–1033, 1998.

[36] P. L. Nunez, R. Srinivasan, A. F. Westdorp, et al., "EEG coherency I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales," *Electroencephalography and Clinical Neurophysiology*, vol. 103, no. 5, pp. 499–515, 1997.

[37] W. D. Penny, S. J. Roberts, E. A. Curran, and M. J. Stokes, "EEG-based communication: a pattern recognition approach," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 214–215, 2000.

[38] D. A. Peterson and M. H. Thaut, "Model and feature selection in microarray classification," in *Proc. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '04)*, pp. 56–60, La Jolla, Calif, USA, October 2004.

[39] G. Pfurtscheller, C. Neuper, C. Guger, et al., "Current trends in Graz brain-computer interface (BCI) research," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 216–219, 2000.

[40] M. Pontil and A. Verri, "Support vector machines for 3D object recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 6, pp. 637–646, 1998.

[41] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[42] B. Raman and T. R. Ioerger, "Enhancing learning using feature and example selection," Tech. Rep. Departement of Computer Science, Texas A & M University, College Station, Tex, USA.

[43] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1371–1382, 2003.

[44] A. J. Smola, B. Schölkopf, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[45] B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds., *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, Mass, USA, 1999.

[46] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognition Letters*, vol. 10, no. 5, pp. 335–347, 1989.

[47] A. C. Tang and B. A. Pearlmutter, "Independent components of magnetoencephalography: localization and single-trial response onset detection," in *Magnetic Source Imaging of the Human Brain*, L. Kaufman and Z. L. Lu, Eds., pp. 159–201, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2003.

[48] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[49] R. Vigario, J. Sarela, V. Jousmaki, M. Hamalainen, and E. Oja, "Independent component approach to the analysis of EEG

and MEG recordings," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 5, pp. 589–593, 2000.

[50] J. Weston, A. Elisseeff, B. Schölkopf, and M. E. Tipping, "Use of the zero-norm with linear models and kernel methods," *Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1439–1461, 2003.

[51] L. D. Whitley, J. R. Beveridge, C. Guerra-Salcedo, and C. R. Graves, "Messy genetic algorithms for subset feature selection," in *Proc. 7th International Conference on Genetic Algorithms (ICGA '97)*, T. Baeck, Ed., pp. 568–575, Morgan Kaufmann, East Lansing, Mich, USA, July 1997.

[52] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[53] J. R. Wolpaw, D. J. McFarland, G. W. Neat, and C. A. Forneris, "An EEG-based brain-computer interface for cursor control," *Electroencephalography and Clinical Neurophysiology*, vol. 78, no. 3, pp. 252–259, 1991.

[54] J. R. Wolpaw, D. J. McFarland, and T. M. Vaughan, "Brain-computer interface research at the Wadsworth center," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 222–226, 2000.

[55] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," in *Feature Extraction, Construction and Selection: A Data Mining Perspective*, H. Liu and H. Motoda, Eds., pp. 117–136, Kluwer Academic, Boston, Mass, USA, 1998.

[56] E. Yom-Tov and G. F. Inbar, "Feature selection for the classification of movements from single movement-related potentials," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 10, no. 3, pp. 170–177, 2002.

David A. Peterson is a Ph.D. candidate in the Computer Science Department at Colorado State University (CSU) and part of the Cognitive Neuroscience Group affiliated with CSU's Center for Biomedical Research in Music. He received a B.S. degree in electrical engineering and a B.S. degree in finance from the University of Colorado at Boulder. He did business data network consulting for Accenture (previously Andersen Consulting) prior to returning to academia. His research is on biomedical applications of machine learning, with an emphasis on classification and feature selection. He has published research in areas as diverse as mammalian taste coding, brain oscillations associated with working memory, and the interaction of model and feature selection in microarray classification. His current interests are in cognitive, EEG-based brain-computer interfaces and the influence of rhythmic musical structure on the electrophysiology of verbal learning.

James N. Knight is currently a Ph.D. student at Colorado State University. He received his M.S. degree in computer science from Colorado State University and his B.S. degree in math and computer science from Oklahoma State University. His research areas include signal processing, reinforcement learning, high-dimensional data modeling, and the application of Markov chain Monte Carlo methods to problems in surface chemistry.

Michael J. Kirby received the B.S. degree in mathematics from MIT (1984), the M.S. degree (1986) and Ph.D. degree (1988) both from the Division of Applied Mathematics, Brown University. He joined Colorado State University in 1989 where he is currently a Professor of mathematics and computer science. He was an Alexander Von Humboldt Fellow (1989–1991) at the Institute for Information Processing, University of Tuebingen, Germany, and received an Engineering and Physical Sciences Research Council (EPSRC) Visiting Research Fellowship (1996). He received an IBM Faculty Award (2002) and the Colorado State University, College of Natural Sciences Award for Graduate Student Education (2002). His interests are in the area geometric methods for modeling large data sets including algorithms for the representation of data on manifolds and data-driven dimension estimation. He has published widely in this area including the textbook *Geometric Data Analysis* (2001), Wiley & Sons.

Charles W. Anderson received the B.S. degree in computer science in 1978 from the University of Nebraska, and the M.S. and Ph.D. degrees in computer science in 1982 and 1986, respectively, from the University of Massachusetts, Amherst. From 1986 through 1990, he was a Senior Member of Technical Staff at GTE Labs in Waltham, Mass. He is now an Associate Professor in the Department of Computer Science at Colorado State University in Fort Collins, Colo. His research interests are in neural networks for signal processing and control. Specifically, he is currently working with medical signals and images and with reinforcement learning methods for the control of heating and cooling systems. Additional information can be found at http://www.cs.colostate.edu/~anderson.

Michael H. Thaut is a Professor of neurosciences and the Chair of the Department of Music, Theatre, and Dance at Colorado State University. He is also the Head of the Center for Biomedical Research in Music. His research focuses on rhythm perception and production and its application to movement rehabilitation in trauma, stroke, and Parkinson's patients. Recent expansion of his research agenda includes applications of the rhythmic structure of music to cognitive rehabilitation in multiple sclerosis. He received his Ph.D. degree in music from Michigan State University and holds degrees in music from the Mozarteum in Salzburg, Austria, and psychology from Muenster University in Germany. He has served as a Visiting Professor of kinesthesiology at the University of Michigan, a Visiting Scientist at Duesseldorf University Medical School, and a Visiting Professor at Heidelberg University. The author and coauthor of primary textbooks in music therapy, his works have appeared in English, German, Italian, Spanish, Korean, and Japanese.