Optimizing Feature Selection and Extraction: Genetic Algorithms in EEG Analysis for Meditation Type Classification

Abhishek Jain[1], Suhel Ahamed [2]
[1,2]Department of IT, Central University (GGU) Bilaspur Chhattisgarh India

Abstract:
Dimensionality reduction is a critical preprocessing step in data analysis, aimed at addressing the curse of dimensionality by reducing the number of features while retaining relevant information. Genetic algorithms (GAs) have emerged as effective tools for optimizing feature selection and extraction in this context. This paper provides a comprehensive review and evaluation of genetic algorithms for dimensionality reduction. It explores the underlying principles, methodologies, applications, strengths, and limitations of GAs in reducing the dimensionality of complex datasets. Furthermore, this work examines the empirical performance of GA-based approaches for EEG domains for meditation type. The paper concludes with a discussion on challenges, future directions, and potential research opportunities in genetic algorithm-based dimensionality reduction techniques.

1. Introduction

1.1 Background
The exponential growth of data has led to high-dimensional datasets in various fields, posing challenges in terms of computational complexity, overfitting, and interpretability. Dimensionality reduction techniques aim to alleviate these issues by transforming the data into a lower-dimensional space while preserving relevant information [1,2,3].

1.2 Problem Statement
The paper aims to investigate the effectiveness of genetic algorithms (GAs) for dimensionality reduction and provide insights into their applications, strengths, and limitations. It also aims to identify challenges and propose potential research directions to enhance GA-based dimensionality reduction techniques.

1.3 Objectives
The objectives of this research paper are:

- To provide a comprehensive review of genetic algorithms and their relevance to dimensionality reduction.
- To analyze different GA-based approaches for feature selection and extraction.
- To evaluate the empirical performance of genetic algorithm techniques across various domains.
- To identify challenges and limitations associated with GA-based dimensionality reduction.
- To suggest potential research directions and areas for improvement.

2. Genetic Algorithm for Dimensionality Reduction

2.1 Genetic Algorithm Basics
Genetic algorithms are population-based optimization techniques inspired by the process of natural evolution. They involve the iterative generation of candidate solutions, selection of individuals based on their fitness, genetic operators such as crossover and mutation, and the evaluation of fitness to guide the search for optimal solutions[4].

2.2 Genetic Operators for Dimensionality Reduction
Genetic operators play a crucial role in GA-based dimensionality reduction[5,6]. Selection operators, such as tournament selection or roulette wheel selection, determine the individuals that will reproduce and contribute to the next generation. Crossover operators, such as one-point or multi-point crossover, combine genetic information from two parent solutions to produce offspring. Mutation operators, such as bit-flip or swap mutation, introduce random changes to the genetic material, promoting exploration of the search space[7,8].

2.3 Encoding and Representation

The choice of encoding and representation scheme is essential in GA-based dimensionality reduction. Binary encoding, real-value encoding, or permutation-based encoding can be employed based on the nature of the problem and the desired representation[9,10,11,12].

3. GA-based Dimensionality Reduction Techniques

3.1 Feature Selection

Feature selection methods aim to identify a subset of relevant features from the original feature set. GA-based feature selection approaches utilize various evaluation metrics, such as fitness functions based on classification accuracy, information gain, or correlation, to guide the search for an optimal feature subset[13,17].

3.2 Feature Extraction

Feature extraction techniques encompass the transformation of the original feature space into a representation of lower dimensions. Principal Component Analysis (PCA) stands as a widely employed GA-based method for feature extraction, striving to identify orthogonal projections of the data that amplify the variance. Other GA-based methods for feature extraction encompass Linear Discriminant Analysis (LDA) and Non-negative Matrix Factorization (NMF)[12,9].

3.3 Hybrid Approaches

Hybrid approaches combine genetic algorithms with other optimization or machine learning techniques to enhance dimensionality reduction. For example, Genetic Programming (GP) can be integrated with GAs to evolve feature selection or extraction functions. Additionally, hybridization with clustering algorithms or fuzzy systems can improve the performance of GA-based dimensionality reduction methods[14,16].

4. Methodology

Dimensionality reduction using a genetic algorithm involves finding an optimal subset of features or variables that best represent the data while minimizing the dimensionality. Here's a step-by-step methodology for using a genetic algorithm for dimensionality reduction:

1. Data Preprocessing:
   - Extract statistical features from the signal, such as mean, standard deviation, skewness, kurtosis, etc.
   - Normalize the features to bring them to a similar scale if required.
2. Define the Problem:
   - Specify the objective of dimensionality reduction, such as maximizing classification accuracy using KNN.
   - Determine the evaluation metric for KNN classification, such as accuracy, precision, recall, or F1 score.
3. Encoding:
   - Represent each candidate solution (feature subset) as a binary string, where each bit corresponds to the inclusion or exclusion of a feature.
   - Determine the length of the binary string based on the total number of features.
4. Initialization:

- Generate an initial population of random feature subsets.
- Specify the population size, which is the number of candidate solutions.

5. Fitness Evaluation:
- Evaluate the fitness of each individual in the population.
- Perform KNN classification using the selected features and the specified evaluation metric to measure the quality of feature subsets.
- Fitness can be defined as the classification accuracy, precision, recall, or F1 score achieved by the KNN classifier on the training or validation set.

6. Selection:
- Select individuals from the population for reproduction based on their fitness.
- Use selection methods like tournament selection or roulette wheel selection to favor individuals with higher fitness.

7. Genetic Operators:
- Apply genetic manipulations (mutation and crossover) to the chosen individuals to produce offspring.
- Crossover: Combine genetic material from two parent feature subsets to create new feature subsets.
- Mutation: Introduce random changes in feature subsets to maintain diversity in the population.

8. Offspring Generation:
- Generate a new population by combining the offspring with the surviving individuals from the previous generation.

9. Fitness Evaluation:
- Evaluate the fitness of the new population by performing KNN classification on the training or validation set.

10. Termination Criteria:
- Determine the termination condition, such as reaching a maximum number of generations or a predefined fitness threshold.
- If the termination condition is met, go to Step 11. Otherwise, go back to Step 6.

11. Final Solution:
- Select the best individual (feature subset) from the final population based on fitness.
- This individual represents the optimal subset of features for dimensionality reduction.

12. Evaluation:
- Apply the selected feature subset to the KNN classifier.
- Evaluate the performance of the KNN classifier on the test set using the chosen evaluation metric.
- Analyze the classification results and assess the effectiveness of the dimensionality reduction process.

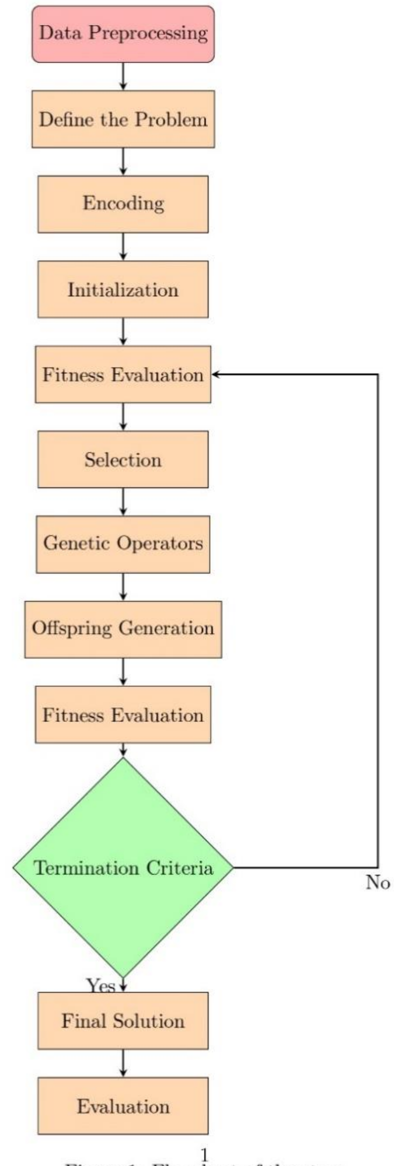

Figure 1: Flowchart of the steps

5. Result and discussion

Using a 64-channel Biosemi system and a Biosemi 10-20 head cap montage at a sampling rate of 2048 Hz, we gathered data. The Biosemi ActiView data collecting system for measuring impedance was used to keep all electrodes within a 15 offset tolerance. There are 24 participants in this meditation experiment. Every two minutes, the subjects were stopped during their meditation to gauge their level of focus and wandering thoughts. They were labelled as two classes of expert and non-expert.

	Predicted Class	
	Class 1	Class 2
Actual Class 1	8	4
Actual Class 2	3	9

Following are the calculation:

a) TP: 8

b) TN: 9

c) FP: 4
d) FN: 3

- Accuracy: (8 + 9) / (8 + 9 + 4 + 3) = 17 / 24 ≈ 0.7083 (70.83%)
- Precision (Class 1): 8 / (8 + 4) = 0.6667 (66.67%)
- Precision (Class 2): 9 / (9 + 3) = 0.75 (75%)
- Recall (Class 1): 8 / (8 + 3) = 0.7273 (72.73%)
- Recall (Class 2): 9 / (9 + 4) = 0.6923 (69.23%)
- F1 Score (Class 1): 2 * (0.6667 * 0.7273) / (0.6667 + 0.7273) ≈ 0.6957 (69.57%)
- F1 Score (Class 2): 2 * (0.75 * 0.6923) / (0.75 + 0.6923) ≈ 0.719 (71.9%)

6. Evaluation of GA-based Dimensionality Reduction

6.1 Performance Metrics

The evaluation of GA-based dimensionality reduction techniques requires appropriate performance metrics. Commonly used metrics include classification accuracy, reduction ratio, information gain, clustering quality, and visualization effectiveness[15,4,7].

6.2 Empirical Performance

Empirical evaluations of GA-based dimensionality reduction techniques involve conducting experiments on diverse datasets and comparing the performance with other state-of-the-art methods. Case studies across domains such as bioinformatics, image analysis, text mining, and financial data analysis demonstrate the effectiveness of GA-based approaches[8,9,11,13].

7. Challenges and Future Directions

7.1 Scalability and Efficiency

As the dimensionality of datasets increases, the scalability and computational efficiency of GA-based techniques become a challenge. Future research should focus on developing parallel and distributed algorithms to handle large-scale datasets efficiently.

7.2 Overfitting and Generalization

GA-based dimensionality reduction techniques may encounter overfitting issues, especially in the presence of noisy or redundant features. The exploration-exploitation trade-off needs to be carefully balanced to ensure generalization performance[4,6,10].

7.3 Hybridization and Ensemble Methods

Exploring hybridization with other optimization techniques and ensemble methods can potentially improve the performance and robustness of GA-based dimensionality reduction methods. Techniques such as particle swarm optimization, differential evolution, and ensemble-based approaches can be investigated[10,11,16,19,20].

7.4 Interpretability and Visualization

Interpretability and visualization of the reduced feature space are important considerations in dimensionality reduction. Future research should focus on developing techniques that preserve interpretability while optimizing the reduction process.

Conclusion

This research paper presented a comprehensive review and evaluation of genetic algorithms for dimensionality reduction. It discussed their methodologies, applications, strengths, and limitations. The empirical performance of GA-based techniques across various domains was analyzed. Furthermore, challenges and potential research directions were identified to enhance GA-based dimensionality reduction techniques.

References:

[1] C. Cimpanu, L. Ferariu, T. Dumitriu, and F. Ungureanu, ‘Multi-Objective Optimization of Feature Selection procedure for EEG signals classification’, E-Health and Bioengineering Conference, 2017.
[2] F. M. Noori, N. Naseer, N. K. Qureshi, S. O. Gilani, H. Nazeer, and R. A. Khan, ‘Optimal feature selection from fNIRS signals using genetic algorithms for BCI’, Neuroscience Letters, 2017.
[3] N. Bidi and Z. Elberrichi, ‘Feature selection for text classification using genetic algorithms’, International Conference on Modelling, Identification and Control, 2016.
[4] H.-C. Lee, L.-W. Ko, H.-L. Huang, J.-Y. Wu, Y.-T. Chuang, and S.-Y. Ho, ‘Statistical analysis and classification of EEG-based attention network task using optimized feature selection’, IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain, 2014.
[5] S.-F. Liang, H.-C. Wang, and W.-L. Chang, ‘Combination of EEG complexity and spectral analysis for epilepsy diagnosis and seizure detection’, EURASIP Journal on Advances in Signal Processing, 2010.
[6] S. R. Ahmad, A. A. Bakar, and M. R. Yaakub, ‘Metaheuristic algorithms for feature selection in sentiment analysis’, Sai, 2015.
[7] M. Esmaeili, M. Zahedi, and N. Hafezi-Motlagh, ‘Performance Analysis of PSO and GA Algorithms in Order to Classifying EEG Data’, 2015.
[8] B. Paulchamy, ‘Efficient Removal of Artifacts from EEG SIGNAL Using Enhanced Hybrid Learning Method’, 2017.
[9] D. Kimovski, J. Ortega, A. Ortiz, F. G. Montoya, F. Manzano-Agugliaro, and R. Baños, ‘Parallel alternatives for evolutionary multi-objective optimization in unsupervised feature selection’, Expert Systems With Applications, 2015.
[10] M. Z. Baig, N. Aslam, H. P. H. Shum, and L. Zhang, ‘Differential Evolution Algorithm as a Tool for Optimal Feature Subset Selection in Motor Imagery EEG’, Expert Systems With Applications, 2017.
[11] T. Wen and Z. Zhang, ‘Effective and extensible feature extraction method using genetic algorithm-based frequency-domain feature search for epileptic EEG multiclassification’, Medicine, 2017.
[12] S. Bhattacharyya, A. Sengupta, T. Chakraborti, A. Konar, and D. N. Tibarewala, ‘Automatic feature selection of motor imagery EEG signals using differential evolution and learning automata’, Medical & Biological Engineering & Computing, 2014.
[13] A. Majkowski et al., ‘Selection of EEG signal features for ERD/ERS classification using genetic algorithms’, 2017 18th International Conference on Computational Problems of Electrical Engineering (CPEE), 2017.
[14] W.-Y. Hsu, ‘Improving classification accuracy of motor imagery EEG using genetic feature selection’, Clinical Eeg and Neuroscience, 2014.
[15] M. Cerrada, R.-V. Sánchez, R. V. Sanchez, D. Cabrera, G. Zurita, and C. Li, ‘Multi-Stage Feature Selection by Using Genetic Algorithms for Fault Diagnosis in Gearboxes Based on Vibration Signal’, Sensors, 2015.
[16] T. Brandejsky, ‘The Use of Local Models Optimized by Genetic Programming Algorithms in Biomedical-Signal Analysis’, Handbook of Optimization, 2013.
[17] M.-L. Wang, C.-W. Lin, N. M. Mayer, M.-H. Hu, and P.-Y. Lee, ‘An brain-computer interface for video content analysis system for perceive emotions by using EEG’, International Conference on Consumer Electronics-Taiwan, 2016.
[18] M. Adamczyk, ‘Genetics of human sleep EEG’, 2015.
[19] Rehmat Khan, Rohit Raja (2016) Introducing L1- Sparse Representation Classification for facial expression, Published in Imperial Journal of Interdisciplinary Research (IJIR), Vol. 2, Iss. 4, pp. 115-122, ISSN: 2454-1362.
[20] Nikita Rawat, Rohit Raja (2016), A Survey on Vehicle Tracking with Various Techniques”, International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Vol. 5 Iss. 2, pp. 374-377, ISSN: 2278-1323.