UBMK'19

Feature Selection with Evolving, Fast and Slow Using Two Parallel Genetic Algorithms

1st Uzay Cetin
Computer Engineering Department
Istanbul Bilgi University
Istanbul, Turkey
uzay00@gmail.com

2nd Yunus Emre Gundogmus*†
* *Tam Faktoring*
† *Statistics Department*
Istanbul Marmara University
Istanbul, Turkey
yemregun@gmail.com

Introduction

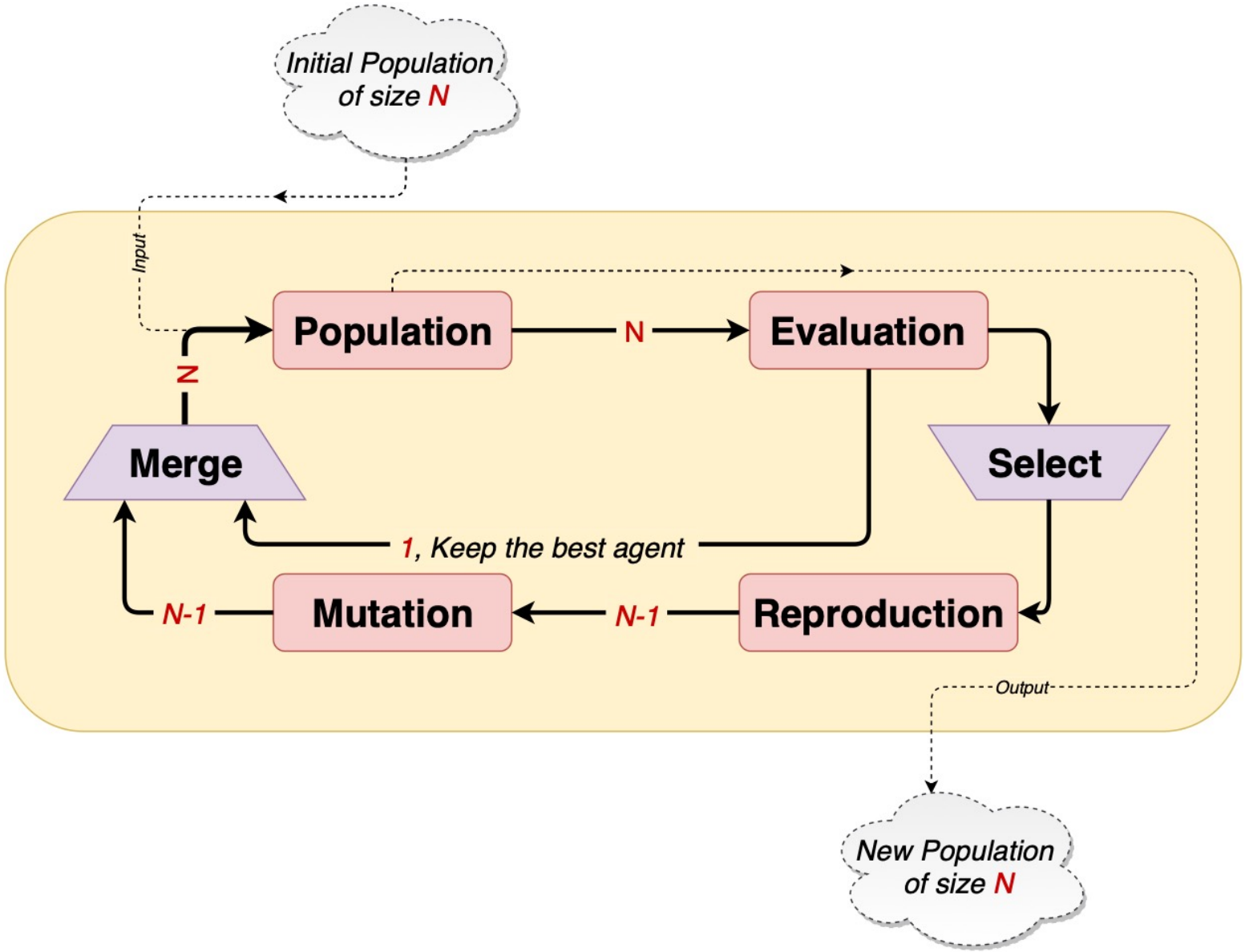
Problem:

- **Feature selection** is one of the most challenging issues in machine learning, especially while working with **high dimensional data**.

Proposed Solution :

- **Evolving Fast and Slow**,a new feature selection algorithm design based on using **two parallel genetic algorithms** having **high and low mutation rates**, respectively.

Standard Genetic Algorthm



Encoding

Genetic Algorthm For Feature Selection

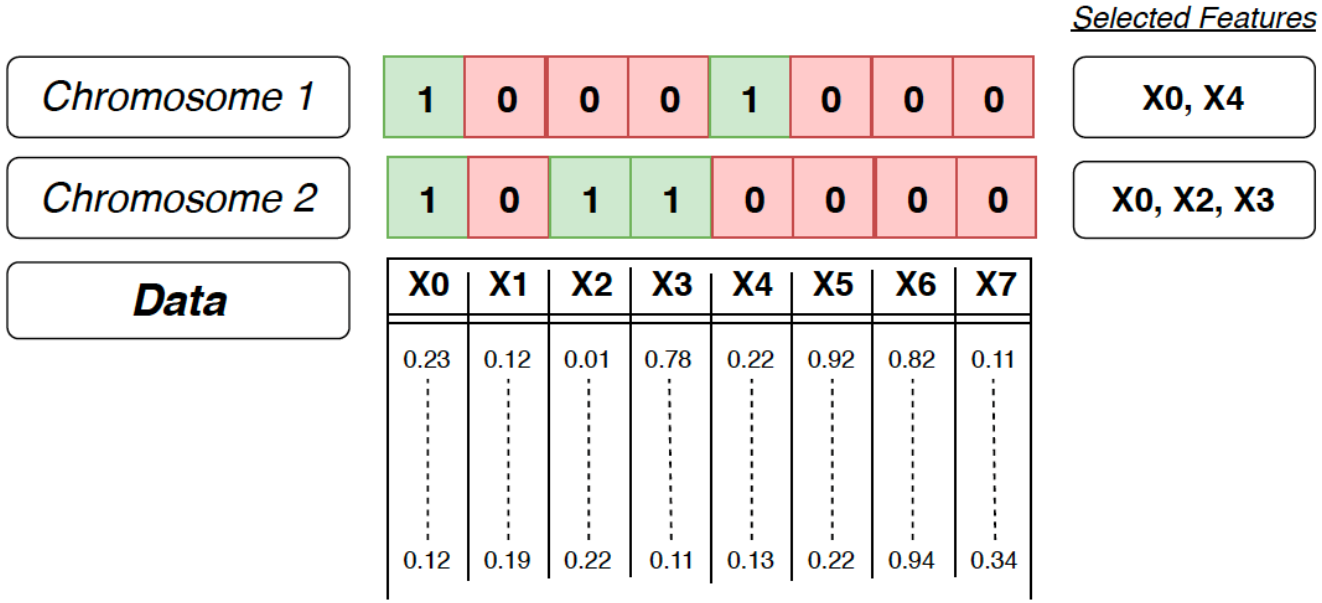


Fig. 2. Binary encoding for the chromosomes where 1 represents selected feature and 0 represents disregarded feature. In this example, *chromosome 1* has value of one at indices 0 and 4 which means 0^{th} and 4^{th} features are selected. In a similar manner, *chromosome 2* selects 0^{th}, 2^{nd} and 3^{rd} features.

Evaluation & Fitness

Genetic Algorthm For Feature Selection

$$fitness(x) = \alpha \times score(x) + (1 - \alpha) \times (1 - \frac{N_x}{N_{all}})$$

Score(x) : Cross validation accuracy score *(to be maximized)*
Nx / Nall : Ratio of selected features *(to be minimized)*

chromosome	score(x)	N_x	fitness(x)
x_1	0.80	400	0.70
x_2	0.80	1000	0.40
x_3	0.82	1000	0.41

TABLE I
FOR $\alpha = 0.5$ AND $N_{all} = 1000$

Reproduction

Genetic Algorthm For Feature Selection

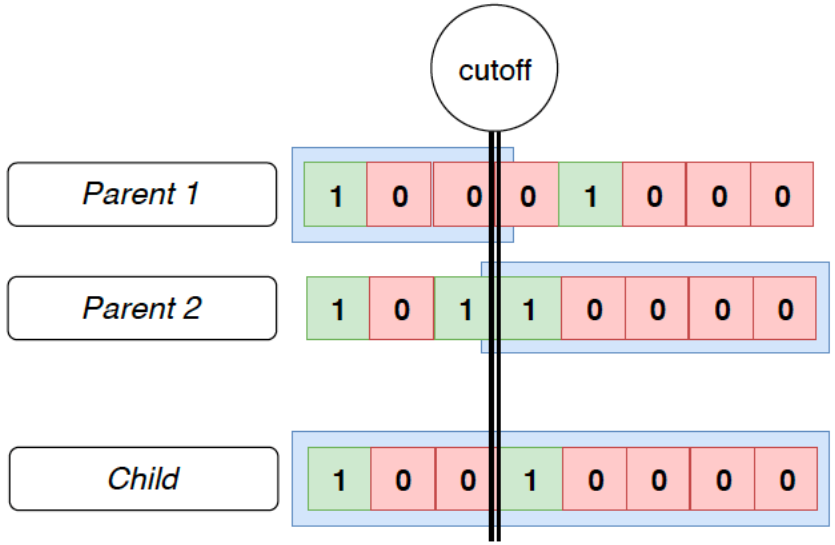


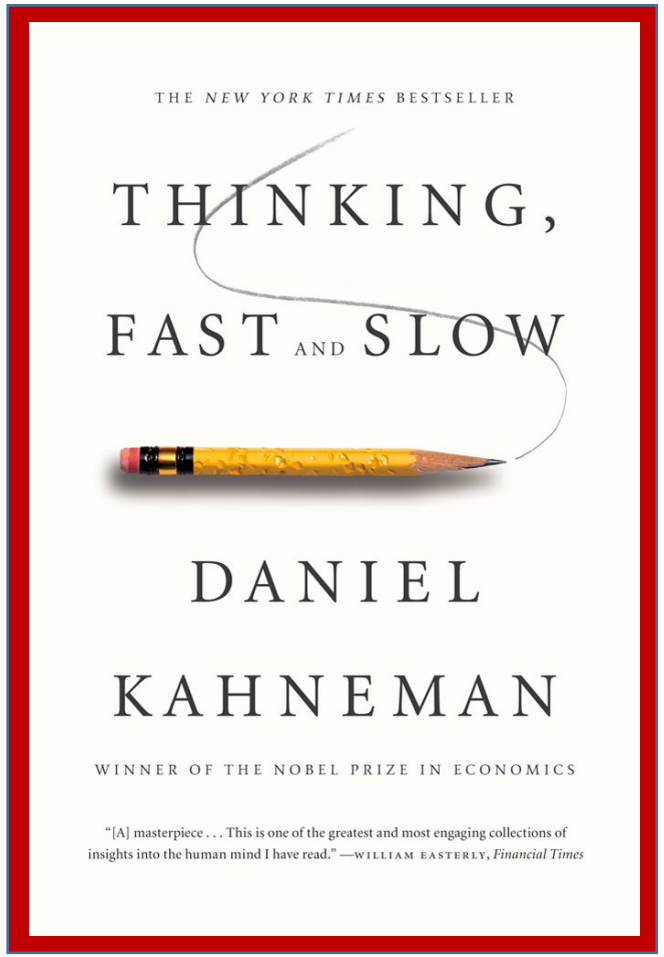
Fig. 3. Child gets the first part of the first parent's chromosome (from beginning to the cutoff point) and gets the second part of the second parent's chromosome (from the cutoff point to the end) during recombination.

Parents selected acc to fitness values. Recombination is followed by mutation.

EVOLVING, FAST AND SLOW

inspired from **Thinking, Fast and Slow** written by the world's most influential living psychologist Nobel Laureate Daniel Kahneman.

> "System 1 (Thinking Fast) continuously generates suggestions for System 2 (Thinking Slow): impressions, intuitions, intentions, and feelings. If endorsed by System 2, impressions and intuitions turn into beliefs, and impulses turn into voluntary actions."



EVOLVING, FAST AND SLOW

Exploration and Exploitation in Unison

New parallel architecture is a combination of

- an automatic system that evolves fast
- and an effortful system that evolves slow.

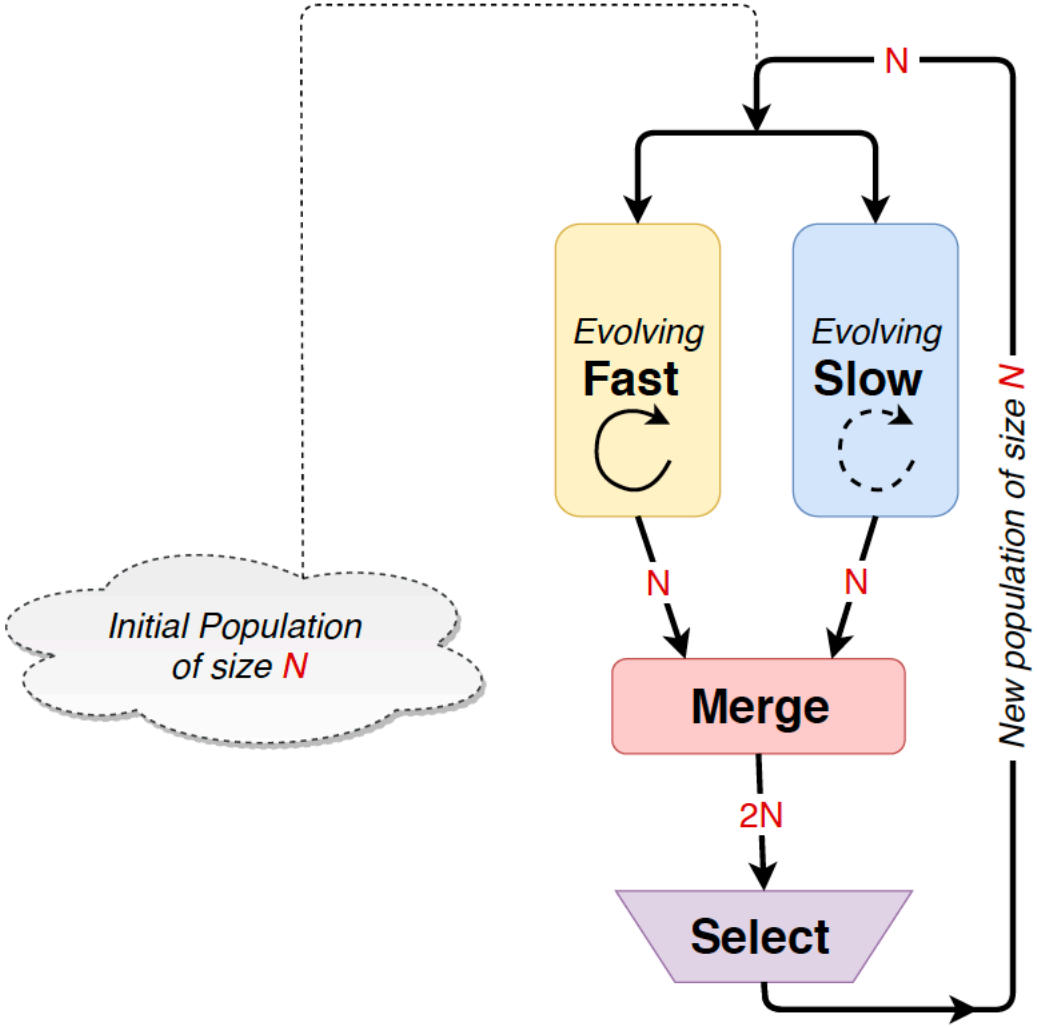


Fig. 4. Evolving, fast and slow with two parallel genetic algorithms having high and low mutation rates, respectively.

Results and Discussion

Toy Dataset

- 10000 sample, 50 feature. Y equals to 1 or 0 depending only the first 10 features. The remaining 40 features are noise.
- An efficient Feature selection algorithm should select only the first ten.

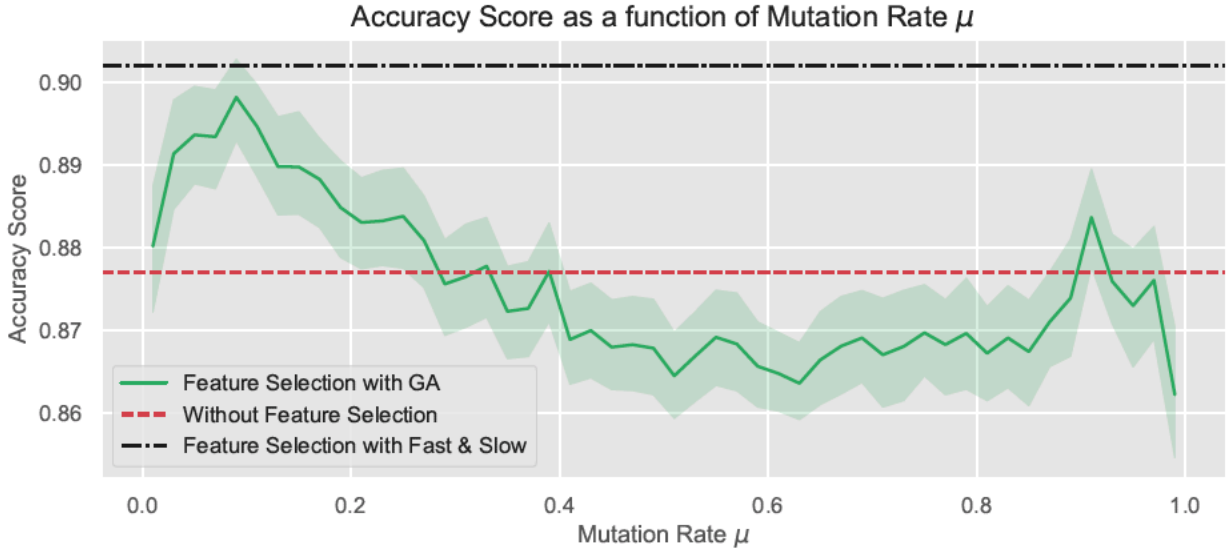


Fig. 5. Average accuracy score on the generated toy dataset.

Parameters:

- Alpha =1,
- number of generations = 20,
- average over 50 runs,
- mutation rates slow 0.1 and fast 1

Turkish political climate dataset

- opinion of the individual voters on a variety of political issues related to Turkish politics.
- Dataset has 885 rows and 14 columns.

TABLE II
POLITICAL DATASET WITH 2 PARTIES, AKP AND CHP

Political Dataset	Accuracy	Number of Features
Fast & Slow	80.93 ± 0.061	1.55 ± 0.739
GA $\mu = 0.10$	78.14 ± 0.072	1.4 ± 0.663
GA $\mu = 0.90$	$80.62 \pm 0,064$	2.3 ± 0.842
Without Feature Selection	85.95 ± 0.005	14

TABLE III
POLITICAL DATASET WITH 6 PARTIES

Political Dataset	Accuracy	Number of Features
Fast & Slow	38.55 ± 0.060	1.7 ± 0.714
GA $\mu = 0.10$	37.53 ± 0.049	1.1 ± 0.300
GA $\mu = 0.90$	36.97 ± 0.057	1.7 ± 0.714
Without Feature Selection	43.67 ± 0.013	14

Results are computed over 20 realizations for $\alpha = 0.9$ with Random Forest Classifier [15] in Tables II and III. The second row of the table shows the results for *Evolving Fast and Slow* where $\mu_{Slow} = 0.10$ and $\mu_{Fast} = 0.90$. The third and fourth

Tam Faktoring Cheque Dataset

- Results are again computed over 20 realizations for alpha = 0.9 for the financial dataset

Tam Faktoring Cheque Dataset has 50 different columns [14]. Each column represent different features, some of which are *cheque value, due date, customer's previous credit information, credit application count, all open credits balance, KKB credit score, open credit cards debt, etc. (all the customer data masked in accordance with KVKK(Personal Data Protection Law)).*

TABLE IV
FINANCIAL DATASET

Financial Dataset	Accuracy	Number of Features
Fast & Slow	67.73 ± 0.029	14.7 ± 1.873
GA $\mu = 0.10$	67.46 ± 0.023	13.65 ± 2.006
GA $\mu = 0.90$	67.89 ± 0.029	16.0 ± 2.0
Without Feature Selection	68.83 ± 0.006	50

Conclusions and Future Research

- We proposed a new feature selection algorithm design called ***Evolving Fast and Slow*** for **classification problems** based on using two parallel genetic algorithms having high and low mutation rates, respectively.
- Our approach can easily be ***extended to regression problems***.
- There are **several benefits** of genetic algorithms, first of all it is **inherently parallel** and can be easily **distributed**. It **always finds a solution**. With correct hyper-parameter settings, solutions found gets better.
- We leave a systematic search of hyper-parameters for different kind of real datasets as a future work.
- One of the key advantage of this work, is that this method ***can decide the number of features to be selected on its own***, whereas most of the existing feature selection algorithms requires the number of selected features as an input.
- Apart from feature selection***, algorithm selection can also be done via genetic algorithms***. (Auto ML)