High Performance GP-Based Approach for fMRI Big Data Classification

Extended Abstract

Amirhessam Tahmassebi*
Department of Scientific Computing, Florida State University
Tallahassee, Florida 32306-4120
at15b@my.fsu.edu

Amir H. Gandomi
BEACON Center for the Study of Evolution in Action, Michigan State University
East Lansing, Michigan 48824
gandomi@msu.edu

Anke Meyer-Bäse
Department of Scientific Computing, Florida State University
Tallahassee, Florida 32306-4120
ameyerbaese@fsu.edu

ABSTRACT

We consider resting-state Functional Magnetic Resonance Imaging (fMRI) of two classes of patients: one that took the drug N-acetylcysteine (NAC) and the other one a placebo before and after a smoking cessation treatment. Our goal was to classify the relapse in nicotine-dependent patients as treatment or non-treatment based on their fMRI scans. 80% accuracy was obtained using Independent Component Analysis (ICA) along with Genetic Programming (GP) classifier using High Performance Computing (HPC) which we consider significant enough to suggest that there is a difference in the resting-state fMRI images of a smoker that undergoes this smoking cessation treatment compared to a smoker that receives a placebo.

CCS CONCEPTS

• **Computing methodologies** → **Classification and regression trees**; **Cross-validation**; *Feature selection*; • **Mathematics of computing** → **Multivariate statistics**;

KEYWORDS

fMRI Big Data, Classification, Generic Programming, High Performance Computing

ACM Reference format:
Amirhessam Tahmassebi, Amir H. Gandomi, and Anke Meyer-Bäse. 2017. High Performance GP-Based Approach for fMRI Big Data Classification. In *Proceedings of PEARC17, New Orleans, LA, USA, July 09-13, 2017,* 4 pages.
https://doi.org/10.1145/3093338.3104145

1 INTRODUCTION

Smoking cigarettes is the essential cause of preventable mortality in the United States with around 50% of lifelong smokers dying from illnesses such as heart disease, strokes and cancer [15]. In addition to this, insomnia, tremors and quivering, lightheadedness, high blood pressure, heart attack, and decreasing in bone density are just a few symptoms that nicotine could cause [3]. Nicotine dependency released by tobacco makes people continue smoking cigarettes [16][17]. Developing a cessation treatment [12] with a compound that will reduce a patient's dependency on nicotine as well as the effects of withdrawal could help millions of people quit a dangerous habit. One of these new potentially effective compounds is N-acetylcysteine (NAC) [3][2]. NAC is a white, crystalline compound of the derivative of the amino acid cysteine prodrug [19]. NAC is FDA approved in the United States and has molecular formula of $C_5H_9NO_3S$ and weight of 163.2 [11].

Previously, Genetic Programming (GP) [5] has been used widely for automatic segmentation of 3D MRI data [7], fuzzy feature selection of fMRI data [6], and dynamic casual modeling of fMRI data [9][18]. In this paper, a new approach for classifications, GP has been applied to analyze data from a smoking cessation treatment, where subjects took the drug NAC to reduce their nicotine dependence while still being allowed to smoke in order to keep off the effects of withdrawal. This is the preferred method as more people are likely to try it if they do not have to quit smoking immediately. The goal is to reduce the nicotine dependency to the point that it is easier for the subject to stop. Functional Magnetic Resonance Imaging (fMRI) is a set of noninvasive techniques for functional brain mapping. Areas of high activity are defined to be those where more oxygen-rich blood is flowing [14][13] and the fMRI is able to map these areas. In this study, we have compared some machine learning algorithms with GP to classify the subjects based on whether or not they underwent the treatment. It should be noted that the accuracy of classification will rely heavily on how the data is reduced.

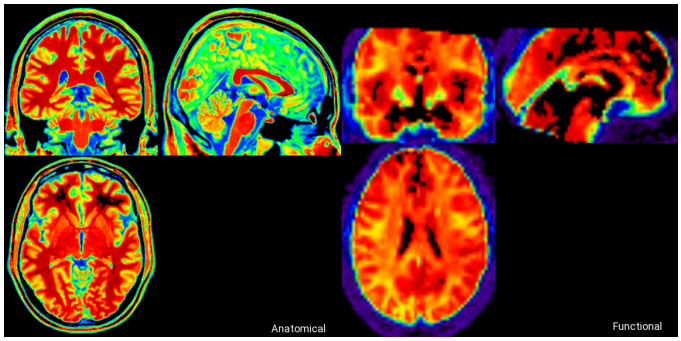

Figure 1: Anatomical and functional MRI data.

*Corresponding Author

PEARC17, July 09-13, 2017, New Orleans, LA, USA

ACM ISBN 978-1-4503-5272-7/17/07.
https://doi.org/10.1145/3093338.3104145

2 SUBJECTS & DATA ACQUISITION

The main goal of this study is to determinate whether or not the drug NAC would decrease nicotine dependency. In this regard, 19 heavy smokers who wanted to quit, took the drug NAC and the other 20 subjects have taken a placebo for two weeks. Then, anatomical and functional scans of their brains have been taken at baseline, and after 2 weeks of NAC treatment. Then, the relapse data are assessed at 6 months past NAC treatment. The treatment study has happened at Amsterdam Center for Addiction and Research. The Amsterdam Addiction Center is equipped with the 3.0 T Intera MRI scanner (Philips Health care, Best, The Netherlands) with a SENSE eight-channel receiver head coil to obtain MRI data. Figure 1 shows the anatomical and functional MRI slices of the brain of a subject. Success over determination of the patients based on fMRI scans would help us to predict relapse and also designing wearable devices to monitor the vital elements of the smokers [10].

3 DATA PRE-PROCESSING

We are given the fMRI data in 4-dimensional spatio-temporal NIFTI (Neuroimaging Informatics Technology Initiative) format. The data contains subject-dependent artifacts due to the long process of the scans, possible movements of the subjects, and physiological noise. The fMRI data analyzing pipeline is made using a combination of Statistical Parametric Mapping (SPM12) and FMRIB Software Library (FSL) to increase the BOLD contrast to noise. Data pre-processing phase contains motion correction, slice timing correction, segmentation, realignment, normalization, smoothing, and co-registration. Figure 2 demonstrates the raw and pre-processed data.

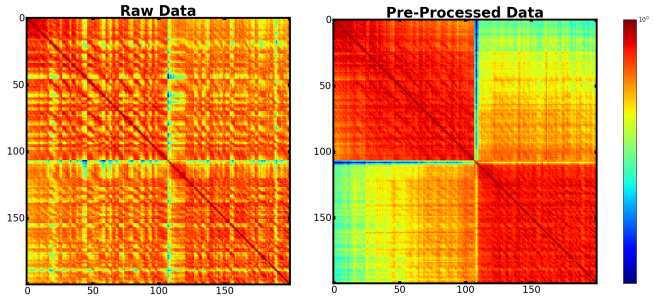

Figure 2: Raw and pre-processed fMRI data.

4 DATA REDUCTIONS

Dealing with a big data problem with 94,720,000 features, with available computational equipment will be computationally expensive and inaccurate. After applying voxel selection scheme, the size of the feature matrix has been reduced to $39 \times 94,720$ which is still a huge number for feature vector for classification. Here, ICA and PCA have been employed for data reduction to find the feature vector to feed the classifier [20], as we increase the numbers of components, the correlation decreases.

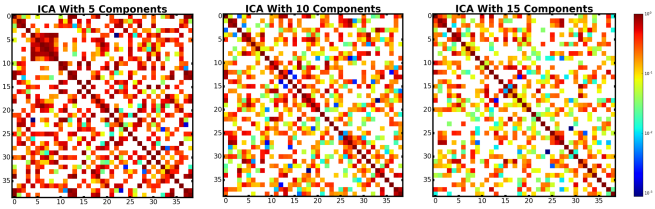

Figure 3: Correlation matrix for ICA.

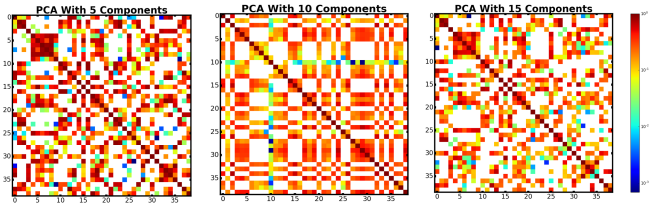

Figure 4: Correlation matrix for PCA.

5 CLASSIFICATION

In this paper, we have compared four different machine learning algorithms with GP. Naive Bayes (NB) with Bernoulli and Gaussian distribution, Logistic Regression (LR), third-Nearest Neighbors (KNN, K=3), and GP have been employed as classifiers.

GP has played the role of the main classifier due to the selection of designs applies on fitness measurement phase. GP has been formulated as a symbolic optimization technique originally based on functional programming language as an evolutionary method to use computer programs for solving a problem following the principle of Darwinian natural selection [4]. We have employed the fitness function during evolution to have the most efficient guided GP.

$$Fitness = \frac{Number\ \ of\ \ patients\ \ classified\ \ correctly}{Number\ \ of\ \ patients\ \ used\ \ for\ \ training}$$

In addition to this, to find the best mathematical formula, a Crossover operator, Subtree Mutations, Point Mutation, Hoist Mutation, and Reproduction operators have been employed in the GP model. Based on the termination criteria, leave-one-out cross validation has been applied and classification accuracy has been used as the output. The parameters setting used in GP classifier would be found in Table 1.

Table 1: Parameters setting for GP classifier.

Parameter	Setting
Population Size	2000
Number of Generations	13000
Hall of Fame	500
Tournament Size	50
P Crossover	0.9
P Subtree Mutation	0.01
P Hoist Mutation	0.01
P Point Mutation	0.01
P Point Replace	0.05
Function Set	*add*, *sub*, *mul*, *div*, *log*, *neg*, *inv*, *abs*
Parsimony Coefficient	0.0005
Max Samples	0.9
Random State	0
Number of Jobs	3

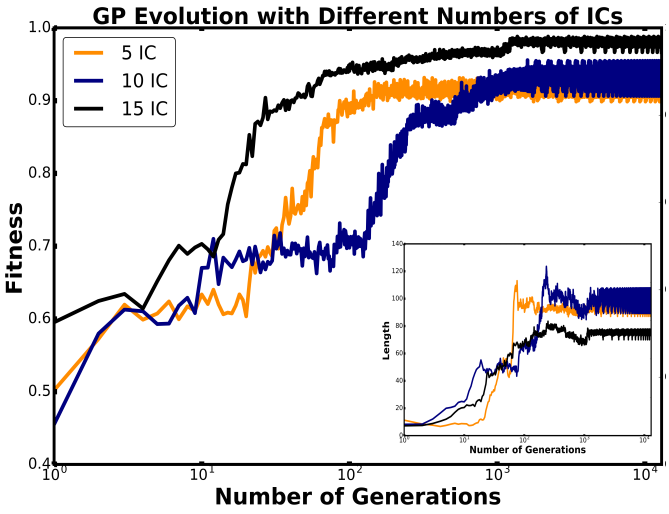

Figure 5: GP evolution with different numbers of ICs.

6 RESULTS

A GP model with 13000 generations and 2000 populations for classifications task with two major data reduction methods, namely ICA and PCA has been developed in Python [8] [1]. Hall of fame has been set to 500 with tournament size of 50. The GP model has been paralleled with three nodes on HPC. For each data reduction scheme, best fitness and average length through generations for different numbers of components have been reported. The axes have been chosen in logarithmic scale. Classification accuracy of GP classifier has been compared with some machine learning algorithms, namely LR, Gaussian NB, Bernoulli NB, and 3rd NN for ICA and PCA. The classification accuracy for all the classifiers could be found in Tables 2 and 3.

Figure 5 shows the GP evolution with ICA. It is obvious that by increasing the number of IC, the average fitness has been increased too. It should be noted that the average length used in the GP model has been decreased for higher numbers of components. That brought us the best classification accuracy with 15 numbers of ICs with average length of 70.

This trend can be seen in Figure 6 where we have presented the GP evolution for different numbers of PCs. The best fitness is also related to 15 components but with 100 as the average length of the GP model. That would be the reason to have lower accuracy than ICA model.

Table 2: Classification accuracy for different numbers of ICs.

IC	LR	Bernoulli NB	Gaussian NB	KNN	GP
5	38.46%	66.66%	41.02%	61.53%	68.71%
10	35.89%	38.46%	48.71%	61.53%	64.10%
15	38.46%	48.71%	41.02%	51.28%	79.48%

Table 3: Classification accuracy for different numbers of PCs.

PC	LR	Bernoulli NB	Gaussian NB	KNN	GP
5	61.53%	58.97%	46.15%	61.53%	64.10%
10	46.15%	51.28%	43.58%	61.53%	64.10%
15	53.84%	61.53%	43.58%	58.97%	68.71%

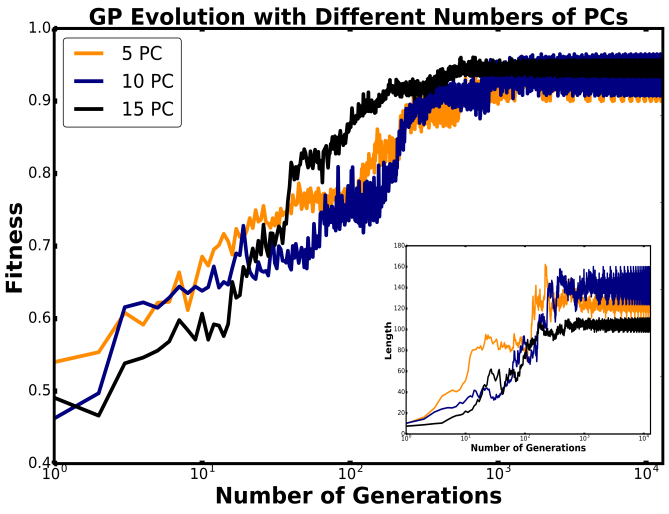

Figure 6: GP evolution with different numbers of PCs.

The best classification accuracy has been found using 15 IC as we can see in the Table 2 for GP. On the other hand, for 5, and 10 IC, the GP model struggles with increasing the depth of the model to increase the fitness factor. It should be noted that the average

lengths for 5 and 10 IC are around 100-110. Looking closer, classification accuracy for the two different data reduction methods might change significantly for a different number of components. Especially, the classification error which differs with data distribution for each method. In addition to this, as we increased the numbers of components, the length of the GP model decreased. On the other hand, by having 13000 numbers of generations, the fitness fluctuates around a constant number. The phenomenon is more obvious for the length of the model too. This could be fixed by decreasing the number of generations or tweaking the GP operators which demands employing optimizations algorithms.

7 CONCLUSIONS

We have compared an evolutionary approach, Genetic Programming model, with multivariate machine learning methods along with ICA and PCA to conduct analyses on high activity regions in the limbic system of the fMRI data. Due to the power of GP methods in classifications and flexible heuristic techniques, GP outperformed the other methods. The best classification accuracy (80%) has been found using GP model along with ICA with 15 numbers of independent components. This result suggests that there is a difference in the resting-state fMRI images of a smoker that undergoes the smoking cessation treatment compared to a smoker that receives a placebo.

REFERENCES

[1] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[2] Anahid Ehtemami. 2016. *Statistical data analysis of resting state fMRI: A study of nicotine addiction treatment.* Ph.D. Dissertation. The Florida State University.

[3] Anahid Ehtemami, Aria Smith, Daniel Fratte, Anke Meyer-Baese, Anna E Goudriaan, Lianne Schmaal, Mieke HJ Schulte, and Olmo Zavala-Romero. 2016. Functional connectivity analysis of resting-state fMRI networks in nicotine dependent patients. In *SPIE Medical Imaging*. International Society for Optics and Photonics, 978827–978827.

[4] Amir Hossein Gandomi and Amir Hossein Alavi. 2011. Multi-stage genetic programming: a new strategy to nonlinear system modeling. *Information Sciences* 181, 23 (2011), 5227–5239.

[5] John R Koza. 1994. Genetic programming II: Automatic discovery of reusable subprograms. *Cambridge, MA, USA* (1994).

[6] Jong-Hyun Lee, Javad Rahimipour Anaraki, Chang Wook Ahn, and Jinung An. 2015. Efficient classification system based on Fuzzy–Rough Feature Selection and Multitree Genetic Programming for intension pattern recognition using brain signal. *Expert Systems with Applications* 42, 3 (2015), 1644–1651.

[7] R Moller and Rene Zeipelt. 2001. Automatic segmentation of 3D-MRI data using a genetic algorithm. In *Medical Imaging and Augmented Reality, 2001. Proceedings. International Workshop on*. IEEE, 278–281.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[9] M Pyka, D Heider, S Hauke, T Kircher, and A Jansen. 2011. Dynamic causal modeling with genetic algorithms. *Journal of neuroscience methods* 194, 2 (2011), 402–406.

[10] Francisco J Romero, Diego P Morales, Encarnación Castillo, Antonio García, Amirhessam Tahmassebi, and Anke Meyer-Baese. 2017. Reconfigurable wearable to monitor physiological variables and movement. In *SPIE Commercial+ Scientific Sensing and Imaging*. International Society for Optics and Photonics, 1021608–1021608.

[11] MHJ Schulte, AE Goudriaan, W Van den Brink, RW Wiers, and L Schmaal. 2015. P. 1. j. 031 The efficacy of N-acetylcysteine on smoking cessation, impulsivity and cue reactivity in heavy smokers. *European Neuropsychopharmacology* 25 (2015), S350.

[12] Mieke HJ Schulte, Janna Cousijn, Tess E den Uyl, Anna E Goudriaan, Wim van den Brink, Dick J Veltman, Thelma Schilt, and Reinout W Wiers. 2014. Recovery of neurocognitive functions following sustained abstinence after substance dependence and implications for treatment. *Clinical psychology review* 34, 7 (2014), 531–550.

[13] Amirhessam Tahmassebi. 2015. *Fluid Flow Through Carbon Nanotubes And Graphene Based Nanostructures.* Ph.D. Dissertation. University of Akron.

[14] Amirhessam Tahmassebi and Alper Buldum. 2015. Fluid flow calculations of Graphene Composites. In *APS March Meeting Abstracts*, Vol. 1. 1140P.

[15] Amirhessam Tahmassebi, Amir H. Gandomi, Ian McCann, Mieke HJ Schulte, Lianne Schmaal, Anna E. Goudriaan, and Anke Meyer-Baese. 2017. An Evolutionary Approach for fMRI Big Data Classification. *IEEE Congress on Evolutionary Computation* (2017).

[16] Amirhessam Tahmassebi, Amir H. Gandomi, Ian McCann, Mieke HJ Schulte, Lianne Schmaal, Anna E. Goudriaan, and Anke Meyer-Baese. 2017. fMRI Smoking Cessation Classification. *IEEE Transactions on Cybernetics* (2017).

[17] Amirhessam Tahmassebi, Amir H. Gandomi, Ian McCann, Mieke HJ Schulte, Lianne Schmaal, Anna E. Goudriaan, and Anke Meyer-Baese. 2017. fMRI Smoking Cessation Classification Using Genetic Programming. *Workshop on Data Science meets Optimization* (2017).

[18] Amirhessam Tahmassebi, Katja Pinker-Domenig, Georg Wengert, Marc Lobbes, Andreas Stadlbauer, Francisco J Romeroa, Diego P Morales, Encarnacion Castillo, Antonio Garcia, Guillermo Botella, et al. 2017. Dynamical graph theory networks techniques for the analysis of sparse connectivity networks in dementia. In *SPIE Commercial+ Scientific Sensing and Imaging*. International Society for Optics and Photonics, 1021609–1021609.

[19] Amirhessam Tahmassebi, Katja Pinker-Domenig, Georg Wengert, Marc Lobbes, Andreas Stadlbauer, Norelle C Wildburger, Francisco J Romeroa, Diego P Morales, Encarnacion Castillo, Antonio Garcia, et al. 2017. The driving regulators of the connectivity protein network of brain malignancies. In *SPIE Commercial+ Scientific Sensing and Imaging*. International Society for Optics and Photonics, 1021605–1021605.

[20] Reza Tavoli, Ehsan Kozegar, Mohammad Shojafar, Hossein Soleimani, and Zahra Pooranian. 2013. Weighted PCA for improving Document Image Retrieval System based on keyword spotting accuracy. In *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*. IEEE, 773–777.