Universidad Nacional de Asunción
Facultad Politécnica

CONSEJO NACIONAL DE CIENCIA Y TECNOLOGÍA
CONACYT
CONT: 064/2017 - PINV15-0257

Machine Learning Algorithm for Features Selection Problem

Authors

Eng. Brenda Quiñonez

PhD. Miguel García-Torres

PhD. Diego Pinto

PhD. Carlos Núñez

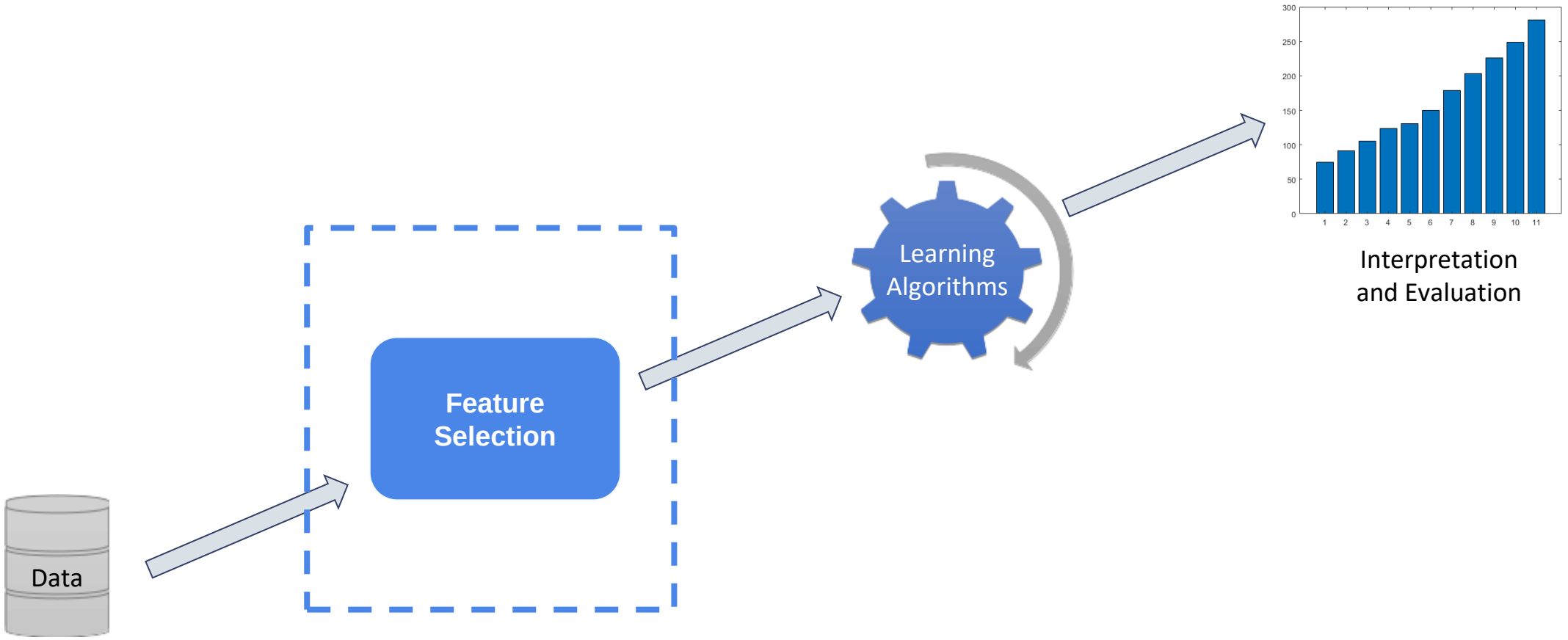
M.Sc. María Elena García

Ph.D. Federico Divina

Index

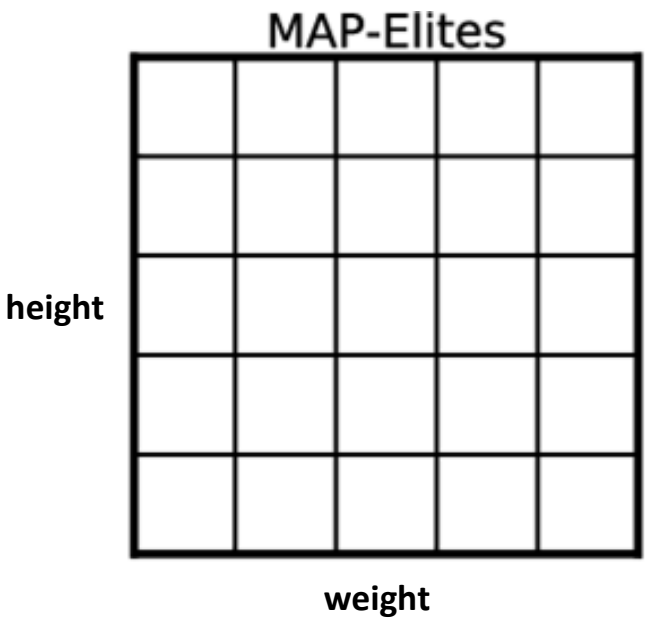
- Feature Selection Problem
- MAP-Elites Algorithm
- Challenges to apply MAP-Elites to Feature Selection
- MAP-Elites-Combinatorial Algorithm
- MAP-Elites Result Experiment
- Future Works

What is Feature Selection?



What is Feature Selection?

MAP-Elites



Grid of cells of 2-dimensions

- Create a search space of user-defined features by discretizing it into a grid.
- The cells are progressively filled with solutions x according to their position in the search space.
- Replacing any solution in the cell only if the new solution is better according to some user-defined quality measure f(x).

MAP-Elites Algorithm

1) Create an empty **N-dimensional map** of cells
2) **for I** iterations
    a) **if (i < G)**
        i) generate a x' = randomSolution()
    b) else randomly select a solution x from the map
        i) create a modified copy of x, x' (via mutation/crossover)
        ii) mapping the solution x' into a cell
    c) compute performance of f(x')
        i) if f(x') is better than cell.f(x)
            (1) set cell to x'
            (2) set cell.f(x) to f(x')

Challenges to Apply MAP-Elites to FS

- How could we represent the binary variables of feature selection problems?
- How could we divide the MAP-Elites search space for the binary variables?
- What evaluation function can we use to qualify the solutions?

MAP-Elites Combinatorial for FS

- Represent the set of solution as a vector with the indexes of the selected features.

- Define the number of cells as an input parameter (NC). Later, NC is used to calculate the number of fixed features per cell (NFF), which are used as cell identifiers.

- By last, we use the accuracy of the classifier as a function to measure the quality of the subset solution.

Combinatorial MAP-Elites Algorithm

1) **NFF = $\log_2$(NC)**
2) **createMap(NC, NFF)**
3) for **I** (with iterator i)
    a) if (i < **G**)
        i) generate a x' = randomSolution()
    b) else randomly select a solution x from the map
        i) create a modified copy of x, x' (via mutation/crossover)
        ii) mapping the solution x' into a cell in the feature space
    c) compute **accuracy of classifier** f(x')
        i) if f(x') is better than cell.f(x)
            (1) set cell to x'
            (2) set cell.f(x) to f(x')

Map-Elites Result Experiment

Dataset	All features	Fitness	Selected features
ionosphere	34	92.02 ± 2.38	12.6 ± .89
glass	9	70.12 ± 4.27	6.0 ± 1.00
anneal	38	96.44 ± 1.60	7.6 ± 1.34
tokyo1	44	92.91 ± 1.08	10.6 ± 2.51
spambase	57	91.76 ± .60	10.6 ± .89
kr-vs-kp	36	90.43 ± 1.46	3.0 ± .00
corral	6	86.90 ± 2.22	5.0 ± .00
breast-cancer	9	71.34 ± 3.95	3.6 ± .89
hypothyroid	29	96.66 ± .29	1.0 ± .00
labor	16	91.21 ± 11.14	5.0 ± .71
vote	16	95.63 ± 1.26	1.0 ± .00

What is Next?

- Compared MAP-Elites with another feature selections algorithms of the state of the art.
- Using another accuracy classifier as a function to qualify the solutions besides Bayes classifier.
- Using differents high-dimensional dataset with more than 2000 features.

Thank you!

Any questions?

Bibliography

[1] Miao, Jianyu Niu, Lingfeng. (2016). A Survey on Feature Selection. Procedia Computer Science. 91. 919-926. 10.1016/j.procs.2016.07.111.

[2] Jean-Baptiste Mouret and Jeff Clune. “Illuminating search spaces by mapping elites”. CoRR. 2015.vol. abs/1504.04909

[3] F. García López, M. García-Torres, B. Melian Batista, J. A. Moreno Pérez, and J. Marcos Moreno-Vegatitle. “Solving feature subset selection problem by a Parallel Scatter Search”. European Journal of Operational Research. 2006.

[4] M. Garcia-Torres, F. Gomez-Vela, B. Melian, J.M. Moreno-Vega. Highdimensional feature selection via feature grouping: A Variable Neighborhood Search approach, Information Sciences, vol. 326, pp. 102-118, 2016.