GA for Feature Selection in Training Cancer Prediction ML Models

A presentation by
Soumaya Adabala
Enrollment Id 24EE62R03
M.Tech Control Systems Engineering
IIT Kharagpur

September 15, 2024

Contents

Introduction

The Problem

“ Curse of dimensionality”
Unnecessary data leads to excessive complications.

- Genetic testing is done to analyzes DNA to find mutations linked to cancer risk.
- Bioinformatics data includes diverse features like gene expressions, and protein sequences.
- Extensive testing may lead to higher medical expenses.
- Complex data can create uncertainty about cancer prognosis.

The problem

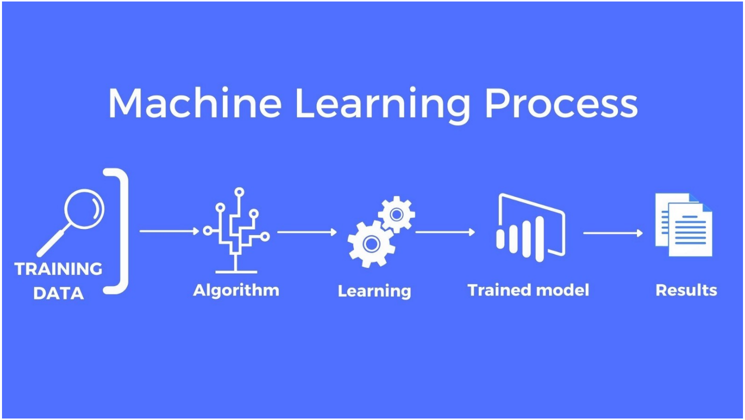

Figure: ML Process

The Solution

- **F**eature Selection: Selecting the right subset of data.

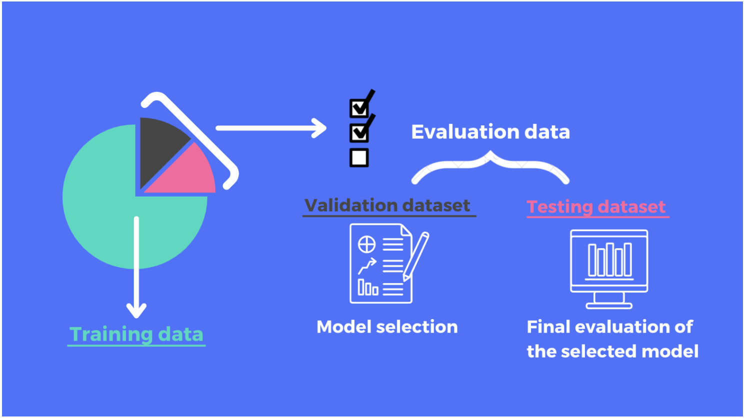

Figure: Data Validation

The Solution

- Genetic Algorithms (GA) are used in feature selection to optimize the selection of relevant features by evolving subsets that maximize model performance.
- GAs help in reducing dimensionality by identifying the most significant features, thereby improving computational efficiency and model accuracy.
- **Why GA?**
  - flexible
  - adaptable
  - robust
  - efficiently handling complex and nonlinear feature interactions.

Flowchart

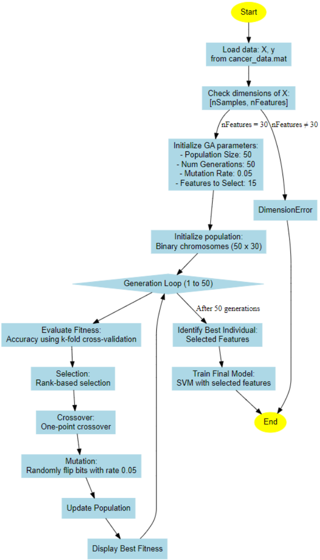

Figure: Flowchaart for feauture selection using GA

Fitness Function

Fitness Function

- **Objective:**
  - Evaluate the performance of feature subsets based on classification accuracy.
- **Inputs:**
  - **Feature Matrix $\mathbf{X}$**: $n \times m$ (samples x features)
  - **Target Labels $\mathbf{y}$**: $n \times 1$
  - **Selected Features $\mathbf{s}$**: $m \times 1$ binary vector (1 if selected, 0 if not)
- **Feature Subset Extraction:**

$$\mathbf{X_s} = \mathbf{X}(:, \mathbf{s})$$

- **Model Training & Evaluation:**
  - Use k-fold cross-validation to evaluate model performance.
  - Train a classifier (SVM Support Vector Machine) and compute accuracy for each fold.
- **Accuracy Calculation:**

$$\text{Accuracy}^i = \frac{\sum_{j=1}^{|\mathbf{X}_{\mathbf{test}}{}^i|} (\text{predicted}_j == \text{actual}_j)}{|\mathbf{X}_{\mathbf{test}}{}^i|}$$

- **Fitness Score:**

$$\text{Fitness}(\mathbf{s}) = \frac{1}{k} \sum_{i=1}^{k} \text{Accuracy}^i$$

Software Demonstration

Results

Command Window

```
Generation 3: Best Fitness = 0.62
Generation 4: Best Fitness = 0.6
Generation 5: Best Fitness = 0.6
Generation 6: Best Fitness = 0.61
Generation 7: Best Fitness = 0.59
Generation 8: Best Fitness = 0.61
Generation 9: Best Fitness = 0.59
Generation 10: Best Fitness = 0.63
Generation 11: Best Fitness = 0.61
Generation 12: Best Fitness = 0.61
Generation 13: Best Fitness = 0.6
Generation 14: Best Fitness = 0.65
Generation 15: Best Fitness = 0.62
Generation 16: Best Fitness = 0.58
Generation 17: Best Fitness = 0.64
Generation 18: Best Fitness = 0.62
Generation 19: Best Fitness = 0.59
Generation 20: Best Fitness = 0.59
Generation 21: Best Fitness = 0.61
Generation 22: Best Fitness = 0.64
Generation 23: Best Fitness = 0.64
Generation 24: Best Fitness = 0.64
Generation 25: Best Fitness = 0.61
Generation 26: Best Fitness = 0.62
Generation 27: Best Fitness = 0.64
Generation 28: Best Fitness = 0.61
Generation 29: Best Fitness = 0.68
Generation 30: Best Fitness = 0.63
Generation 31: Best Fitness = 0.6
Generation 32: Best Fitness = 0.61
Generation 33: Best Fitness = 0.61
Generation 34: Best Fitness = 0.62
Generation 35: Best Fitness = 0.63
Generation 36: Best Fitness = 0.6
Generation 37: Best Fitness = 0.6
Generation 38: Best Fitness = 0.63
Generation 39: Best Fitness = 0.64
Generation 40: Best Fitness = 0.59
Generation 41: Best Fitness = 0.6
Generation 42: Best Fitness = 0.6
Generation 43: Best Fitness = 0.6
Generation 44: Best Fitness = 0.64
Generation 45: Best Fitness = 0.61
Generation 46: Best Fitness = 0.62
Generation 47: Best Fitness = 0.6
Generation 48: Best Fitness = 0.63
Generation 49: Best Fitness = 0.64
Generation 50: Best Fitness = 0.6
Selected Features:
     3     7     8     9    10    11    16    17    18    19    21    24    26    28    29
```

Figure: Results

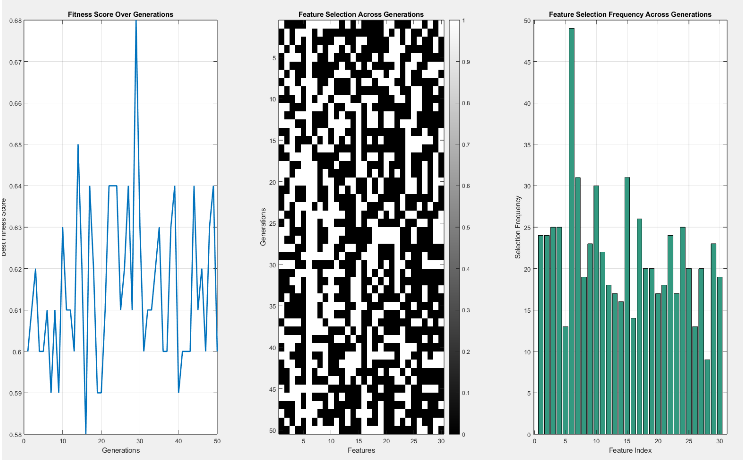

Figure: Results

Conclusions

Conclusions

- Genetic Algorithms optimize feature selection by evolving subsets to enhance model performance, reduce dimensionality, and improve both computational efficiency and accuracy.
- Applications include.
  - Healthcare Diagnosis
  - Customer Segmentation
  - Predictive Maintenance

References

References

- Liu, B. G., Xu, L. J., Wang, Y. H., & Tang, J. H. (2012). *Genetic Algorithm-Based Feature Selection for Classification: A Comparative Study*. Presented at the IEEE International Conference on Systems, Man, and Cybernetics (SMC). Available at IEEE Xplore.
- García, J. S., Gómez, J. A., & Carrillo, A. J. L. (2013). *Feature Selection Using Genetic Algorithms: A Review*. IEEE Transactions on Evolutionary Computation. Available at IEEE Xplore.
- Pal, S. K., Sinha, S. K., & Chaudhuri, B. B. (2011). *Feature Selection for Classification Using Genetic Algorithms*. Pattern Recognition. Available at ScienceDirect.

Thank You