Title
The utility of data-driven feature selection: Re: Chu et al. 2012

Authors
Kerr, Wesley T
Douglas, Pamela K
Anderson, Ariana
et al.

Publication Date
2014

DOI
10.1016/j.neuroimage.2013.07.050

Peer reviewed

NeuroImage 84 (2014) 1107–1110

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Comments and Controversies

The utility of data-driven feature selection: Re: Chu *et al.* 2012☆

Wesley T. Kerr[a,*], Pamela K. Douglas[b], Ariana Anderson[b], Mark S. Cohen[b]

[a] David Geffen School of Medicine at UCLA, USA
[b] University of California, Los Angeles, USA

ARTICLE INFO

Article history:
Accepted 18 July 2013
Available online 25 July 2013

Keywords:
Feature selection
Machine learning
Neuroimaging

ABSTRACT

The recent Chu et al. (2012) manuscript discusses two key findings regarding feature selection (FS): (1) data driven FS was no better than using whole brain voxel data and (2) *a priori* biological knowledge was effective to guide FS. Use of FS is highly relevant in neuroimaging-based machine learning, as the number of attributes can greatly exceed the number of exemplars. We strongly endorse their demonstration of both of these findings, and we provide additional important practical and theoretical arguments as to why, in their case, the data-driven FS methods they implemented did not result in improved accuracy. Further, we emphasize that the data-driven FS methods they tested performed approximately as well as the all-voxel case. We discuss why a sparse model may be favored over a complex one with similar performance. We caution readers that the findings in the Chu et al. report should not be generalized to all data-driven FS methods.

Comment:

Recently, Chu et al. (2012) assessed how feature selection (FS) affected classification accuracy on a series of two class problems using gray matter voxel features. FS techniques are categorized typically as filter based, embedded, or wrapper based methods (Tuv et al., 2009). Within the neuroimaging community, data-driven FS (DD-FS) methods have been used commonly because they are generally effective: univariate *t*-test filtering (*e.g.* Esterman et al., 2009; Johnson et al., 2009 and wrapper-based SVM recursive feature elimination (RFE) approaches (established in Guyon et al., 2002; effective in De Martino et al., 2008; Ecker et al., 2010; Dai et al., 2012).

Chu et al. (2012) presented a principled analysis that compared the performance of these two DD-FS approaches with voxelized features from a region of interest (ROI) based on a biological hypothesis, *t*-test in combination within an ROI constraint, and in the absence of any first stage FS. Their analysis revealed that the DD-FS methods tested were unable to outperform simply using all ~300,000 voxel features for discrimination, similar to results published by Cuingnet et al. (2011) who tested a series of FS methods. While Chu et al. clearly discuss that these results are data specific, their findings nonetheless highlight the essential importance for further analysis of FS methods in neuroimaging applications where the data is both noisy and vast. We emphasize that their findings that DD-FS did not improve accuracies should be limited to a certain class of FS methods, for a limited set of parameter choices and kernels. The sensitivity of SVM accuracy to DD-FS methods with respect to changing kernels was discussed by Song et al. (2011), so we focus on the specific DD-FS methods implemented by Chu et al. We caution readers that their results should not be generalized to other DD-FS methods.

We first discuss the two DD-FS methods that were tested, and point out certain theoretical constraints that are common across both techniques. These limitations are well established in the machine learning (ML) literature, and have been discussed by the primary author of the fundamental RFE manuscript (Guyon and Elisseeff, 2003). Both *t*-test filtering and RFE favor selection of features that maximize accuracy individually, assuming that these will provide the highest discrimination accuracy when used in aggregate (Guyon et al., 2002). Consider however, examples where multiple features provide largely redundant, yet highly diagnostic, information (*i.e.*, spatially adjacent neuroimaging voxels), while other features with lower margins and *t* statistics hold unique information (Haxby et al., 2001). Within this framework, the redundant features will be retained, while the features that provide unique information that could improve performance will be discarded. Both of these factors contribute to a decrease in classification accuracy, rather than an increase, as discussed for neuroimaging data by Kriegeskorte et al. (2006), Pereira and Botvinick (2010) and Björnsdotter et al. (2011).

Further, features that are not themselves diagnostic, but which control for nuisance factors (*e.g.* age-associated atrophy; Farid et al., 2012) are expected to have extremely low univariate $|t|$ values and reduced margins. To determine the utility of each feature in RFE, the multivariate separability vector, w, is projected onto each feature-dimension to get a

☆ Chu C., Hsu A.L., Chou K.H., Bandettini P., Lin C., ADNI Initiative. "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images." NeuroImage. 2012;60(1):59–70.

* Corresponding author at: 760 Westwood Plaza, CHS Suite B8-169, Los Angeles, CA 90095, USA. Fax: +1 310 794 7406.
E-mail address: wesleytk@ucla.edu (W.T. Kerr).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.07.050

univariate margin, w_j. In RFE, features with the smallest univariate margin, $||w_j||$, are excluded iteratively until the desired number of features is achieved. We expect that the margin of nuisance-controlling factors would be greater than noise but smaller than the margin of the diagnostic feature. In this case, the smallest margin and $|t|$ statistic features would be excluded before the diagnostic features by these DD-FS methods because the stopping criterion used by Chu et al. was the *number* of selected input features. The stopping criterion is defined by the criteria used to determine exactly how many features are included in the final model. If one had used the observed training or testing accuracy (as in backward or forward selection) or an arbitrary fixed criterion for $||w_j||$ or $|t|$ to determine the stopping criterion, we would expect that these nuisance features may be included in the final model learned using RFE, but not using t-statistic filtering.

In contrast, the least-squares (ℓ_2) regularization in SVM, itself a multivariate DD-FS method, likely includes these nuisance factors: in regularization, features are selected based on the degree to which they maximize classification accuracy instead of reducing the number of input features using an indirect proxy for classification accuracy. The RFE model is mathematically equivalent to the ℓ_2 SVM model in which the smallest SVM margins are set to be identically zero instead of their small estimated value. Similarly, t statistics assumes that the margins of low $|t|$-statistic features should be zero. This assumption is identical to the sparsity assumption of an ℓ_1 regularized SVM. However, ℓ_1 SVMs only outperform ℓ_2 SVMs when the underlying solution itself is sparse (Liu et al., 2007). By extension, we believe that RFE and t statistic filtering will only outperform ℓ_2 SVM if the best diagnostic model is sparse.

As shown by Chu et al., RFE and t-statistics did not improve performance, suggesting that these assumptions of non-redundancy and sparsity may have been violated. These shortcomings suggest that, while t-statistics and RFE have practical value, they are not general panaceas.

The limited efficacy of RFE, or univariate t statistics, does not predict that alternate unsupervised DD-FS algorithms will, or will not, outperform regularization. Independent and Principal Component Analysis (ICA and PCA), for example, can both in effect project multiple linearly correlated, or redundant, features onto a reduced number of features (Comon, 1994; Hyvarinen, 1999; Jutten and Herault, 1991; Wold et al., 1987). In contrast to RFE and t-statistics, these methods that combine highly correlated and, frequently, spatially continuous voxels into regional features improve generalization substantially (*e.g.* Douglas et al., 2010; Franke et al., 2010, 2012; Hinrichs et al., 2009; McKeown et al., 1997). Both ICA and PCA can control for the variation in highly diagnostic independent or principal components due to nuisance factors. Other DD-FS methods such as information criteria (Ding and Peng, 2005; Peng et al., 2005), genetic algorithms (Yang and Honavar, 1998), and Markov Chain Monte Carlo methods (Green, 1995) select a single representative of each set of redundant diagnostic features. This perspective on DD-FS does not modify the original input features; instead it aims to more efficiently select the minimum subset of non-redundant features that maximizes performance. Numerous other DD-FS approaches employ clever algorithms that overcome some of the limitations of RFE and t statistics (*i.e.* Dietterich, 2000; Freund and Schapire, 1997; Friedman et al., 2000; Kwak and Choi, 2002; Leiva-Murillo and Artes-Rodriguez, 2007; Liu and Setiono, 1997; Setiono and Liu, 1997; Sindhwani et al., 2004; Zhang and Sun, 2002; for a review see Saeys et al., 2007). Therefore, we emphasize again that the findings for RFE and t statistics should not be generalized to all DD-FS methods.

As a second practical point, we consider the conclusion that DD-FS performed worse than the feature selection inherent to SVM. We direct attention to Fig. 9E of the original manuscript, which shows how accuracy changes with decreasing values of the SVM regularization parameter, C, as a function of the DD-FS method employed for the largest sample size. We remind the reader of the original soft margin SVM formulation presented famously by Cortes and Vapnik (1995) that presents the Lagrange functional for the two-class problem as:

$$L(w,b,R,\xi)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i[y_i(x_i\cdot w+b)-1+\xi_i]-\sum_{i=1}^{n}r_i\xi_i+C\sum_{i=1}^{n}\xi_i \quad (1)$$

where $n, i, w, b, Y, X, \alpha, r$, and ξ are the total number of exemplars, exemplar index, margin, intercept, output class vector, input data matrix, support vector Lagrange parameter, soft margin Lagrange parameter and soft margin misclassification penalty, respectively. The linear decision function in the feature space takes the form:

$$I(z)=\mathrm{sign}\left(\sum_{\text{Support Vectors}}\alpha_i x_i\cdot z+b\right) \quad (2)$$

where z is the hyperplane perpendicular to w. If $\alpha_i=0$, then the corresponding sample was classified correctly and is irrelevant to the final solution. If $\alpha_i=C$, then the sample was misclassified, and if $0<\alpha_i<C$, then the sample is located on the margin. If $\alpha_i>0$, the sample is called a support vector (Biggio et al., 2011). When solving for very large values of C, the problem tends towards the hard margin solution that can be solved using quadratic programming. With smaller C, the soft margin functional can be optimized through its dual formulation with quadratic programming.

Within their analysis, Chu et al. assessed their accuracy with several parameter choices of C without cross-validation. The global optimum accuracy was obtained in the absence of FS. However, we would like to emphasize that even for the optimum C case (indicated by C^*), the

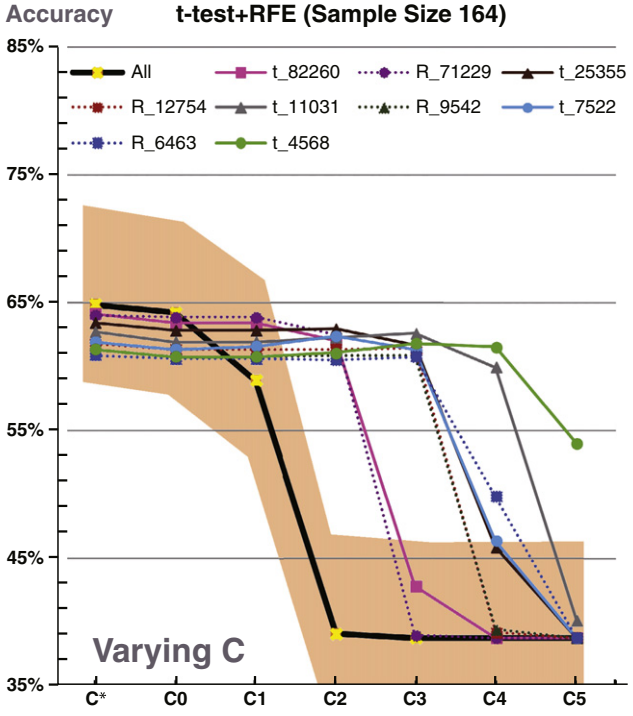

Fig. 1. A reproduction of Chu et al.'s Fig. 9E where the added shading indicates the 95% confidence interval for the no feature selection accuracy using the normal approximation of the binomial distribution. Accuracy using all voxelized features was not significantly higher than data-driven feature selection accuracy at the optimum C, C^*. At multiple non-optimum C values, the accuracy using data-driven feature selection was significantly higher than using all voxelized features.

performance of the other FS algorithms were all within the 95% confidence interval of the no FS approach (Fig. 1). For moderate to small choices of *C*, FS methods systematically outperformed no FS, and were overall less sensitive and more robust to the choice of *C*. As discussed by Chu et al., the selection of this *C* is computationally intense therefore it is frequently simply selected *a priori*.

While we agree that DD-FS does not always improve classification accuracy, it may nevertheless help elucidate the pathology or physiology of the system under study, and can reduce the sensitivity of performance to tuning parameters when applied to the data in a principled manner. Overall, a parsimonious model made possible by DD-FS allows models to be more transparent, and thereby more useful for neuroscientific interpretation (Hanke et al., 2010). This sparsity can be implemented through separate FS methods, or within the SVM itself. While ℓ_2 regularization already applies a degree of sparsity (Cortes and Vapnik, 1995), ℓ_0 regularization imposes a stricter penalty and has been used to interpret dynamic causal modeling features (Brodersen et al., 2011).

In the ML literature, it is common to evaluate methods primarily, or solely, on their classification accuracy. For typical cases, this is entirely appropriate: the goal is to classify, and not to explain. In investigative research, however, the needs are broader and more nuanced. In our own work, we use ML to aid in our understanding of brain function and dysfunction. We have shown previously that in some cases high classification accuracy can be obtained either from nuisance factors (Anderson et al., 2011), or from factors in the data, such as demographics, unrelated to neuroimaging (Colby et al., 2012). While these have the potential to generate clinically meaningful accuracies, they provide limited neurophysiological insight. If, on the other hand, the feature space is selected to project onto well-defined, neurally-oriented subspaces, it is possible to jointly achieve excellent accuracy and explanatory power to aid in neuroscientific discovery. For example, independent components identified from functional MRI data frequently identify the default mode network (Greicius et al., 2004) and have been used for classification (Douglas et al., 2010) as well as the generation of meaningful feature dictionaries (Anderson et al., 2011; Zibulevsky and Pearlmutter, 2001). Although these dictionary elements would vary across subjects and scans, we and others have shown that they are consistent enough to have an identifiable manifestation, an assumption underlying group-ICA methods (Franco et al., in press; Sui et al., 2009). Therefore, these methods accomplish the tasks of feature 'identification' and 'selection' simultaneously.

The goal of feature selection is to minimize the number of estimated parameters in the final machine-learning model to improve performance and generalizability. The concept of balancing the empirical performance of the model to the data with the number of estimated parameters is well established in conventional statistics. For generalized linear models, the pervasive *F* test explicitly divides the explained variance of a model by the number of estimated parameters in the model to calculate the mean squared error. Additionally, the reference *F* distribution for determining significance is wider for models with more estimated parameters. Similarly, the Akaike and Bayesian information criteria (AIC and BIC) explicitly penalize the observed log likelihood of models using a function of the number of estimated parameters. While these criteria cannot formally be applied directly to cross-validation accuracy, our perspective is that the concept behind these criteria is applicable to machine-learning models. Based on that idea, machine-learning models that achieve similar accuracy by operating on a selected set of features are preferred in investigative research over machine-learning models that are saturated with input features. We recognize that, unlike the likelihood or explained variance, cross-validation accuracies do not monotonically increase with the number of estimated parameters. We believe that DD-FS methods, in some situations, can be used effectively to accomplish this dual goal of model simplicity and high empirical cross-validation accuracy.

Despite the shortcomings of the methods tested mentioned herein, we also find it interesting that removal of a vast number of potentially irrelevant features with FS did not offer improvement, despite the theoretical caveats we detail above. It is possible that this lack of improvement is informative in and of itself. We suggest that pre/post FS accuracy should be reported more often, as these results may be helpful in conceptualizing how feature interactions are related to information representation in neural systems.

Because of this improvement in interpretability, we emphasize that FS methods are valuable beyond improving classification accuracy; just as a picture is a thousand words, an interpretable model is oftentimes immensely more valuable than a marginally superior yet uninstructive classification tool.

Acknowledgments

The authors gratefully acknowledge the support received from NIH R33DA026109, NIH T32-GM008185, the UCLA Department of Biomathematics and the William M Keck Foundation. We also thank Jesse Rissman, Nicco Reggente, and Ying Nian Wu for helpful comments during the review process.

Conflict of interest statement

The authors declare no conflict of interest.

References

Anderson, A., Bramen, J., Douglas, P.K., Lenartowicz, A., Cho, A., Culbertson, C., Brody, A.L., Yuille, A.L., Cohen, M.S., 2011. Large sample group independent component analysis of functional magnetic resonance imaging using anatomical atlas-based reduction and bootstrapped clustering. Int. J. Imaging Syst. Technol. 21, 223–231.

Biggio, B., Nelson, B., Laskov, P., 2011. Support vector machines under adversarial label noise. JMLR: Workshop and Conference Proceedings 20, 1–6.

Björnsdotter, M., Rylander, K., Wessberg, J., 2011. A Monte Carlo method for locally-multivariate brain mapping. NeuroImage 56, 508–516.

Brodersen, K.H., Schofield, T.M., Leff, A.P., Ong, C.S., Lomakina, E.I., Buhmann, J.M., Stephan, K.E., 2011. Generative embedding for model-based classification of fMRI data. PLoS Comput. Biol. 7, e1002079.

Colby, J.B., Rudie, J.D., Brown, J.A., Douglas, P.K., Cohen, M.S., Shehzad, Z., 2012. Insights into multimodal imaging classification of ADHD. Front. Syst. Neurosci. 6, 59.

Comon, P., 1994. Independent component analysis, a new concept. Signal Process. 36, 287–314.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 272–297.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehericy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., ADNI, 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56, 776–781.

Dai, D., Wang, J., Hua, J., He, H., 2012. Classification of ADHD children through multimodal magnetic resonance imaging. Front. Syst. Neurosci. 6, 63.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43, 44–58.

Dietterich, T.G., 2000. Ensemble methods in machine learning. Multiple Classifier Systems, 1857, pp. 1–15.

Ding, C., Peng, H.C., 2005. Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. 3, 185–205.

Douglas, P.K., Harris, S., Yuille, A., Cohen, M.S., 2010. Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs disbelief. NeuroImage 56, 544–553.

Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E.M., Brammer, M.J., Murphy, C., Murphy, D.G., the MRC AIMS Consortium 1, 2010. Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. NeuroImage 49, 44–56.

Esterman, M., Chiu, Y.C., Tamber-Rosenau, B.J., Yantis, S., 2009. Decoding cognitive control in human parietal cortex. Proc. Natl. Acad. Sci. 106 (42), 17974–17979.

Farid, N., Girard, H.M., Kemmotsu, N., Smith, M.E., Magda, S.W., Lim, W.Y., Lee, R.R., McDonald, C.R., 2012. Temporal lobe epilepsy: quantitative MR volumetry in detection of hippocampal atrophy. Radiology 264, 542–550.

Franco, A.R., Manell, M.M.V., Calhoun, V., Mayer, A., 2013. Impact of analysis methods on the reproducibility and reliability of resting state networks. Brain Connect (in press).

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., ADNI, 2010. Estimating the age of healthy subjects from T_1-weighted MRI scans using kernel methods: exploring the influence of various parameters. NeuroImage 50, 883–892.

Franke, K., Luders, E., May, A., Wilke, M., Gaser, C., 2012. Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI. NeuroImage 63, 1305–1312.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Stat. 28, 337–374.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.
Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V., 2004. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. Proc. Natl. Acad. Sci. U. S. A. 101, 4637–4642.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422.
Hanke, M., Halchenko, Y.O., Haxby, J.V., Pollmann, S., 2010. Statistical learning analysis in neuroscience: aiming for transparency. Front. Neurosci. 4, 38.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.
Hinrichs, C., Singh, V., Mukherjee, L., Xu, G.F., Chung, M.K., Johnson, S.C., Initi, A.s.D.N., 2009. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. NeuroImage 48, 138–149.
Hyvarinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural. Netw. 10, 626–634.
Johnson, J.D., McDuff, S.G., Rugg, M.D., Norman, K.A., 2009. Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. Neuron 63 (5), 697–708.
Jutten, C., Herault, J., 1991. Blind separation of sources.1. An adaptive algorithm based on neuromimetic architecture. Signal Process. 24, 1–10.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. 103, 3863.
Kwak, N., Choi, C.H., 2002. Input feature selection for classification problems. IEEE Trans. Neural Netw. 13, 143–159.
Leiva-Murillo, J.M., Artes-Rodriguez, A., 2007. Maximization of mutual information for supervised linear feature extraction. IEEE Trans. Neural Netw. 18, 1433–1441.
Liu, H., Setiono, R., 1997. Feature selection *via* discretization. IEEE Trans. Knowl. Data Eng. 9, 642–645.
Liu, Y., Zhang, H.H., Park, C., Ahn, J., 2007. The Lq support vector machine. Contemp. Math. 443, 35–48.
McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.-P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J., 1997. Analysis of fMRI Data by Blind Separation into Independent Spatial Components No. NHRC-REPT-97-42. Naval Health Research Center, San Diego, CA.
Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238.
Pereira, F., Botvinick, M., 2010. Information mapping with pattern classifiers: a comparative study. NeuroImage 56, 476–496.
Saeys, Y., Inza, I., Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517.
Setiono, R., Liu, H., 1997. Neural-network feature selector. IEEE Trans. Neural Netw. 8, 654–662.
Sindhwani, V., Rakshit, S., Deodhare, D., Erdogmus, D., Principe, J.C., 2004. Feature selection in MLPs and SVMs based on maximum output information. IEEE Trans. Neural Netw. 15, 937–948.
Song, S., Zhan, Z., Long, Z., Zhang, J., Yao, L., 2011. Comparative study of SVM methods combined with voxel selection for object category classification on fMRI data. PLoS One 6 (2), e17191.
Sui, J., Adali, T., Pearlson, G.D., Calhoun, V.D., 2009. An ICA-based method for the identification of optimal FMRI features and components using combined group-discriminative techniques. NeuroImage 46, 73–86.
Tuv, E., Borisov, A., Runger, G., Torkkola, K., 2009. Feature selection with ensembles, artificial variables, and redundancy elimination. J. Mach. Learn. Res. 10, 1341–1366.
Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemom. Intell. Lab. Syst. 2, 37–52.
Yang, J.H., Honavar, V., 1998. Feature subset selection using a genetic algorithm. IEEE Intell. Syst. Applic. 13, 44–49.
Zhang, H.B., Sun, G.Y., 2002. Feature selection using Tabu Search method. Pattern Recogn. 35, 701–711.
Zibulevsky, M., Pearlmutter, B.A., 2001. Blind source separation by sparse decomposition in a signal dictionary. Neural Comput. 13, 863–882.