The Use of Genetic Algorithms for Feature Selection

Munshi Imran Hossain

Cytel, India

PSI - 2018

Agenda

- The Problem of Feature Selection
- Genetic Algorithms (GA)
- The Data Science Problem
- Results
- Conclusion

The Problem of Feature Selection

- Large number of features; sometimes greater than 100.
- The number of combinations can be well over a billion!

- Is there a way to search for an optimal set of features in reasonable time and with reasonable computation power?

Different ways to search for this needle

- Evaluate every possible combination to come up with the best combination – the brute force method!
- Step-up/step-down methods that add or remove a feature at a time and evaluate model performance.
- Use genetic algorithms (GA) for searching this huge solution space.

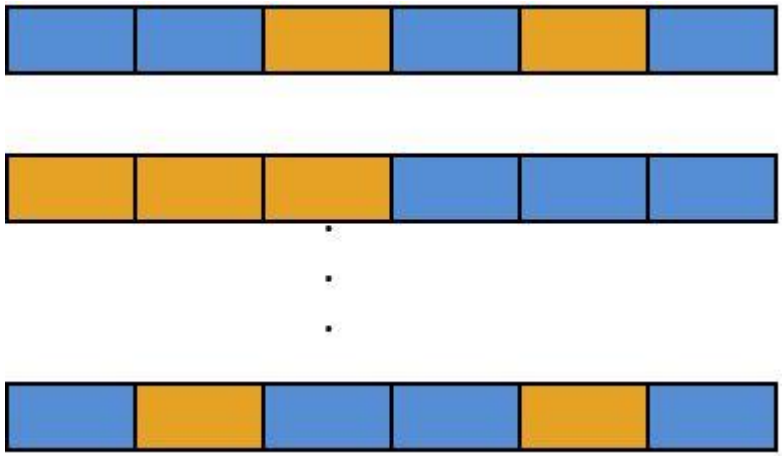
Genetic Algorithms (GA)

- This is a high level simulation of a biologically inspired adaptive system – evolution.
- Using a simple set of rules, this system can have emergent behaviour that makes it useful for various applications.
- GA have been used in applications such as
  - predicting the structure of proteins
  - training neural networks
- Here, I will talk about the use of GA for searching through the feature space to select an optimal set of features.

Terms associated with GA

- **Chromosome** – a potential solution to the problem. A common way to represent solutions is using binary numbers.

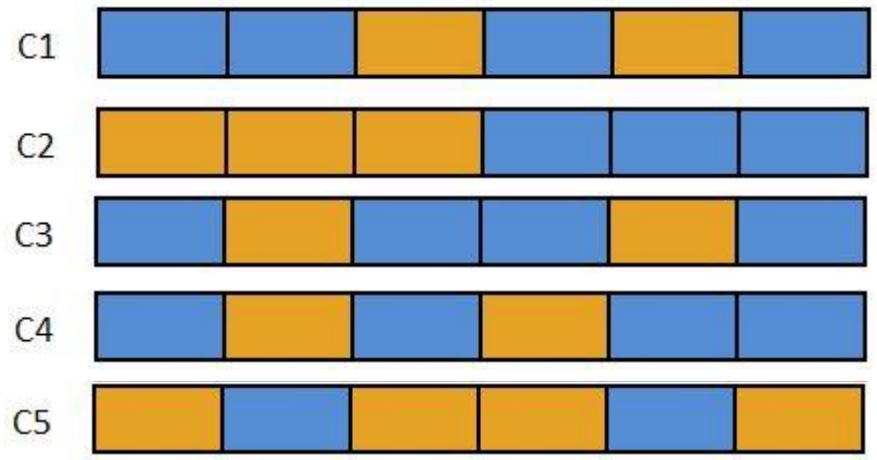
- **Population** – a set of chromosomes belonging to a generation.



Cytel

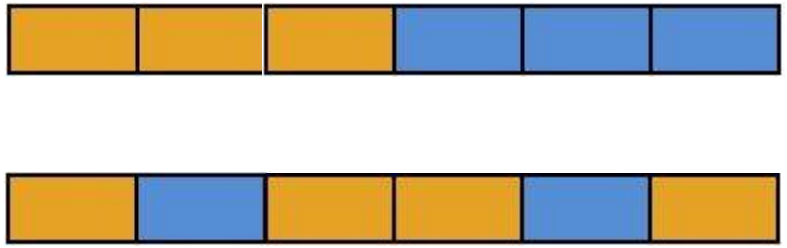
Terms associated with GA

- **Generation** – each iteration of the algorithm.
- **Fitness** – a metric to evaluate how well a particular solution solves the problem.
- **Selection** – a process by which some chromosomes of a population are chosen for generating new solutions.

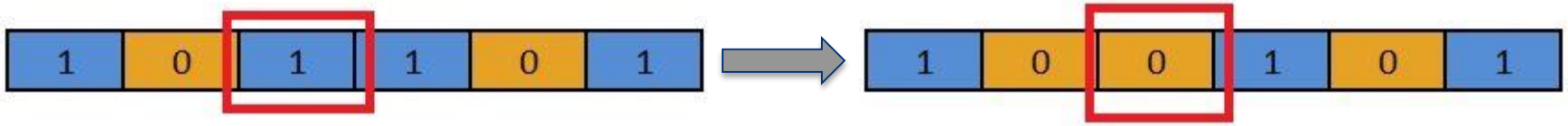


Terms associated with GA

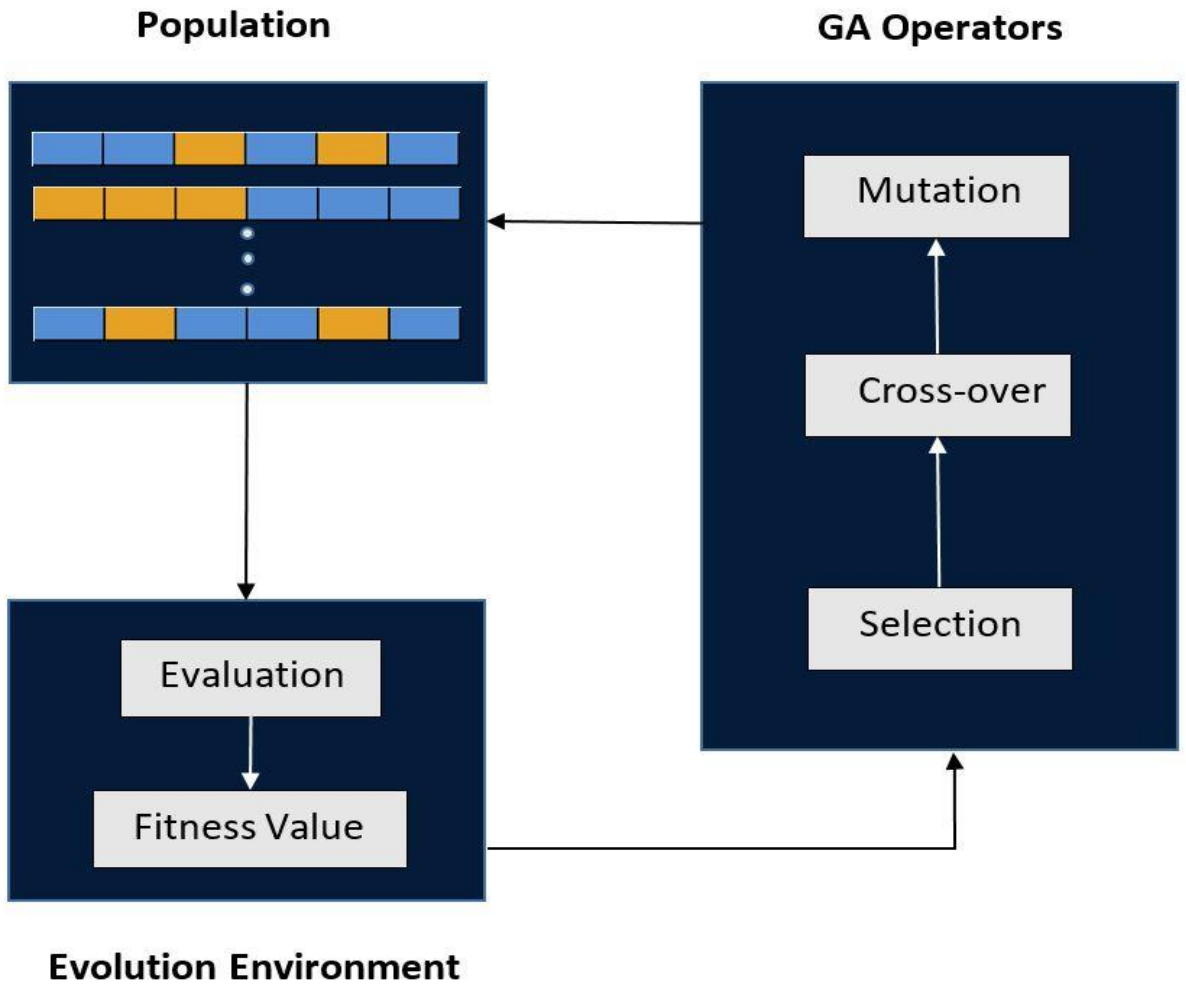
- **Cross-over** – is the process of exchange of information between selected chromosomes.



- **Mutation** – random changes in chromosomes.



Schematic of a GA



The Schema Theorem (John Holland, 1970)

- *Short, low-order schemata with above-average fitness increase exponentially in frequency in successive generations*.
- A **schema** is a template that identifies a subset of strings with similarities at certain string positions. For example, H = 10*1* is a schema for the binary strings
  - 10010,
  - 10011,
  - 10110 and
  - 10111

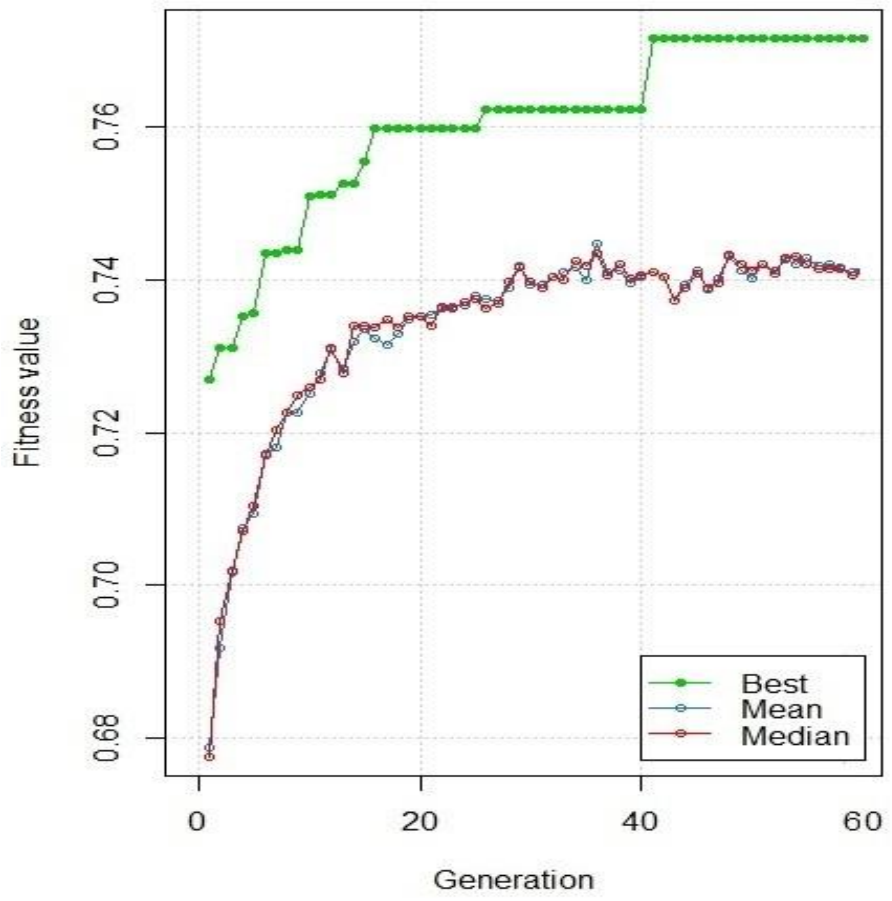
The Data Science Problem

- Data from a device containing a dual-axis accelerometer.
- The accelerometer signals capture motion in the subject which are then used to build models to predict normal or impaired movement in the subject.
- A number of features (over 100) are derived from the processed accelerometer data.
- The number of observations is around 1500.
- Linear Discriminant Analysis (LDA) models were built for classifying signals as normal or impaired.

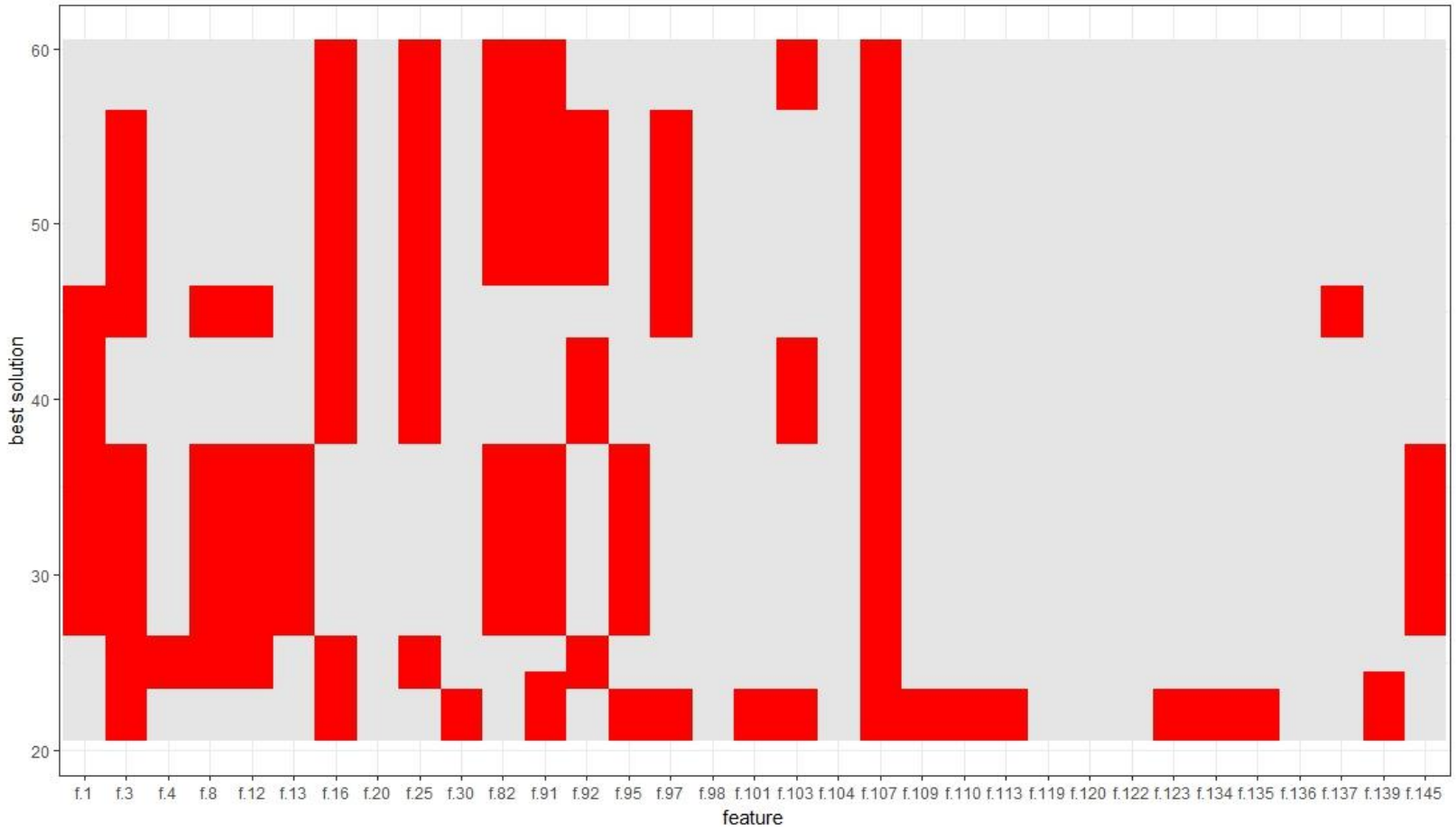
Parameters for the algorithm

PARAMETER	VALUES
# of generations	60
Size of population	50
Fitness function	median AUC
Method of Selection	Roulette wheel sampling
Method of cross-over	Single point crossover
Length of a chromosome	33
Rate of mutation	1/length of a chromosome
Split of data into training & test sets	80:20
Number of simulations for each feature set	10,000
Classification model	LDA

Performance



The best solutions



Conclusion

- GA are not bound by constraints such as continuity and differentiability of the objective function over the entire solution space.
- They may not be able to find the optimal solution, but can help find solutions that are good enough for the problem at hand.
- There are various other evolutionary ideas (such as use of diploid chromosomes) that have been adopted into GA which make them dynamic.
- With computing power becoming cheaper, these can be a viable method for searching in a solution space.

References

1. Coley, David A., An Introduction to Genetic Algorithms for Scientists and Engineers, 1999, World Scientific Publishing Co. Pte. Ltd.

2. Goldberg, David E., Genetic Algorithms in Search, Optimization and Machine Learning, 1989, Pearson Education Inc.

munshiimran.hossain@cytel.com