

Evaluation of Feature Selection Techniques for Analysis of Functional MRI and EEG

Lauren S. Burrell, Otis L. Smart, George Georgoulas, Eric Marsh, and George J. Vachtsevanos

Abstract—The application of feature selection techniques greatly reduces the computational cost of classifying high-dimensional data. Feature selection algorithms of varying performance and computational complexities have been studied previously. This paper compares the performance of classical sequential methods, a floating search method, and the “globally optimal” branch and bound algorithm when applied to functional MRI and intracranial EEG to classify pathological events. We find that the sequential floating forward technique outperforms the other methodologies for these particular data. Previous works have found branch and bound to be a superior feature subset selection technique; however, in this application, the branch and bound algorithm fails to create subsets with better classification accuracy.

Index Terms—Pattern classification, sequential feature selection, branch and bound, fMRI, EEG

I. INTRODUCTION

Feature selection is an integral preliminary step in mining large datasets and in any pattern classification problem. By reducing the dimensionality of the data, the computational burden can be greatly decreased. The goal of feature selection is to reduce the number of features to be processed without sacrificing class discrimination, and therefore classification accuracy.

Feature subset selection algorithms can be divided into three categories: exponential, randomized, and sequential [1]. In exponential search algorithms such as exhaustive search and branch and bound (B&B), the number of subsets grows exponentially with the dimensionality of the search space. Sequential algorithms have reduced computational complexity but tend to become trapped in local minima due to the so-called nesting effect. Sequential forward and backward selection, Plus- l Minus- r selection, and sequential floating selection are examples of such methods. Randomized search methods try to avoid the problem of local minima by adding randomness to the search. Genetic algorithms and simulated annealing fall into this category.

The main objective of this paper is to compare techniques for creating feature subsets to be used in classifying biological signals, namely intracranial electroencephalograms

(iEEG) and functional magnetic resonance imaging (fMRI) data. In particular, iEEG and fMRI have become very useful in medicine to study epilepsy because they have the ability to detect and localize epileptic activity. However, in epilepsy literature there is no state of the art for selecting features for the inherent pattern classification problem in these data, i.e. determining which areas of the brain are dysfunctional. The sequential algorithms are chosen for this study due to their low computational burden. These suboptimal routines are then measured against an “optimal” search method, B&B, to evaluate the algorithms when applied to a relatively small number of features in fMRI and iEEG epilepsy data. Evolutionary algorithms and other randomized search methods are beyond the scope of this study and will be considered in future work.

II. PRELIMINARIES

A. Data

Functional MRI indirectly measures neural activity through detection of changes in blood flow, blood volume, and oxygen consumption. As regional brain function increases, there is a corresponding increase in regional cerebral blood flow (CBF). During arterial spin labeling (ASL) perfusion scans, arterial blood water near the area of interest is electromagnetically labeled by a radiofrequency pulse to allow tracking of the CBF. Immediately after imaging of the pre-labeled spins, control images without labeled spins are acquired. Through pairwise subtraction of these label and control pairs, the effects of labeling can be determined and CBF images can be created.

Resting perfusion fMRI scans from five temporal lobe epilepsy patients are evaluated in this study. Four of the five patient datasets contain forty minutes of scanning data, while the final contains thirty minutes worth of scans. The patients are scanned at 3 Tesla with a repetition time of three seconds. The images are realigned to account for subject motion during scanning and reduce artifacts [2], smoothed with a Gaussian kernel to increase signal-to-noise ratio [3], and pairwise subtracted to obtain cerebral blood flow (CBF) images. Finally, the images are normalized to the standard Montreal Neurological Institute (MNI) brain template to ensure that all brain images conform to the same space [2]. All realignment, co-registration, normalization, and smoothing are performed with Statistical Parametric Mapping software (SPM2) [2], [4], [5]. These are standard steps in perfusion image preprocessing. The resulting CBF datasets consist of 300 or 400 images depending on the length of the scan.

L.S. Burrell, O.L. Smart, and G. Georgoulas are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA (phone: 404-894-9356; fax: 404-894-4130; email: lsburrell@gatech.edu, gte851r@mail.gatech.edu, ggeorgoulas@mail.gatech.edu).

E. Marsh is with the Division of Neurology, The Childrens Hospital of Philadelphia, University of Pennsylvania School of Medicine, Philadelphia, PA, 19104, USA (email: marshe@email.chop.edu).

G.J. Vachtsevanos is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA (fax: 404-894-7583; email: gjv@ece.gatech.edu).

For this experiment, analysis of the brain images is limited to voxels in the mesial temporal lobes. For each patient, a neurologist marks the temporal lobe from which the epileptic activity is suspected to originate. Using the WFU Pickatlas toolbox for Matlab, normalized masks of the left and right mesial temporal lobes are created [6], [7]. The time course, i.e. voxel intensity as a function of time, is then extracted from each voxel in the regions of interest, as illustrated in Figure 1.

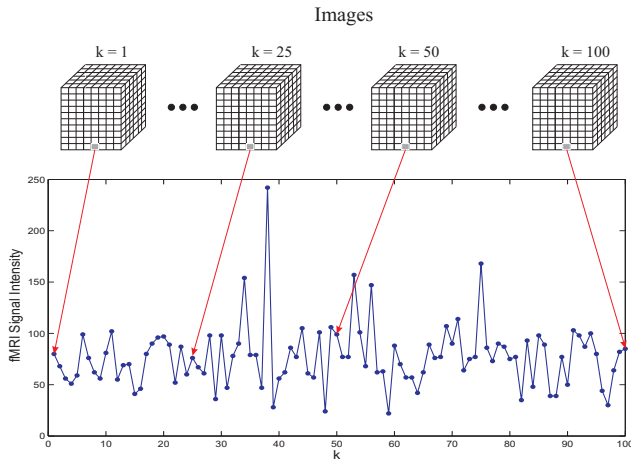


Fig. 1. fMRI voxel time course extraction.

EEG has become a primary diagnostic tool to identify the causes and symptoms of epileptic seizures. The EEG voltage can be considered a random process that is grossly equivalent to a superposition of the extra-cellular potentials of neurons near the EEG electrode. It is suspected that neurons constantly interact in a disorganized manner—as in a brain without epilepsy—before an epileptic seizure begins but recurrently interact in a more synchronous fashion at the onset of a seizure [8]. Surface (or scalp) electrodes, which are the typical use of the EEG, are affixed to the scalp of the epilepsy patient and record electrical activity from the brain at a depth of only about 1 cm below the surface of the brain, while intracranial EEG electrodes are surgically implanted within the brain tissue of epilepsy patients and is reserved for pre-surgical evaluation of epilepsy and safety considerations according to stringent standard protocols [9], [10], [11]. The iEEG provides better spatial resolution and a higher signal-to-noise ratio with fewer artifacts than scalp electrodes. Moreover, study of the iEEG has led to converging evidence that certain abnormal electrographic waveforms (e.g., fast ripples, high frequency epileptiform oscillations) may localize [12], [13], [14] or even predict [15], [16], [17], [18] the onset of epileptic seizures to ultimately guide effective treatments for epilepsy patients. Data recordings from five epilepsy patients are considered in this study. An example of iEEG signals is shown in Figure 2.

For the fMRI and iEEG data, each patient provides informed consent for participation in data collection under the approval of the Internal Review Board.

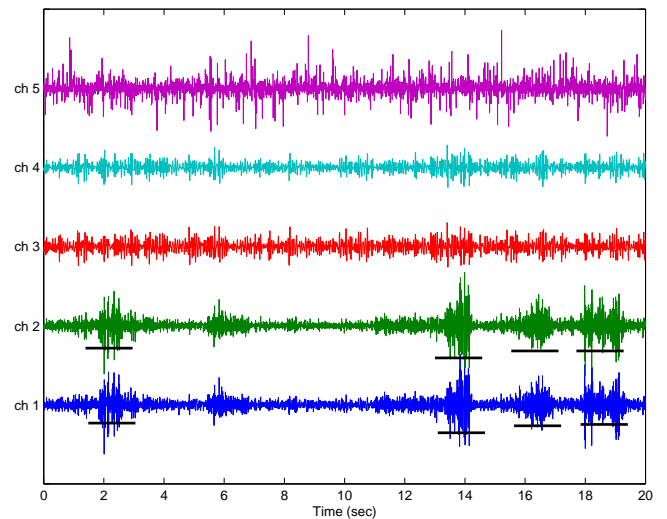


Fig. 2. A 20-second segment of multi-electrode iEEG with pathologic activity (*underscores*) that occurs within intervals of typical activity.

B. Feature Selection

Feature selection algorithms search for the best feature subset that reduces the feature space dimensionality with the smallest loss in classification accuracy. In other words, for a set of D features, the algorithm chooses a subset of size $d < D$, which has the greatest ability to discriminate between classes. The goodness of a particular feature subset is evaluated using an objective function, $J(Y_m)$, where Y_m is a feature subset of size m . The two types of objective functions are filters and wrappers. Filters rate features based on general characteristics, such as interclass distance or statistical independence, without employing any mining algorithms [19]. Wrappers, on the other hand, evaluate subsets based on their predictive accuracy when employing a particular classifier. While filters are advantageous because they execute quickly and yield a more general solution, they tend to create large subsets. Wrappers are slower but typically do not suffer from the problem of overfitting like filters do. They also tend to result in subsets with higher classification accuracy because they are trained to work with a specific classifier.

The main advantage of sequential algorithms is their relatively low computational burden of $O(d^2)$ [1]. A naive approach to feature selection would rank the features and select the top d to create the best subset. However, this procedure overlooks the possibility of features with poor individual objective function values performing better in tandem because of complementary information [19]. Sequential forward selection (SFS) is a greedy search algorithm that determines an “optimal” set of features for extraction by starting from an empty set and sequentially adding a single feature in the superset to the subset if it increases the value of the chosen objective function. Sequential backward selection (SBS) is similar to SFS but works in the opposite direction. The search initializes with the full superset set, X , of D features and removes a single feature that improves (or minimally worsens) an objective function to obtain the

best subset of features. The problem with these sequential approaches is that they gravitate toward local minima due to the inability to re-evaluate the usefulness of features that were previously added or discarded. Pseudocode for SFS and SBS is shown in Figure 3 and Figure 4, respectively. Plus- l minus- r selection attempts to avoid the nesting problem by performing l forward selection steps followed by r backward selections and looping until the desired number of features is found. The drawback of this approach is that the optimal choices for l and r are unknown. Sequential floating selection methods improve on this technique by dynamically determining the best values for l and r so as to maximize $J(Y_m)$. Pseudocode for sequential floating forward selection (SFFS) can be found in Figure 5.

```

1. Initialize feature set
    $Y_0 = \{\emptyset\}; m = 0$ 
2. Select the next best feature
    $x^+ = \arg \max_{x \notin Y_m} [J(Y_m + x)]$ 
3. Update feature set
    $Y_{m+1} = Y_m + x^+$ 
4. While  $m < d$ 
    $m = m + 1$ 
   Go to Step 2

```

Fig. 3. Sequential forward selection pseudocode.

```

1. Initialize feature set
    $Y_0 = X; m = 0$ 
2. Remove the worst feature
    $x^- = \arg \max_{x \in Y_m} [J(Y_m - x)]$ 
3. Update feature set
    $Y_{m+1} = Y_m - x^-$ 
4. While  $m < d$ 
    $m = m + 1$ 
   Go to Step 2

```

Fig. 4. Sequential backward selection pseudocode.

```

1. Initialize feature set
    $Y_0 = \{\emptyset\}; m = 0$ 
2. Find the best feature and update  $Y_m$ 
    $x^+ = \arg \max_{x \notin Y_m} [J(Y_m + x)]$ 
    $Y_m = Y_m + x^+; m = m + 1$ 
3. Find the worst feature
    $x^- = \arg \max_{x \in Y_m} [J(Y_m - x)]$ 
4. If  $J(Y_m - x^-) > J(Y_m)$  then
    $Y_{m+1} = Y_m - x^-; m = m + 1$ 
   Go to Step 3
else
   Go to Step 2

```

Fig. 5. Sequential floating forward selection pseudocode.

Exponential algorithms, such as exhaustive search and B&B, have the potential to find optimal solutions under certain conditions and assumptions. The main drawback of these algorithms is that they have complexity $O(2^d)$ [1]. This high complexity becomes prohibitively expensive for large feature sets. Exhaustive search evaluates every subset of size d to find the optimal one. The B&B algorithm works by constructing a search tree like the one in Figure 6. The chosen objective function must meet the monotonicity condition, which is defined as follows:

$$J(Y_s) \geq J(Y_t) \text{ for any } Y_s \supseteq Y_t$$

This monotonicity requirement allows many objective function evaluations to be avoided without missing the global optimum [20]. Starting from the root of the tree (top node), $J(Y)$ is computed at each node as the tree is traversed. When a node (feature subset Y) is found that has an objective function value lower than the current bound, i.e. the maximum $J(Y)$ so far for a leaf node, that branch of the tree is eliminated from the search, thereby potentially reducing the total number of computations. The monotonicity requirement ensures that the values of the leaf nodes of that branch cannot be better than the current bound, so those nodes need not be evaluated.

III. METHODS

A set of features, twelve for the fMRI data and fourteen for the iEEG, is extracted from each patient dataset. All features are mathematical expressions inspired by several analysis domains (e.g., time, frequency, statistics, information theory). For a single feature of an iEEG time-series, feature extraction involves a moving window of length W that causally slides across the signal and evaluates a function of the signal, which describes the feature to extract. The procedure slides the window T points in time, processing a new interval of data with the feature in each shift and ultimately returning a feature vector. For each voxel in an fMRI image, a window size equal to the length of the time course is used to create a feature vector with a single feature value for each time course. In addition, a binary classification vector for the fourteen (twelve) feature vectors of the iEEG (fMRI) data is constructed in which a zero corresponds to feature values of non-epileptic activity and a value of one corresponds to epileptic activity.

Each feature selection algorithm in Section II.B is executed for varying feature subset sizes with the objective function, feature vectors, and classification vectors as the algorithm inputs. Since the features are to ultimately be used in a pattern detection system, the classification accuracy defined by Equation (1) is taken as the objective function for each feature selection method. A k -nearest-neighbor (k -NN) rule is used as the classifier in the system where five neighbors ($k = 5$) is chosen.

In the following expression for the classification accuracy, which is the number of correctly classified events normalized

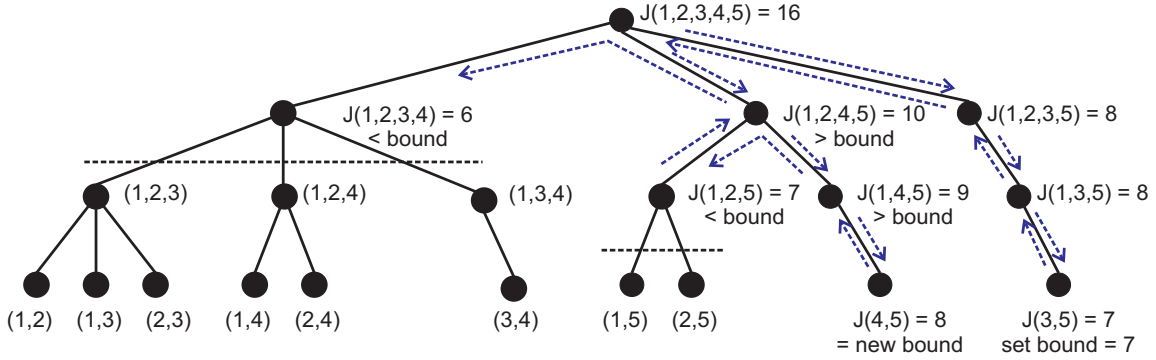


Fig. 6. Example of a branch and bound search tree for $D = 5$ and $d = 2$. The dashed black lines represent the points at which branches can be pruned in order to reduce the computational burden without loss of optimality, and the arrows represent the order of tree traversal starting from the right.

by the total number of events,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

the values TP, TN, FP, and FN are numbers of true positives, true negatives, false positives, and false negatives, respectively. Positives indicate epileptic activity, and negatives denote non-epileptic activity.

For each set of data, using the classification accuracy as an objective function necessitates separating the feature vectors and classification vector into training and testing sets for the k -NN. The counts for TP, TN, FP, and FN are determined after a testing set is classified and compared to the corresponding a priori classification vector of the testing set.

In the training set, a balanced proportion of values from the epileptic (1) and non-epileptic (0) classes is used because training a classifier with a balanced set provides a “guaranteed” measure of future classification accuracy, whereas training an unbalanced classifier may prove much better or much worse if the testing set is unbalanced [21]. A 95% confidence interval for the accuracy is approximated using Equation (2) after five Monte Carlo simulations are run. Each simulation randomly draws a different training and testing data set, which is used for each selection method, before the accuracy is calculated and recorded. Five simulations are considered sufficient because of the low variance in the recorded accuracies (Figures 7 and 8). In Equation (2), \bar{x} , s , and n are the mean, standard deviation, and number of the recorded accuracies, respectively.

$$CI_{95\%} = \bar{x} \pm 2 \cdot \frac{s}{\sqrt{n}} \quad (2)$$

IV. RESULTS

Figures 7 and 8 illustrate classification accuracies greater than 90% (80%) in patients with high (low) signal to noise ratio in the fMRI and iEEG data, respectively. The error bars denote the confidence interval in the computed accuracy for the number of selected features. Comparing the classification accuracy of the feature selection algorithms across subjects

for each the fMRI and iEEG signal, it is clear that SFFS is the best technique. Figure 7 shows that the performance of the selection algorithms is comparable, but SFFS slightly edges all the alternatives for each patient with SFFS being the next most accurate choice. The similarity in the accuracies is attributed to the features used to process the fMRI data, which share difficulty in separating functional and dysfunctional classes. On the other hand, Figure 8 illustrates that SFFS is considerably better in all patients except the following two: 1) a patient with high signal-to-noise ratio (Patient E03), for which only a few features are needed—as noted by the decreasing trend as the number of features increases; and 2) a patient with poor signal-to-noise ratio (Patient E09), for which the B&B algorithm can achieve the best classification accuracy, but only if 13 of the 14 features are used.

The figures reflect that the SFFS often chooses 6-8 features (approximately half or more of the features in each dataset) to achieve the maximum expected classification accuracy in each set of data, necessitating a means to fuse the selected features for further reduction in computational complexity. Furthermore, it is important to state that the subset of features selected by SFFS did not necessarily include the top individually ranked features and varied from patient to patient in size and elements of the set.

As a final note, it is interesting to find that the B&B algorithm typically does not outperform any of the other methods in both the fMRI and iEEG data, despite claims about the optimality of the algorithm in other works [20]. However, this observation is not surprising when the following notions are considered: 1) the No Free Lunch Theorem [22] cautions against generalizing the supposed superiority of an approach in one application across other domains of analysis; 2) a flaw in the search procedure of the B&B algorithm is that it may prune a branch due to a lower classification accuracy compared to the current bound although the accuracy could increase again at the leaf nodes if the monotonicity property is violated; 3) an advantage in SFFS is that each iteration of the algorithm provides an opportunity to increase or decrease the accuracy by adding or subtracting features from an intermediate subset.

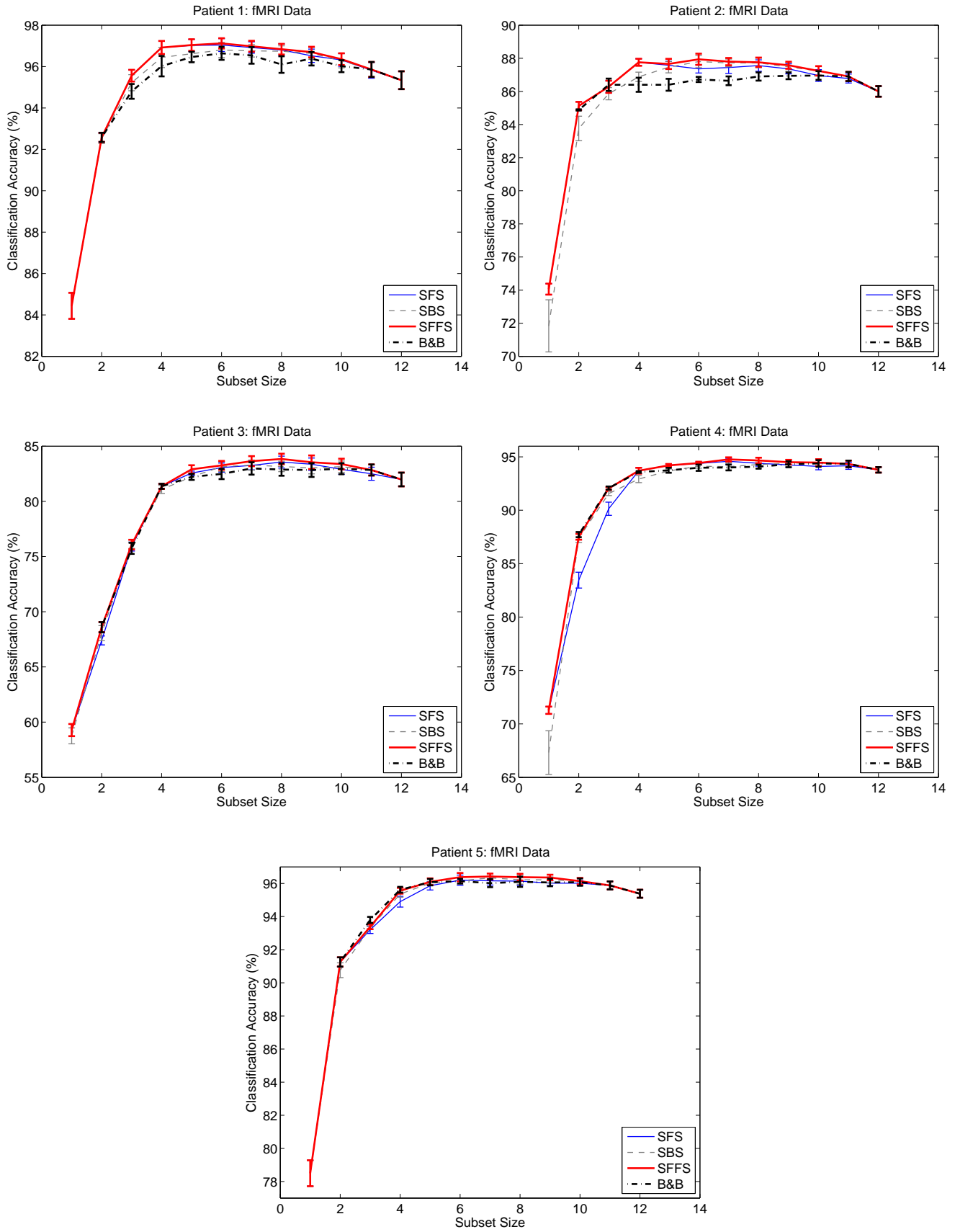


Fig. 7. Classification accuracy as a function of subset size for the fMRI data.

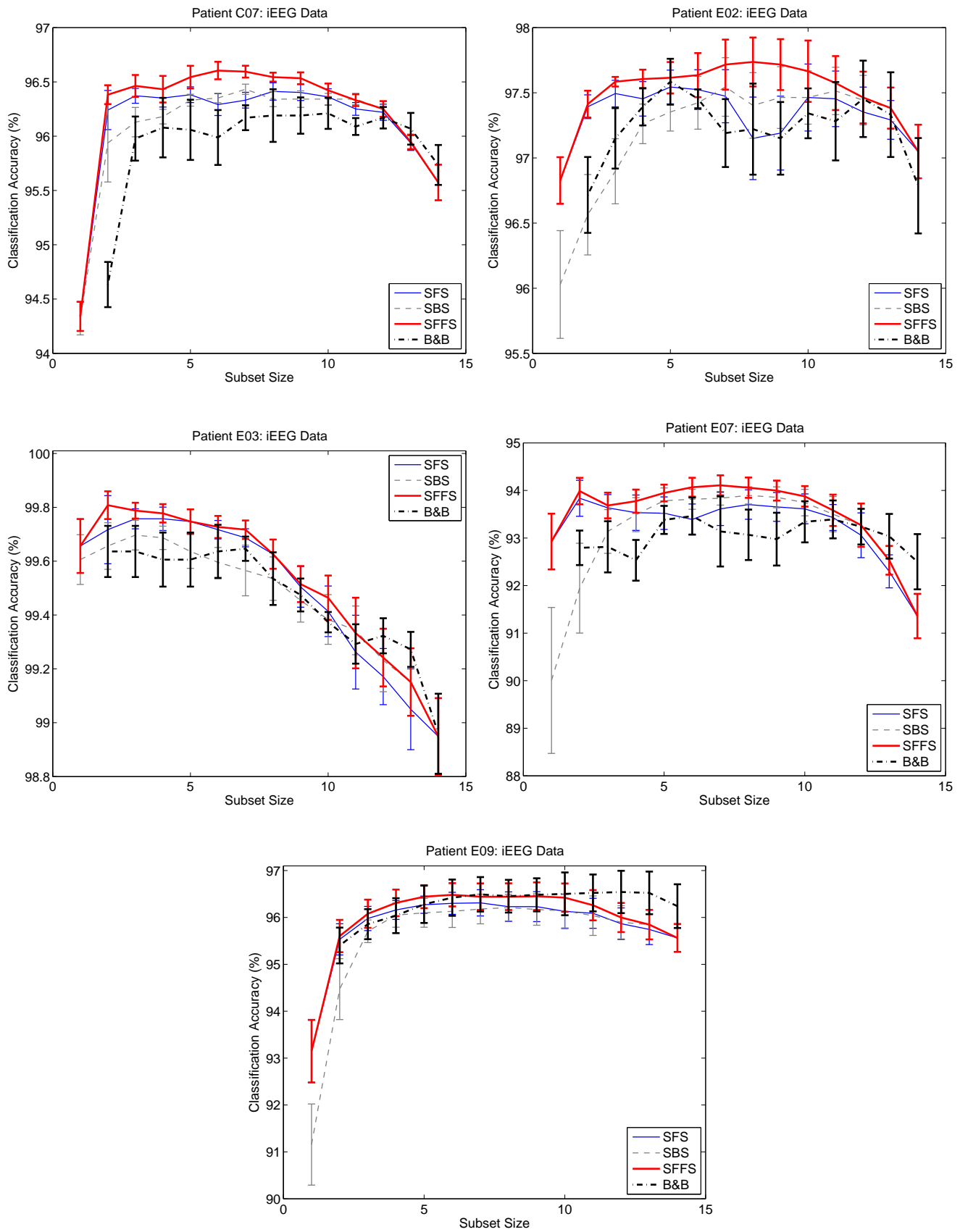


Fig. 8. Classification accuracy as a function of subset size for the iEEG data.

V. CONCLUSIONS

This study finds that the choice of an algorithm for feature selection is important in analyzing fMRI or iEEG data that records activity from epilepsy patients. In terms of classification accuracy, the SFFS algorithm proves to be the best option for the automatic selection of features that discern functional and dysfunctional activity in both fMRI and iEEG. For the fMRI data, SFS is a reasonable alternative that at times marginally sacrifices classification accuracy for a smaller subset of features; and for the iEEG data, either SFS or SBS are viable alternatives for the same reason. Furthermore, the results of this work contradict that claim in several sources that the B&B algorithm is an optimal search algorithm for feature selection, but resonates with the concept of the No Free Lunch Theorem to suggest that B&B may be optimal for selecting features in applications besides the analysis of fMRI and iEEG signals from patients with epilepsy. More importantly, the presented work demonstrates that classical feature selection can be successfully applied to pattern classification problems involving fMRI and iEEG for epilepsy patients.

ACKNOWLEDGMENT

The authors would like to thank the Center for Functional Neuroimaging at the University of Pennsylvania, Children's Hospital of Philadelphia, and the Hospital of the University of Pennsylvania for providing data for this investigation.

REFERENCES

- [1] J. Doak, *An Evaluation of Feature Selection Methods and Their Application to Computer Security (Tech. Rep. CSE-92-18)*. University of California at Davis, 1992.
- [2] R. S. J. Frackowiak, K. J. Friston, C. Frith, R. Dolan, C. J. Price, S. Zeki, J. Ashburner, and W. D. Penny, *Human Brain Function*, 2nd ed: Academic Press, 2003.
- [3] J. Wang, Z. Wang, G. K. Aguirre, and J. A. Detre, "To smooth or not to smooth? ROC analysis of perfusion fMRI data," *Magnetic Resonance Imaging*, vol. 23, pp. 75-81, 2005.
- [4] K. J. Friston, J. Ashburner, C. D. Frith, J. B. Poline, J. D. Heather, and R. S. J. Frackowiak, "Spatial registration and normalization of images," *Human Brain Mapping*, vol. 3, pp. 165-189, 1995.
- [5] K. J. Friston, A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Human Brain Mapping*, vol. 2, pp. 189-210, 1994.
- [6] J. A. Maldjian, P. J. Laurienti, and J. H. Burdette, "Precentral gyrus discrepancy in electronic versions of the Talairach atlas," *NeuroImage*, vol. 21, pp. 450-455, 2004.
- [7] J. A. Maldjian, P. J. Laurienti, R. A. Kraft, and J. H. Burdette, "An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets," *NeuroImage*, vol. 19, pp. 1233-1239, 2003.
- [8] M. R. Mehta, C. Dasgupta, and G. R. Ulal, "Kindling of focal epilepsy occurs due to learning: a neural network model," *International Journal of Neural Systems*, vol. 6, pp. 107-111, 1995.
- [9] J. Engel, ed. "Surgical Treatment of the Epilepsies", 1st ed. New York: Raven Press, 1987, vol. 1.
- [10] L. F. Quesney, "Intracranial EEG Investigation in Neocortical Epilepsy," *Advanced Neurology*, vol. 84, pp. 253-74, 2000.
- [11] L. F. Quesney, M. Constain, T. Rasmussen, A. Olivier, and A. Palmieri "Presurgical EEG Investigation in Frontal Lobe Epilepsy," *Epilepsy Research Supplement*, vol. 5, pp. 55-69, 1992.
- [12] G. A. Worrell, Landi Parish, Stephen D. Cranstoun, Rachel Jonas, Gordon Baltuch, and Brian Litt "High Frequency Oscillations and Seizure Generation in Neocortical Epilepsy," *Brain*, vol. 127, pp. 1-11, April 2004.
- [13] A. Bragin, C. Wilson, J. Almajano, I. Mody, I., and J. Engel Jr., "High Frequency Oscillations after Status Epilepticus: Epileptogenesis and Seizure Genesis," *Epilepsia*, vol. 45, no. 9, pp. 1017-1023, 2004.
- [14] A. Bragin, I. Mody, C. L. Wilson, and J. Engel, Jr., "Local generation of fast ripples in epileptic brain," *Journal of Neuroscience*, vol. 22, no. 5, pp. 2012-21, 2002.
- [15] B. Litt, R. Esteller, J. Echaz, R. Shor, C. Bowen, and G. Vachtsevanos, "Pre-Ictal Prodromes Predict Seizures in a Patient With Mesial Temporal Lobe Epilepsy," in Proc. *American Epilepsy Society Meeting*. Orlando, Florida, December, 1999.
- [16] B. Litt and J. Echaz, "Prediction of Epileptic Seizures," *The Lancet Neurology*, vol. 1, no. 1, pp. 22-30, May 2002.
- [17] H. Witte, L. D. Iasemidis, and B. Litt, "Special Issue on Epileptic Seizure Prediction," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 5, pp. 537-539, May 2003.
- [18] B. Litt, R. Esteller, J. Echaz, M. D'Alessandro, R. Shor, and T. Henry, "Epileptic Seizures May Begin Hours in Advance of Clinical Onset: A Report of Five Patients," *Neuron*, vol. 30, pp. 51-64, April 2001.
- [19] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.
- [20] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, pp. 900-912, 2004.
- [21] J. Echaz, "Classifier-Based Performance Metrics and the Effects of Prior Probability Mismatches," IntelliMedix, Inc., Technical Report, March 2000.
- [22] R. Duda, P. Hart, and D. Stork, *Pattern Classification (2nd Edition)*: Wiley-Interscience, 2000.