Network-Guided Particle Swarm Optimization for Feature Selection in High-Dimensional Microarray Data

Your Name
Department / Institute Name
your.email@domain.com

Abstract

Feature selection is a critical task in high-dimensional microarray data classification, where the number of features significantly exceeds the number of samples. Traditional evolutionary feature selection methods often suffer from redundancy and instability due to the lack of structural information among features. In this work, we propose a Network-Guided Particle Swarm Optimization (NetG-PSO) approach that incorporates feature dependency information through a correlation-based feature network. The proposed method leverages network-aware initialization and velocity control to guide the swarm toward compact and informative feature subsets. Experiments conducted on the colon cancer microarray dataset demonstrate that NetG-PSO significantly reduces the number of selected features while maintaining or improving classification accuracy compared to standard Binary PSO, Differential Evolution-based feature selection, and a network-agnostic PSO variant. Furthermore, a post-selection pruning strategy using Random Forest feature importance yields a highly compact gene signature of 20 features with an accuracy of 84.21%. These results confirm the effectiveness of integrating network structure into swarm-based feature selection.

Keywords: Feature Selection, Particle Swarm Optimization, Feature Network, Microarray Data, Evolutionary Computation

1 Introduction

High-dimensional datasets are common in bioinformatics, particularly in microarray gene expression analysis, where thousands of genes are measured for a limited number of samples. This high feature-to-sample ratio often leads to overfitting, increased computational cost, and reduced model interpretability. Feature selection aims to identify a small subset of informative features that improves classification performance while reducing redundancy.

Evolutionary algorithms such as Genetic Algorithms, Particle Swarm Optimization (PSO) [6, 4], and Differential Evolution (DE) [7] have been widely used for feature selection due to their ability to perform global search in large spaces. However, most existing approaches treat features as independent entities and ignore inherent relationships among them. In biological data, genes often interact and form functional groups, suggesting that exploiting feature dependencies may improve selection quality.

To address this limitation, we propose a Network-Guided Particle Swarm Optimization (NetG-PSO) framework that integrates feature interaction information through a correlation-based network. By embedding network structure into swarm initialization and velocity updates, the proposed method promotes the selection of coherent and non-redundant feature subsets. A comprehensive overview of feature selection techniques can be found in [5].

2 Dataset Description

The proposed approach is evaluated on the colon cancer microarray dataset [2], a widely used benchmark in feature selection research. The dataset consists of gene expression profiles from colon tissue samples, containing a small number of samples and a large number of gene features. Prior to feature selection, the data are normalized using z-score normalization to ensure consistent feature scaling.

3 Proposed Methodology

3.1 Feature Network Construction

A feature network is constructed to capture dependencies among features. Each feature is represented as a node, and weighted edges are established between pairs of features whose Pearson correlation coefficient exceeds a predefined threshold. The resulting network encodes feature relationships and serves as prior knowledge for the optimization process. Feature relationships are particularly important in biological data, where genes often interact in functional groups [1].

3.2 Network-Guided Particle Swarm Optimization

In the proposed NetG-PSO framework, each particle encodes the feature subset as a real-valued position vector. A sigmoid transfer function maps particle positions to binary feature selection decisions. Unlike standard PSO, network information is incorporated into both the initialization and velocity update mechanisms. Features with higher network centrality are more likely to be selected during initialization, while velocity updates are adjusted to favor features belonging to highly connected network regions.

3.3 Fitness Function

The fitness of a particle is defined as a weighted combination of classification accuracy and feature subset size:

$$Fitness = \alpha \cdot Accuracy - (1 - \alpha) \cdot \frac{|S|}{D} \tag{1}$$

where $|S|$ is the number of selected features, D is the total number of features, and α controls the trade-off between accuracy and sparsity.

3.4 Post-Selection Pruning

To further enhance interpretability, a post-selection pruning step is applied using Random Forest feature importance [3]. The initially selected feature pool is ranked by importance, and only the top-k features are retained for final evaluation.

4 Experimental Setup

The proposed method is compared against Binary PSO (BPSO), Differential Evolution-based Feature Selection (DE-FS), a network-agnostic PSO variant, and a no-feature-selection baseline. All methods are evaluated using the same training-testing splits and classifier configuration. Performance is measured using classification accuracy, precision, recall, F1-score, and the number of selected features. Each experiment is repeated multiple times with different random seeds to assess stability.

5 Results and Discussion

5.1 Comparison of Feature Counts

Figure 1 compares the average number of selected features across methods. NetG-PSO selects substantially fewer features than BPSO and DE-FS, demonstrating the effectiveness of network guidance in reducing redundancy.

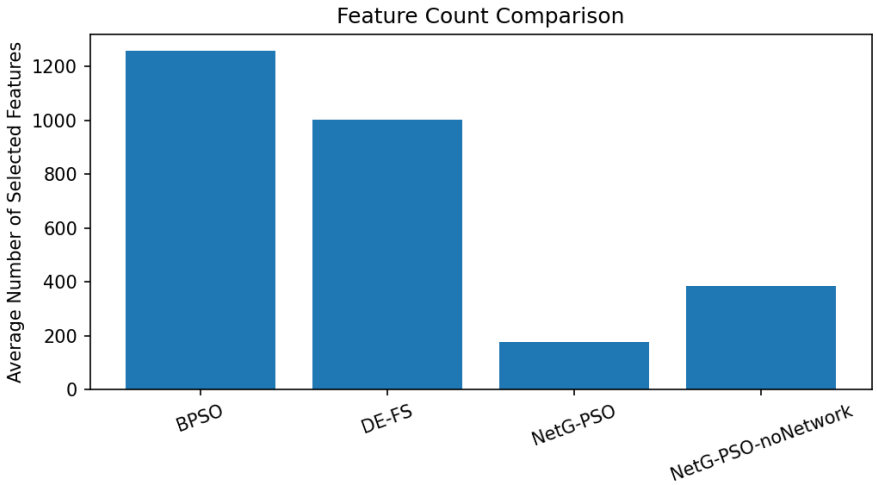

Figure 1: Average number of selected features by different methods.

5.2 Classification Accuracy Comparison

Figure 2 presents the distribution of test accuracy for all methods. NetG-PSO achieves competitive accuracy with improved stability compared to baseline approaches.

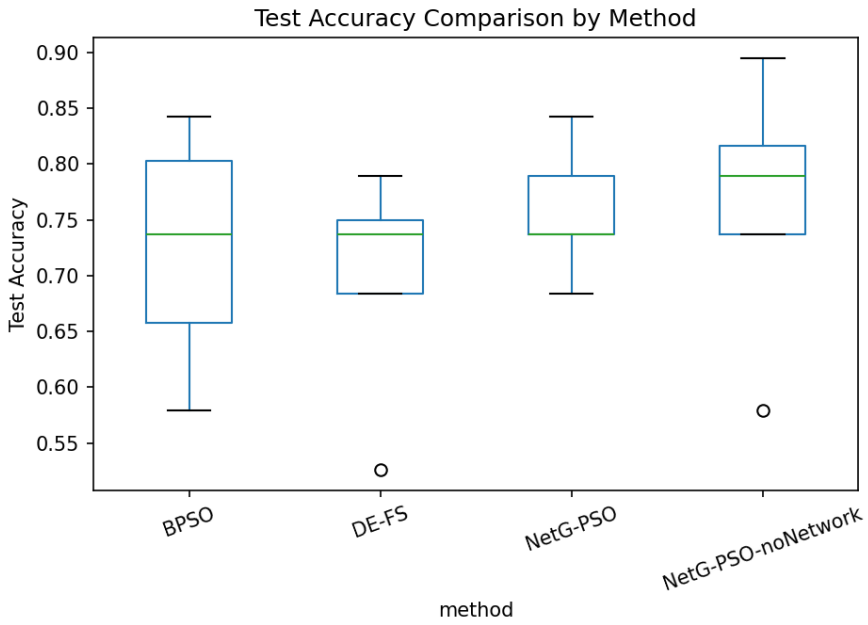

Figure 2: Test accuracy comparison across methods.

5.3 Accuracy vs Feature Count

The trade-off between accuracy and feature subset size is illustrated in Figure 3. NetG-PSO achieves higher accuracy with significantly fewer features.

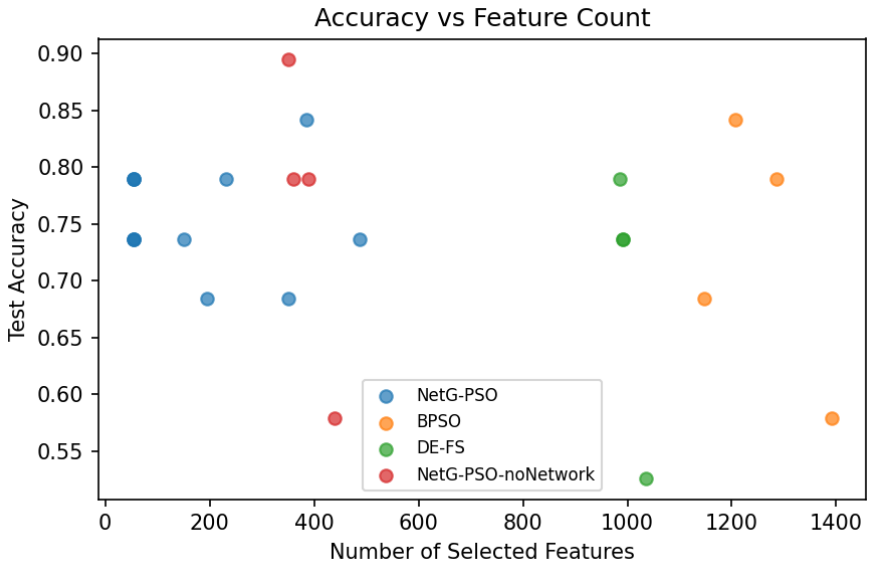

Figure 3: Accuracy versus number of selected features.

5.4 Post-Pruning Analysis

After Random Forest-based pruning, NetG-PSO achieves a classification accuracy of 84.21% using only 20 features, highlighting the potential for highly compact and interpretable gene signatures.

6 Conclusion

This work presents a Network-Guided Particle Swarm Optimization approach for feature selection in high-dimensional microarray data. By incorporating feature dependency information through a correlation-based network, the proposed method effectively reduces feature redundancy while maintaining high classification accuracy. Experimental results demonstrate that NetG-PSO outperforms traditional evolutionary feature selection methods and yields compact feature subsets after post-selection pruning. Future work may explore extending the proposed framework to multi-objective optimization and additional biological datasets.

References

[1] Constantin F Aliferis et al. Machine learning models for microarray data. *Methods in Microarray Data Analysis*, pages 87–102, 2003.

[2] Uri Alon, Niv Barkai, Daniel A Notterman, K Gish, S Ybarra, D Mack, and AJ Levine. Gene expression profiles of colon cancer. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Russell Eberhart and Yuhui Shi. Particle swarm optimization: Developments, applications and resources. *Proceedings of the 2001 Congress on Evolutionary Computation*, 1:81–86, 2001.

[5] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[6] James Kennedy and Russell Eberhart. Particle swarm optimization. *Proceedings of ICNN'95 - International Conference on Neural Networks*, 4:1942–1948, 1995.

[7] Rainer Storn and Kenneth Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.