RELIABLE CLASSIFICATION OF TWO-CLASS CANCER DATA USING EVOLUTIONARY ALGORITHMS

Kalyanmoy Deb∗, A. Raji Reddy

Kanpur Genetic Algorithms Laboratory (KanGAL), Indian Institute of Technology Kanpur, Kanpur 208 016, India

Lucia Helmeciova

PROBLEM INTRODUCTION

Gene Expression:

Gene expression is measured using DNA microarrays, where mRNA levels from cells are captured on a chip and quantified to reflect how actively genes are being.

Problem Statement:

High dimensionality: A vast number of genes (often thousands) are measured, but available sample sizes (patients) are typically much smaller.

The need to identify which genes are most relevant for distinguishing between different types of cancer or disease states.

Current Challenges:

Overfitting and noise: Many genes may not be relevant, adding noise to the classification process.

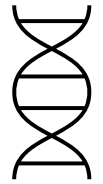
Computational complexity: Finding the most informative subset of genes from thousands possible is computationally challenging and resource-intensive.

Objective of the Study:

Utilize evolutionary algorithms, specifically multiobjective optimization, to effectively reduce the gene subset size while maximizing classification accuracy.

Developing a reliable method for gene subset selection can lead to better diagnostic tools and more personalized treatment options for cancer patients.

I M P L E M E N T A T I O N



Filter all negative and small gene expression values and set them to 0 => only those with large expression variability remain.

Define multiobjective optimization: => minimize size of subset

=> minimize number of mismatches in training using LOOCV

=> minimize number of mismatches in testing using LOOCV

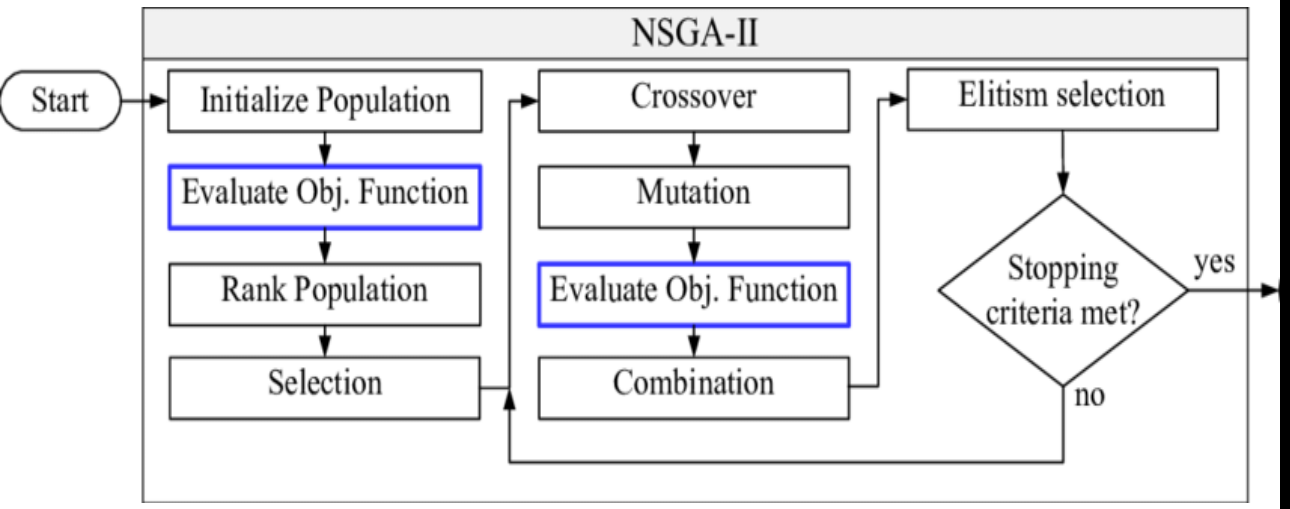
Code the genes in dataset into l-bit strings, where presence of genes is marked 1 and absence 0

Initialize population such that only 10% of string are 1, do it randomly.

Calculate multiobjective criteria values.

Use NSGA II

NON-DOMINATED SORTING PROCESS



Classification of Solutions:

- **Initial Identification:** The process starts by identifying solutions in the population that are not dominated by any other. These solutions make up the first "front".
- **Rank Assignment:** Each solution in the first front is assigned a rank of 1, indicating top priority in terms of optimality.

Layering of Solutions:

- **Successive Fronts:** After setting aside the first front, the next set of solutions that are not dominated by the remaining solutions is identified. This set forms the second front.
- **Continued Process:** This layering continues, with each subsequent set of non-dominated solutions forming a new front and receiving the next higher rank. The process repeats until all solutions are ranked.

NSGA-II ALGORITHM

- NSGA-II (Non-dominated Sorting Genetic Algorithm II) is a widely used evolutionary algorithm for solving multi-objective optimization problems efficiently.

1. **Initialization:**

   Parent Population (Pt): Start with a parent population of size N.

   Genetic Operators: Apply genetic operators such as single-point crossover and bit-wise mutation to create an offspring population Qt also of size N.

2. **Combination:**

   Combined Population (Rt): Combine the parent (Pt) and offspring (Qt) populations to form a new population Rt of size 2N.

3. **Non-Dominated Sorting:**

   Sorting Procedure: Use a non-dominated sorting approach to classify the solutions in Rt.

   Create New Parent Population (Pt+1): Select solutions from Rt starting with the best non-dominated front, followed by the second and third, etc., until the population size exceeds N.

4. **Handling Excess Solutions:**

   Overflow: If the last non-dominated front to be considered has more solutions than available slots in Pt+1, only select those with the highest diversity.

   Crowding Distance: Calculate a simple crowding distance for each solution, which is the Euclidean distance in the objective space between neighboring solutions. Select solutions with the largest crowding distances to maximize diversity in the parent population.

5. **Iteration:**

   Continuation: Repeat the above steps for a user-defined number of iterations T, or until convergence criteria are met.

CASE STUDY: LEUKEMIA, LYMPHOMA, AND COLON CANCER

Dataset: 72 Leukemia (composed of 62 bone marrow and 10 peripheral blood)

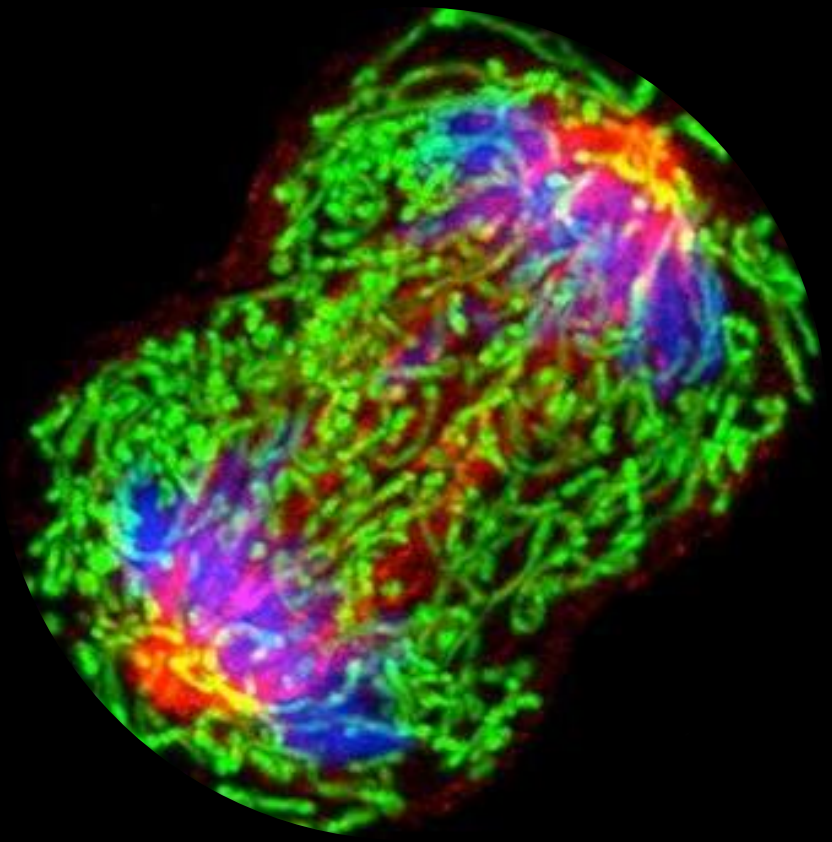
96 Lymphoma (42 samples of (DLBCL) and 54 samples of other types)

62 Colon (22 normal and 40 Colon cancer samples)

Results: Leukemia: 4 genes, 100% accuracy on test

Lymphoma: 5 genes 100% accuracy

Colon: 7 genes, one mismatch (not combination that would be 100% accurate)



FURTHER EXPERIMENTS

- *Modified domination criterion for multiple gene subset sizes:*

Biased-domination definition differs from the original dominance definition in that any two solutions with identical *fj* values will not dominate each other = more solutions.

This approach produced 36 equally 100% accurate gene sets in Leukemia.

- *Multimodal MOEAs for multiple solutions:*

Specifically designed to capture and maintain diversity not just across the objective space but also within the same level of performance in the objective space.

Acknowledges and utilizes the fact that different combinations of decision variables (genes) can lead to the same performance outcome, which is particularly important in biological contexts where different genes may have different biological implications.

RESOURCES:

- DEB, Kalyanmoy; REDDY, A. Raji. Reliable classification of two-class cancer data using evolutionary algorithms. *BioSystems*, 2003, 72.1-2: 111-129.
- EHRENREICH, Armin. DNA microarray technology for the microbiologist: an overview. *Applied microbiology and biotechnology*, 2006, 73.2: 255-273.