Mutual Information Iterated Local Search: A Wrapper-Filter Hybrid for Feature Selection in Brain Computer Interfaces

Jason Adair [0000-0003-0198-9095], Alexander E. I. Brownlee [0000-0003-2892-5059], and Gabriela Ochoa [0000-0001-7649-5669]

Computing Science and Mathematics, University of Stirling, Stirling, United Kingdom
{jason.adair, alexander.brownlee, gabriela.ochoa}@stir.ac.uk

Abstract. Brain Computer Interfaces provide a very challenging classification task due to small numbers of instances, large numbers of features, non-stationary problems, and low signal-to-noise ratios. Feature selection (FS) is a promising solution to help mitigate these effects. Wrapper FS methods are typically found to outperform filter FS methods, but reliance on cross-validation accuracies can be misleading due to overfitting. This paper proposes a filter-wrapper hybrid based on Iterated Local Search and Mutual Information, and shows that it can provide more reliable solutions, where the solutions are more able to generalise to unseen data. This study further contributes comparisons over multiple datasets, something that has been uncommon in the literature.

Keywords: Brain Computer Interface, Mutual Information, Evolutionary Search, Iterated Local Search

1 Introduction

Brain Computer Interfaces (BCI) allow neurological signals to be decoded to enable the control of external devices. The most common mode of recording signals for use in such devices is electroencephalography (EEG). This involves the placement of electrodes on the scalp of a user, to record the electrical activity within the underlying brain region. With these signals, practitioners can analyse the working state of the patient's brain to detect seizure triggers, sleep patterns, or even allow the control of external devices — a life-altering application for people in need of prosthetic limbs, or with muscle-degenerative disorders.

EEG-based devices are safer, more accessible, and more cost-effective than invasive techniques, but they do come with substantial caveats: enough energy must be produced by the brain region to pass through three centimetres of bone and soft tissue. This requires at least 6cm^2 of neural material to be active, greatly reducing the spatial resolution of the signal. The problem is further compounded by multiple sources of noise: eye movements, muscle contractions, and cardiac

rhythms. Pre-processing the data by band pass filtering, and using a technique known as *Feature Extraction* can help emphasize characteristics in the data that are useful for constructing effective, predictive BCI models. However, as BCI datasets typically consist of a large number of variables, with few instances, it is prudent to reduce the feature set to avoid overfitting, decrease training times, and remove noisy or redundant features.

Filter methods rank features according to a statistical measure, which creates generalised models, but fail to exploit the nature of the machine learning algorithm intended for use. Wrapper methods use the classifier as a 'black-box' evaluation of feature subsets, achieving high classification accuracies, but are prone to over-fitting on training sets. We propose a hybrid of both categories of feature selection algorithms: filters and wrappers. This hybridisation is defined here as *Minimal Redundancy Maximal Relevance Iterated Local Search (MRMR-ILS)*; a metaheuristic known as Iterated Local Search that utilises a mutual information measure to guide the perturbation operator, while allowing a normal wrapper-based local search heuristic.

The aim of this study is to provide a new hybrid-heuristic that is capable of finding small feature subsets that generalise more effectively than typical wrapper approaches, and are more accurate than those found by filter approaches. Specifically, the contributions of this study are:

1. A new filter-wrapper hybrid combining Iterated Local Search with the Minimum Redundancy Maximal Relevance mutual information measure.
2. Results based on three different datasets, originating from three different motor-imagery based problems.
3. Analysis of interactive effects between mutual information measures, error rates obtained from training sets, and the predictive accuracy on unseen data.

This paper is structured in the following manner: a general background to BCI based feature selection is given in Section 2, with interest to mutual information, wrapper methods, and their hybrids. This is followed by our proposed algorithm in Section 3 and the methodology used to evaluate it in Section 4. We then present our results and discussion in Section 5, and conclusion in Section 6.

2 Background

Extra-cranial BCI recordings are notoriously noisy; while we are only interested in the energy generated from the neurons that correspond to the task at hand, the signals also contain unrelated neural processes, muscle movements, and other sources of information that can negatively impact the performance of our classifiers. This can render data recorded from some frequencies, or even entire channels, redundant to our needs. Selecting the data to retain, and what to disregard, is a non-trivial task. However, obtaining near optimal feature subsets reduces the

dimensionality of the data, decreases the training and prediction time costs, creates simpler models, and increases the predictive accuracy [1]. Feature Selection algorithms can be divided into three groups: Filter, Embedded, and Wrapper methods.

2.1 Filters

Filter based methods rank variables according to a criterion, independently of the classifier. Examples of these performance measures include the Pearson correlation coefficient [2], Fisher score [3], and measures based in Information Theory [4]. The advantages of such techniques tend to be that they are typically less computationally expensive, simpler to implement, and resulting feature subsets are more generalisable as they are not tied to a specific classifier [5]. That being said, they lack the ability to exploit specific characteristics of the machine learning algorithms intended for use, and therefore rarely obtain the highest classification accuracies.

The following concept definitions explain the mutual information aspects of the algorithm presented by this paper.

Entropy is an integral concept within Information Theory, defining the uncertainty of a variable. A key measurement of this is Shannon's entropy [6]:

$$H(X) = -\sum_{x} p(x) \log p(x) \,. \tag{1}$$

Entropy is calculated by the summation of all the probability distributions ($p(x)$) of values (x) of the set X, multiplied by the natural log of those probability distributions.

Mutual Information is the unique information shared between two variables. Using entropy, it is possible to quantify the conveyable information from a variable, however, what is often of interest, is how much variables 'overlap' in what they have recorded. This is especially useful when we want to consider how effective one variable is at predicting another; higher shared information suggests that they are measuring a similar source of information:

$$I(X : Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y) \,. \tag{2}$$

To do this, we consider how much information is conveyed by each variable as individuals, in comparison with how much information is conveyed when they are paired.

Maximum Relevance Minimal Redundancy Mutual Information can detect even non-linear interactions between variables, but it is limited due to it being a univariate approach. This is a source of weakness in applications such

as feature selection, as we frequently find multivariate interactions between variables and their labels. To solve this, Peng, Long and Ding [7] proposed the minimal-redundancy-maximal-relevance approach. It seeks to address two conditions; maximisation of selected features *Relevance*, and minimisation of their *Redundancy*:

$$maxD(S,c), D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c)\,. \tag{3}$$

where $I(x_i; c)$ is the mutual information between each selected feature (x_i) in the subset (S) and the class (c).

$$minR(S), D = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j)\,. \tag{4}$$

where $I(x_i; x_j)$ is the mutual information between each pair of selected features within the selected subset (S).

$$max\Phi(D,R), Phi = D - R\,. \tag{5}$$

MRMR seeks to maximise the distance between the Relevance (D) and Redundancy (R).

2.2 Wrappers

Wrapper methods select feature subsets by utilising the classifier as a 'black box' fitness function. By iteratively evaluating feature subsets using the machine learning algorithm to make predictions on the training set, models created from the selected subset can be tested on their validity. The simplest, and one of the most common forms of this in feature selection is *Sequential Forward Selection* (SFS). Every feature is used to train a model, and then used to make predictions on the available instances. The one with the highest predictive accuracy is selected as the first feature in the subset. Each following feature is found by appending the existing subset, and evaluating it as before [8]. Simple heuristics such as SFS are successful, but rarely provide state-of-the-art results due to their inability to detect feature interactions; when a feature is added, no consideration is given to previously selected features and their counter-dependencies. More complex heuristics such as Genetic Algorithms [9], Particle Swarm Optimisation [10], and Ant Colony Optimisation have been widely used in Brain Computer Interface feature selection with great success [11].

2.3 Hybrid Approaches

A relatively uncommon approach in BCI is the combination of filters and wrappers in hybrid methods. A common form of this is a two-stage approach: a filter method is first applied to remove the most redundant features, before a wrapper is applied to the remaining features. A variation of this is seen in Gan [12], where *Sequential Forward Floating Search (SFFS)* was combined with *MRMR*

by using the mutual information approach to select a set of candidate features for addition and removal at each phase. This reduced the computational training cost of utilising the classifier across all the candidate features. *Ant Colony Optimisation* was combined with *Differential Evolution* in Khushaba *et al* [13]. This technique used a mutual information evaluation function as the Selection Measure in ACO, and evaluated each of the ants using a Linear Discriminate classifier.

In other feature selection fields, hybridised approaches involving mutual information are somewhat more prevalent. Mutual information was used to reduce the search space in advance of running a *Genetic Algorithm* in Tan, Fu, Zhang, and Bourgeois [14] and *Particle Swarm Optimisation* in Ali and Shahzad [15]. It has also been successfully used within memetic algorithms as a local search method to refine the solutions found by PSO in Particle Swarm Optimisation Backwards Elimination (PSOBE) [16] and in Genetic Algorithms [17]. A common observation however, is that mutual information is almost always used as a local search operator in these cases, and to our knowledge, has not been used in the explorative phase of a metaheuristic prior to this paper.

3 Proposed Method

Here we introduce the existing Iterated Local Search (ILS) algorithm, followed by our contribution, the MRMR-ILS.

3.1 Iterated Local Search

Iterated Local Search is an iterative search based algorithm that has demonstrated interesting results across a variety of domains [18], but with almost no application to the BCI domain. The ILS used in this paper consists of a layered search; a local search, in the form of a hillclimber, and a diversification mechanism, in the form of a strong mutation, known as a perturbation. A solution is either randomly generated or provided to the algorithm. A hillclimber is then used to search the local space; a candidate solution is created by performing a single point mutation on the current solution. This is achieved by randomly choosing one of the selected features in the current solution, and replacing it with an unselected feature. This is then evaluated by performing 10-fold cross-validation using the training set and obtaining the average prediction error rates on each of the folds.

Cross-validation is a technique used to assess algorithm performance when using limited quantities of data. K-Fold cross-validation is commonly used within the machine learning community to evaluate the performance of models. To do this, training data is subdivided into K sets. K-1 sets are used to train a classifier, and is then used to predict the labels of the 'left out' set and its error rate is measured. This is then performed on all K sets, and the average error rate returned.

3.2 Minimal Redundancy Maximal Relevance-Iterated Local Search

In the Mutual Information-based Iterated Local Search (*MRMR-ILS*) algorithm proposed by this paper, the stochastic perturbation stage of the ILS is replaced by an information-measure based selection process. Instead of randomly selecting features for replacement, features are selected for retention based on the information they share with each other, and the label. The mRMR score for each feature is calculated and those that score most highly, that is, those that have the highest relevance with the label, and lowest information overlap with other features within the selected solution, are retained. The remaining features are randomly replaced with unselected features.

4 Methodology

The experimental methodology is presented in the following order; dataset descriptions, feature extraction method, solution size, classification algorithms used, fitness function, search algorithm parametres, and benchmark methods for comparison.

4.1 Datasets

The datasets provided by the Berlin Brain Computer Interface Competitions have been some of the most prevalent in literature over the past few years. Two of these datasets were used in this paper; Berlin BCI competition II, datasets III and IV[1]. Both of these datasets have proven popular in literature due to their challenging, but well-defined, nature. The third dataset was acquired by the RIKEN Centre of Advanced Intelligence Project[2]. It does not appear as frequently in literature as the BCI competitions, but was chosen as diversity is essential to foster amongst the state-of-the-art benchmarks. The following section will describe the paradigms used in each dataset, the conditions of their recording, and any pre-processing steps required before feature extraction.

Dataset A - Berlin BCI competition II Datasets III

Over a set of 280 9-second trails, a participant was asked to imagine left and right hand movements to control an on-screen cursor. Three electrodes were placed on the participants scalp, and a blank screen displayed. The first two seconds were a resting phase, followed by an auditory signal and cross being displayed in the centre of the screen to focus the participant's attention. On the fourth second, the cross became an arrow, signifying the motor-imagery (left or right hand movements) that the participant was required to imagine. Three electrodes placed at C3, C4, and Cz, and sampled at 128Hz. The signal was then bandpass filtered between 0.5 and 30Hz. The first 140 instances were assigned as 'training data', and the remaining 140 as 'testing data'.

[1] `http://www.bbci.de/competition/ii/#datasets`
[2] `http://www.bsp.brain.riken.jp/~qibin/homepage/Datasets.html`

Dataset B - Berlin BCI competition II Datasets IV
A set of 28 EEG electrodes were used to record a single subject during a self-paced finger movement task. The participant was asked to sit at a computer with their hands in a typical position at the keyboard. The participant was then allowed to press keys at a rate of one per second, in a self-determined order. In total, 416 instances were collected; 316 of which were designated as training, and 100 were provided, unlabelled, as testing data. This results in 416 instances of 500 ms, stopping 130 ms before the key-press, each labelled with either 'right' or 'left' hand. The sampling was performed at 1000 Hz, band-pass filtered between 0.05 and 200 Hz, before being down sampled to 100 Hz. The electrodes were arranged according to the international 10/20-system.

Dataset C - Riken - Subject A Sessions one and two from Subject A were taken from the RIKEN EEG Datasets homepage. A subject was asked to sit in a chair and pay attention to a blank screen. After 2 seconds, an arrow pointing left or right appeared and, for the following three seconds, the user imagined the corresponding left or right hand movements. The recording was obtained via six channels, sampled at a rate of 256Hz, which was then band-pass filtered between 2 and 30Hz. In total, 264 instances were recorded: session one was selected as the training dataset with 130 trials, with the remaining 134 trials from session two serving as the testing data.

4.2 Features

Power Spectral Densities were selected in the following experiments as they preserve spatial and frequency dimensionality, and by epoching the data, some temporal resolution is preserved. This type of feature can provide practical insight into the problem: allowing understanding as to where the key regions of interest are in terms of which electrodes and frequencies provides the richest information.

4.3 Solution Size

As noted by Chandrashekar and Sahin [19], there are no ideal methods to choose the size of the subset for selection. For this reason, we selected a solution size for the *Iterated Local Search (ILS)* and *Minimal Reduncancy Maximal Relevance-Iterated Local Search (MRMR-ILS)* for Berlin BCI Competition II Dataset III based on Rejer [9]. As there is no background literature that utilises Power Spectral Densities in this way, to the authors' knowledge, for Berlin BCI Competition II Dataset IV, and RIKEN Subject A, preliminary exploration was required.

4.4 Classifiers

The key aim of BCI paradigms is simply to produce an effective model to classify some aspect of neural recordings. The creation of such a model relies heavily on

which machine learning algorithm that was chosen. In this paper, we evaluate two such algorithms:

- *K-Nearest-Neighbours (KNN)*, while commonly used in other fields, have been largely neglected within the BCI literature due to their known sensitivity to the 'Curse of Dimensionality' [20]. They were selected for use in this paper for exploration, and to support our deliberate selection of small feature subsets.

- *Support Vector Machines (SVM)* are commonly used in BCI literature, and often obtain the best accuracies. This is thought to be due to their ability to handle larger feature sets, and their resistance to overfitting [21].

Fitness Function The fitness of a proposed feature subset was evaluated using k-fold cross-validation of the training data. $K = 10$ was selected due to preliminary experimentation revealing a noisy fitness function, originating mainly from the randomly chosen splits in cross-validation. While 10-folds creates an expensive fitness function, it is required in such datasets where we find high-dimensionality, with low number of samples and poor signal-to-noise ratios [22].

Search Algorithm Parametres Each algorithm was run 25 times, with 100,000 evaluations of the classifier set as the termination criteria. In each run, there were 100 perturbation 'kicks', and local searches were limited to 1000 evaluation first-improvement hillclimbers.

Benchmark Methods

Filters - Two mutual information filter methods were evaluated using a greedy forward-search to select the feature subset size, as used in Lan *et al* [23]. Mutual Information Feature Selection (*MIFS*), relies on selecting features that increase the selected subsets mutual information with the class label. *MRMR* seeks to maximise the selected subsets mutual information with the class label (relevance), while minimising the mutual information between features (redundancy).

Wrappers - Two wrapper approaches were selected for comparison: *Sequential Forward Search*, a greedy algorithm that selects the next best feature as evaluated by the classifier; and *Iterated Local Search*, a two layer search involving perturbations and local searches. SFS is a very popular technique, and is often used as an exploratory measure in feature selection. ILS has been used in a wide variety of different search areas, but is almost unheard of within BCI.

Embedded - Least Absolute Shrinkage and Selection Operator (LASSO) (or L1 regularisation) performs feature selection by reducing the sum of the absolute values of the model parametres below an upper bound. It does this by shrinking the coefficients of the features, often to zero, effectively deselecting them. It can provide two feature subsets: Sparse, and Mean Squared Error (MSE). This method provides relatively poor cross-validation error rates on the training set, but tend to be reasonably more generalisable.

5 Results and Discussion

Table 1: Results of each feature selection algorithm while using the KNN Classifier. Number of selected features, cross-validation error rates, and accuracy is shown for Datasets A, B and C. Figures in bold denote the highest performing algorithm for each measure.

Dataset	Algorithm	Selected f	CVE	Accuracy
A	**MIFS**	20	0.410476	0.6
	MRMR	43	0.329524	0.728571
	LASSO (Sparse)	8	0.2186	0.7143
	LASSO (MSE)	29	0.1993	0.7143
	SFS	14	0.1357	0.7357
	ILS	**6**	0.1110	**0.7918**
	MRMR ILS	**6**	**0.1057**	0.7896
B	**MIFS**	10	0.483861	0.56
	MRMR	34	0.475422	0.52
	LASSO (Sparse)	11	0.4269	0.5500
	LASSO (MSE)	13	0.4222	0.5500
	SFS	15	0.2816	0.6200
	ILS	**6**	0.2716	0.6164
	MRMR ILS	**6**	**0.2707**	**0.6464**
C	**MIFS**	6	0.517179	0.619403
	MRMR	30	0.477179	0.522388
	LASSO (Sparse)	4	0.2408	0.6045
	LASSO (MSE)	15	0.2615	0.5672
	SFS	14	**0.1385**	0.5896
	ILS	4	0.1539	0.5997
	MRMR ILS	4	0.1492	**0.6085**

Table 1 and 2 present results obtained using the KNN and SVM classifiers respectively. The list of measures are: the number of features selected by each algorithm (Selected f); the average final solutions' fitnesses (cross-validation error rate on training data; *CVE*); and their accuracy on the unseen, testing data. The datasets were labeled: *A* - Berlin BCI Competition II Dataset III; *B* - Berlin BCI Competition II Dataset IV; *C* - Subject A from the Riken dataset.

When using a KNN classifier, we see in Table 1 that the MRMR-ILS finds solutions with the lowest cross-validation error rates on two datasets: A (10.56%) and B (27.07%). On dataset C, it achieved the second lowest (14.92%), falling only just behind the SFS (13.85%). In all three cases, the MRMR ILS outperformed the unguided ILS. These cross validation error rates reflected the

algorithms' performance on unseen data by achieving the highest accuracy on datasets B (64.64%) and C (60.85%), with the second highest accuracy on dataset A (78.96%).

Table 2: Results of each feature selection algorithm while using the SVM Classifier. Number of selected features, cross-validation error rates, and accuracy is shown for Datasets A, B and C. Figures in bold denote the highest performing algorithm for each measure.

Dataset	Algorithm	Selected f	CVE	Accuracy
A	**MIFS**	20	0.374048	0.607143
	MRMR	43	0.258095	0.792857
	LASSO (Sparse)	8	0.1493	0.7929
	LASSO (MSE)	29	0.1757	0.7929
	SFS	8	0.0857	0.8071
	ILS	**6**	0.0846	**0.8423**
	MRMR ILS	**6**	**0.0843**	0.8269
B	**MIFS**	10	0.415295	0.52
	MRMR	34	0.399684	0.58
	LASSO (Sparse)	11	0.3095	0.6700
	LASSO (MSE)	13	0.3168	0.6200
	SFS	**9**	0.2532	0.6200
	ILS	12	**0.2422**	0.6836
	MRMR ILS	12	0.2439	**0.6948**
C	**MIFS**	6	0.407692	0.537313
	MRMR	30	0.28	0.567164
	LASSO (Sparse)	4	0.2377	0.6045
	LASSO (MSE)	15	0.1508	**0.6567**
	SFS	17	0.1000	0.5970
	ILS	15	0.0735	0.6197
	MRMR ILS	15	**0.0772**	0.6391

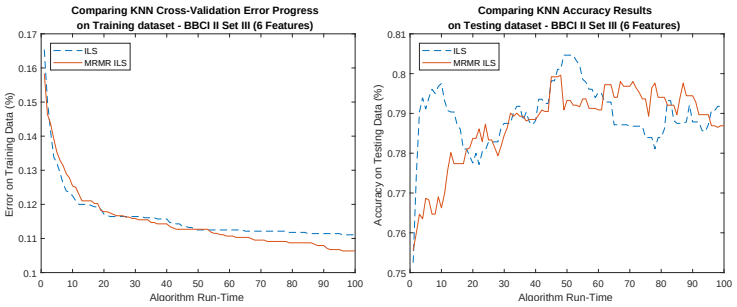
In Table 2, the SVM classifer produces results with a similar pattern as using the KNN, with the MRMR ILS achieving the lowest cross-validation error rates in dataset A and C (8.843% and 7.72% respectively), and falling behind the ILS by just 0.17% on dataset B. Classification accuracies on unseen datasets in this case are slightly more nuanced; the MRMR-ILS achieved the highest accuracy on dataset B (69.48%). In datasets A and C, it achieved the second highest accuracies (84.23% and 65.67%) to ILS and the MSE LASSO solutions (84.23% and 65.67%) respectively.

The graphs in Figures 1a, 2a, and 3a show the average incumbent solution fitness based on the cross-validation error rates over each iteration of the ILS and

MRMR-ILS algorithms. In a post-hoc analysis, we extracted these incumbent solutions and evaluated their predictive accuracy on the testing data, plotted in 1b, 2b, and 3c. We can see that the relationship between the MRMR-ILS fitness function, and the performance on unseen data is much stronger than that observed in the ILS. In order to find a real-world feature subset for BCI applications, it is imperative that the estimated accuracy provided by the fitness function in our algorithms correlates as closely as possible to accuracy rates obtained from new, unseen data. We further explore this in Table 3, in which the Pearson's correlation coefficient is calculated for the cross-validation error rates and accuracies of the incumbent solutions. In five of the six test cases, there is a substantially higher correlation between the predicted accuracy (CVE rate) and the accuracy on the unseen data in the MRMR ILS than that of the ILS. The most notable examples of this is the use of KNN in dataset A, and the use of SVM in dataset C, where the correlations seen within the solutions of the ILS have weak negative correlations (-0.1464 and -0.3787), which is heavily contrasted against the strong negative correlations in those of the MRMR ILS (-0.8954 and -0.7203).

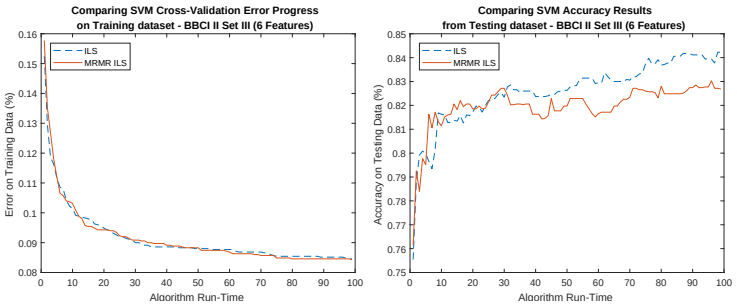
Table 3: Correlations between Cross Validation Error Rates and Accuracy of Solution during ILS and MRMR-ILS Search. Figures in bold denote the highest performing algorithm for each measure.

Classifier	**Dataset**	**Algorithm**	
		ILS	**MRMR ILS**
KNN	**A**	-0.1464	**-0.8954**
	B	-0.7598	**-0.8871**
	C	-0.9224	**-0.9686**
SVM	**A**	**-0.9370**	-0.9100
	B	-0.8348	**-0.8619**
	C	-0.3787	**-0.7203**



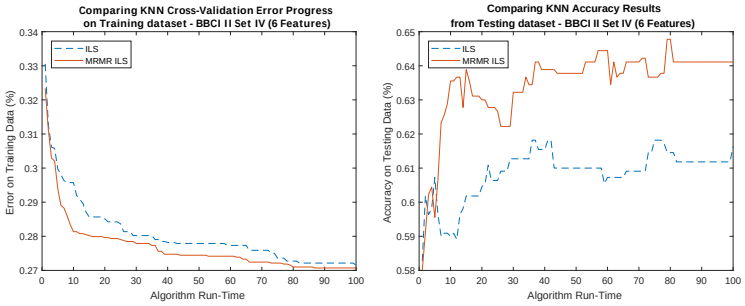
(a) Cross-Validation Error Rates (b) Accuracy on unseen data

Fig. 1: Comparison between ILS and MRMR-ILS over each iteration of the algorithms for the KNN classifier on dataset *A* - BCI Competition II dataset III



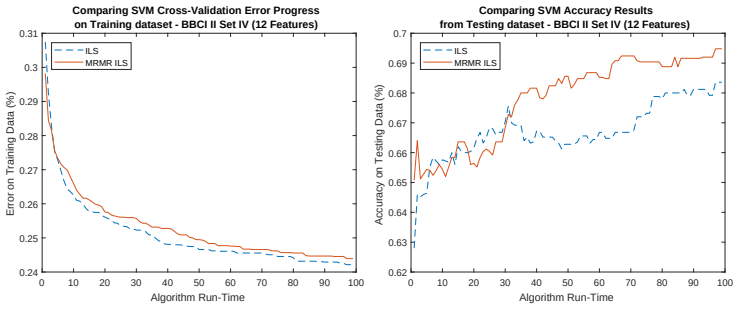
(a) Cross-Validation Error Rates (b) Accuracy on unseen data

Fig. 2: Comparison between ILS and MRMR-ILS over each iteration of the algorithms for the SVM classifier on dataset *A* - BCI Competition II dataset III



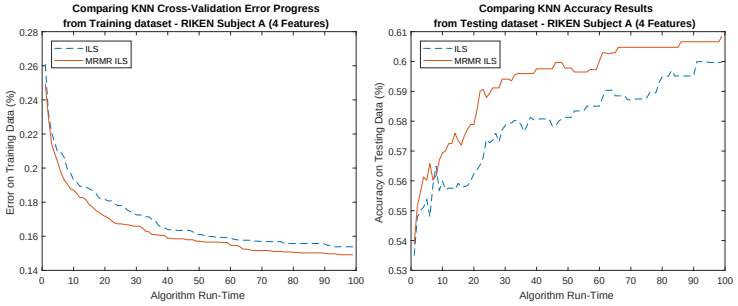
(a) Accuracy on unseen data (b) Accuracy on unseen data

Fig. 3: Comparison between ILS and MRMR-ILS over each iteration of the algorithms for the KNN classifier on dataset *B* - BCI Competition II dataset IV



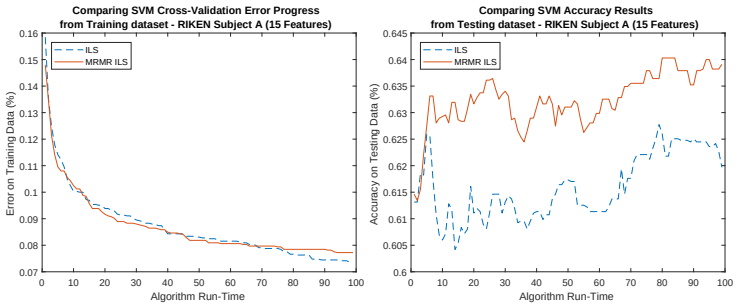
(a) Cross-Validation Error Rates (b) Accuracy on unseen data

Fig. 4: Comparison between ILS and MRMR-ILS over each iteration of the algorithms for the SVM classifier on dataset B - BCI Competition II dataset IV



(a) Cross-Validation Error Rates (b) Accuracy on unseen data

Fig. 5: Comparison between ILS and MRMR-ILS over each iteration of the algorithms for the KNN classifier on dataset C - RIKEN Subject A



(a) Cross-Validation Error Rates (b) Accuracy on unseen data

Fig. 6: Comparison between ILS and MRMR-ILS over each iteration of the algorithms for the SVM classifier on C - RIKEN Subject A

6 Conclusion

This paper proposed MRMR-ILS; a hybrid Filter-Wrapper method involving mutual information for feature selection. Evaluations over three datasets using KNN and SVM classifiers demonstrated that feature subsets found by our method were typically of higher quality, with lower error rates on training sets and higher accuracy on testing data, than those found by the compared traditional methods. What is of additional interest, is the quality of the solutions found during the search process of the MRMR-ILS in comparison to those of the ILS. Relying solely on the cross-validation error rates allowed feature subsets to be discovered that were highly effective for creating models that represent the training data but, when tested on unseen data, their performance was unpredictable. When MRMR was incorporated into the algorithm, the search was partially constrained to areas in the search space rich in mutual information. This resulted in models that generalised to unseen data in a much more consistent manner. Further experimentation should seek to compare the MRMR-ILS with other mutual information based hybrid methods from the wider feature selection literature, and investigate the relationship between mutual information, cross-validation error rates, and predictive accuracy on unseen data.

Acknowledgements Work funded by UK EPSRC grant EP/J017515 (DAASE).

References

1. Xue, B., Zhang, M., Browne, W.N., Yao, X.: A Survey on Evolutionary Computation Approaches to Feature Selection. IEEE Transactions on Evolutionary Computation 20(4), 606–626 (2016)
2. Vega, R., Sajed, T., Mathewson, K.W., Khare, K., Pilarski, P.M., Greiner, R., Sanchez-Ante, G., Antelis, J.M.: Assessment of feature selection and classification methods for recognizing motor imagery tasks from electroencephalographic signals. Artificial Intelligence Research (1), 37–51 (2016)
3. Cabrera, A.F., Farina, D., Dremstrup, K.: Comparison of feature selection and classification methods for a brain-computer interface driven by non-motor imagery. Medical and Biological Engineering and Computing 48(2), 123–132 (2010)
4. Ang, K.K., Chin, Z.Y., Wang, C., Guan, C., Zhang, H.: Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. Frontiers in Neuroscience 6(MAR), 1–9 (2012)
5. Alotaiby, T., El-Samie, F.E.A., Alshebeili, S.A., Ahmad, I.: A review of channel selection algorithms for EEG signal processing. EURASIP Journal on Advances in Signal Processing 2015(1), 66 (2015)
6. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal 27(July 1928), 379–423 (1948)
7. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
8. Ciaccio, E.J., Dunn, S.M., Akay, M.: Biosignal pattern recognition and interpretation systems: Part 2 of 4: Methods for feature extraction and selection. IEEE Engineering in Medicine and Biology Magazine 12(January), 106–113 (1993)

9. Rejer, I.: Genetic algorithm with aggressive mutation for feature selection in BCI feature space. Pattern Analysis and Applications 18(3), 485–492 (2014)
10. Wei, Q., Wang, Y.: Binary Multi-Objective Particle Swarm Optimization for Channel Selection in Motor Imagery Based Brain-Computer Interfaces. 2011 4th International Conference on Biomedical Engineering and Informatics (BME I) pp. 667–670 (2011)
11. Atyabi, A., Luerssen, M., Fitzgibbon, S.P., Powers, D.M.W.: Use of Evolutionary Algorithm-Based Methods in EEG Based BCI Systems. Swarm Intelligence for Electric and Electronic Engineering (May 2016), 326–344 (2012)
12. Gan, J.Q., Awwad Shiekh Hasan, B., Tsui, C.S.L.: A filter-dominating hybrid sequential forward floating search method for feature subset selection in high-dimensional space. International Journal of Machine Learning and Cybernetics 5(3), 413–423 (2014)
13. Khushaba, R.N., Al-ani, A., Alsukker, A., Al-jumaily, A.: A combined ant colony and differential evolution feature selection algorithm. Ant Colony Optimization ... (2008)
14. Tan, F., Fu, X., Zhang, Y., Bourgeois, A.G.: A genetic algorithm-based method for feature subset selection. Soft Computing 12(2), 111–120 (2008)
15. Ali, S.I., Shahzad, W.: A feature subset selection method based on symmetric uncertainty and Ant Colony Optimization A Feature Subset Selection Method based on Symmetric Uncertainty and Ant Colony Optimization (November 2015) (2012)
16. Nguyen, H.B., Xue, B., Liu, I., Zhang, M.: Filter based backward elimination in wrapper based PSO for feature selection in classification. Proceedings of the 2014 IEEE Congress on Evolutionary Computation, CEC 2014 pp. 3111–3118 (2014)
17. Zhu, Z., Jia, S., Ji, Z.: Towards a memetic feature selection paradigm. IEEE Computational Intelligence Magazine 5(2), 41–53 (2010)
18. Lourenco, H.R., Martin, O.C., Stutzle, T.: Iterated Local Search: Framework and Applications, pp. 363–397. Springer US, Boston, MA (2010)
19. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers and Electrical Engineering 40(1), 16–28 (2014)
20. Lotte, F., Congedo, M., Anatole, L., Lotte, F., Congedo, M., Anatole, L.: A Review of Classification Algorithms for EEG-based BCI (2007)
21. Ramos, A.C., Vellasco, M.: Feature Selection Methods Applied to Motor Imagery Task Classification (Mi) (2016)
22. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Appears in the International Joint Conference on Articial Intelligence (IJCAI) 5, 1–7 (1995)
23. Lan, T., Erdogmus, D., Adami, A., Pavel, M., Mathan, S.: Salient EEG channel selection in brain computer interfaces by mutual information maximization. Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference 7, 7064–7 (2005)