Dynamic Genotype Reduction for Narrowing the Feature Selection Search Space

Sašo Karakatič, Iztok Fister Jr., Dušan Fister

Faculty of Electrical Engineering and Computer Science,
University of Maribor, Koroška cesta 46, 2000 Maribor, Slovenia

Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko

Introduction

- The common **problem** of wide-datasets in data mining (DM): too many features with sometimes not enough of instances (cases).
    - Multicollinearity of features
    - Redundant features
    - Irrelevant features

- **Consequence**: pattern extraction (e.g. classification algorithms) perform worse and training of models takes a long time.

Introduction

- **Feature selection** (FS) is a process that reduces the dimensionality of data – selects only appropriate and informative features.
  - Filter based – performs FS separately from DM
  - Wrapper based – combines FS and DM
  - Embedded based – combination of both

- **Nature-inspired** (NIA) meta-heuristic optimization methods have been successfully used for FS.
  - Problems: takes a long time for FS and can get stuck in local optima (sub-optimal set of selected features).

Contribution

- Dynamically reducing gene representation to eliminate redundant, low-variance, inexpensive or irrelevant features completely during the optimization.

- Attaining the feature selection search space to narrow accordingly, keeping the more expensive features for further space search and at the same time decreasing the difficulty of the problem.

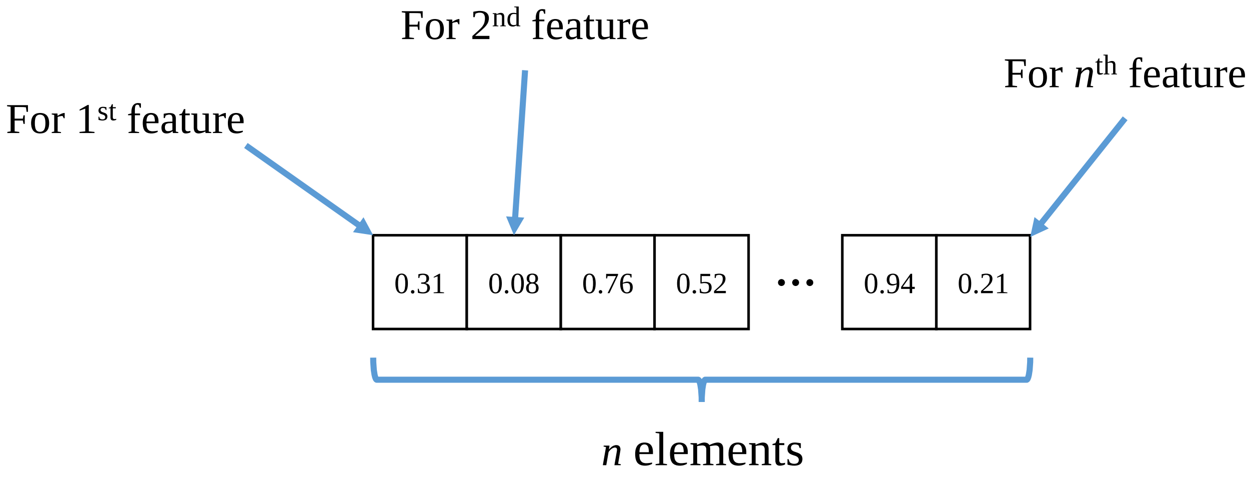
Dynamic feature selection (DynFS)

- Solving the large search-space problem with dynamic reduction of it.
  - Search space with least potential is iteratively disregarded during the optimization process. Thus, minimizing the chances of getting stuck in local optima and maximizing the focus in search-space with the most potential.

- Hypotheses:
  - (1) potentially improve the optimization process, or make it less difficult,
  - (2) subsequently shorten the computation time needed for the search toward the global optimum.

Genotype representation

- **Regular FS genotype**: fixed-length array of *n* real values, where *i*-th element represent one feature from the dataset.
- The values in the array are in the interval [0, 1], and threshold *T* helps the mapping of each value to either presence (above *T*) or absence (below *T*) of that *i*-th feature.



if $T == 0.5$

1st value is $\boxed{0.31}$, which is below *T* and thus the 1st feature is not selected.

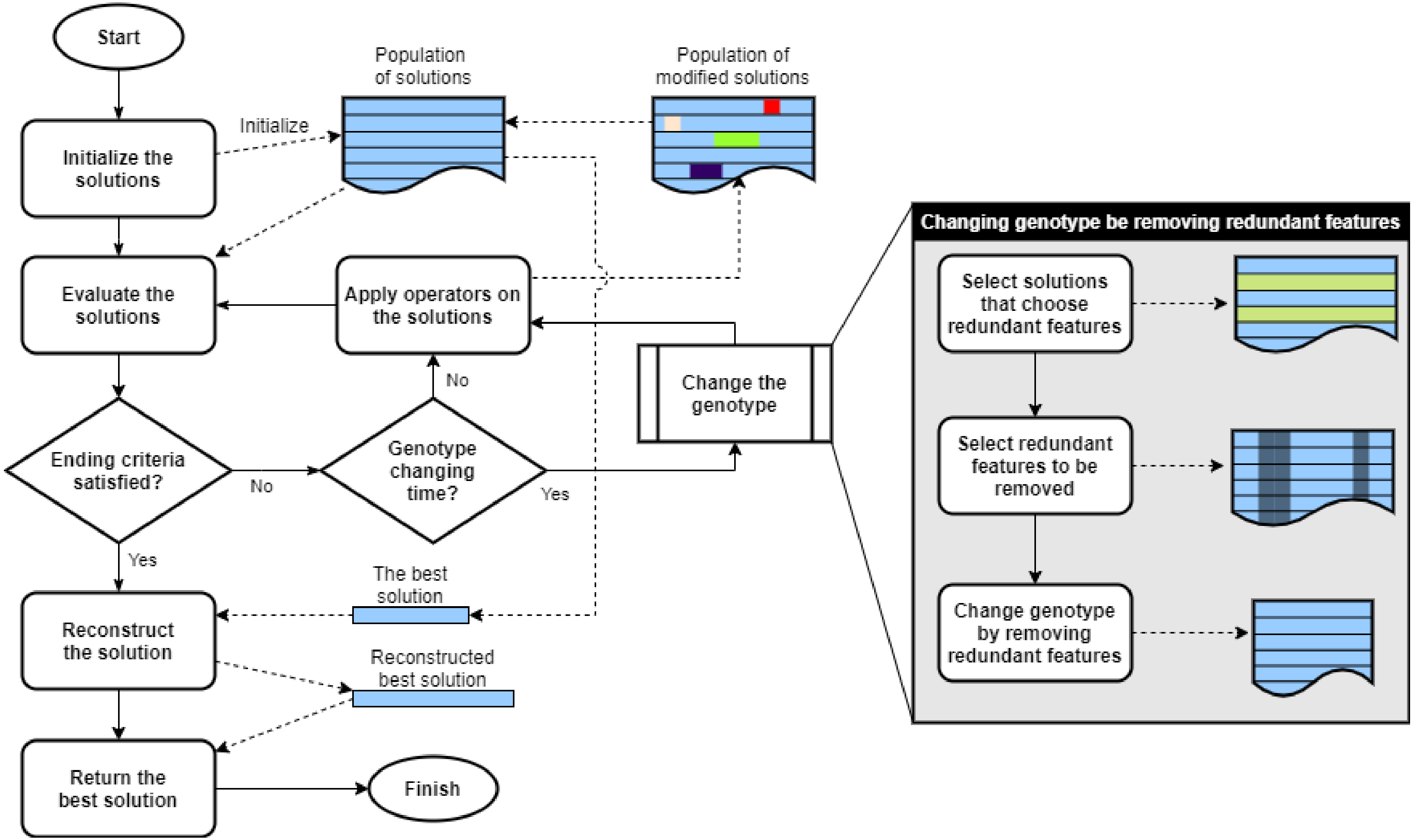
Dynamic genotype

1. The best *W* performing solutions (FS sets) are selected and they vote for *F* worst performing features.
    - The voting rules are: whether the feature is present in a particular solution, it gets a vote from that solutions.

2. Features with the least votes represent the features that are least common in best performing feature sets and are considered as the features that make the FS worse off.

3. Worst performing features are removed from the genotype – the genotype size becomes smaller, as does the search-space.

Dynamic genotype - overview



Nature-inspired optimization with jDE

- NIA algorithm was **self-adaptive differential solution jDE** implemented in **EvoPreprocess** and **NiaPy**.

- Each solution is extended with scale factor F and crossover rate CR that undergo the variation operators, mathematically represented as:

$$\mathbf{x}_i^{(t)} = (x_{i,1}^{(t)}, x_{i,2}^{(t)}, ..., x_{i,M}^{(t)}, F_i^{(t)}, CR_i^{(t)}).$$

- Both parameters are modified according to the following equations:

$$F_i^{(t+1)} = \begin{cases} F_l + \text{rand}_1 \cdot (F_u - F_l) & \text{if } \text{rand}_2 < \tau_1, \\ F_i^{(t)} & \text{otherwise}, \end{cases}$$

$$CR_i^{(t+1)} = \begin{cases} \text{rand}_3 & \text{if } \text{rand}_4 < \tau_2, \\ CR_i^{(t)} & \text{otherwise}, \end{cases}$$

Experiment setup

- Arrhythmia dataset
  - 279 features, with values ranging from negative to positive integer and floating-point values.
  - The 452 instances of the dataset are classified into 16 classes, with the majority falling into the “normal” class and the rest incorporating any cardiac disorders.
  - Split into 5 folds for cross-validation.

No.	Feature	No.	Feature
1.	Age	*Of channels**	
2.	Gender	28.-39.	DII
3.	Height	40.-51.	DIII
4.	Weight	52.-63.	AVR
5.	QRS duration	64.-75.	AVL
6.	P-R interval	76.-87.	AVF
7.	Q-T interval	88.-99.	V1
8.	T interval	100.-111.	V2
9.	P interval	112.-123.	V3
Vector angles		124.-135.	V4
10.	QRS	136.-147.	V5
11.	T	136.-147.	V5
12.	P	148.-159.	V6
13.	QRST	*Of channel DI*	
14.	J	160.	JJ wave
15.	Heart rate	161.	Q wave
Of channel DI		162.	R wave
16.	Q wave average width	163.	S wave
17.	R wave average width	164.	R’ wave
18.	S wave average width	165.	S’ wave
19.	R’ wave average width	166.	P wave
20.	S’ wave average width	167.	T wave
21.	No. of intrinsic deflections	168.	QRSA
22.	Existence of ragged R wave	169.	QRSTA
23.	Existence of diphasic derivation of R wave	*Of channels*** 170.-179.	 DII
24.	Existence of ragged P wave	180.-189.	DIII
25.	Existence of diphasic derivation of P wave	190. 199. 200.-209.	AVR AVL
26.	Existence of ragged T wave	210. 219.	AVF
27.	Existence of diphasic derivation of T wave	220.-279.	V1, V2, V3 V4, V5, V6

Note: “*” means similar as for 16.-27. features and “**” means similar as 160.-169. features.
Source: UCI Machine Learning Repository.

Experiment setup

- jDE with fitness function of (*1- F_1score*), where

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

- DynFS settings:
  - Genotype cuttings at [512, 256, 126, 64]
  - Total of 960 generations.
  - W = 50 best performing solutions vote for the removal of 10% of features.

- Other parameter setting:

Parameter	**Symbol**	**Value**
Genotype changing margins	M	512, 256, 126, 64
Total no. of generations	GEN	960
No. of best performing solutions	W	50
Removal rate	R	10 %
Initial crossover rate	CR	0.8
Initial scaling factor	F	1.0
No. of classification folds	k	5
Training-validation split		50%-50%

- CART classification used for the evaluation of solutions.
- Comparison to EvoFS – nature-inspired FS without dynamic genotype.

Results

	CART		EvoFS		DynFS	
Fold	**Acc**	F_1	**Acc**	F_1	**Acc**	F_1
1	96.70	66.67	**98.90**	**90.91**	**98.90**	**90.91**
2	**95.60**	50.00	**95.60**	**60.00**	**95.60**	50.00
3	94.44	44.44	**96.67**	**66.67**	95.56	50.00
4	**97.78**	**80.00**	93.33	57.14	95.56	60.00
5	93.33	50.00	**95.56**	33.33	**95.56**	**60.00**
Mean	95.57	58.22	96.01	61.61	**96.23**	**62.18**
Std. dev.	0.016	0.132	0.018	0.185	**0.013**	**0.150**
Avg. rank	2.2	2.2	**1.4**	1.8	**1.4**	**1.6**

- DynFS achieved the best results in three out of five folds accuracy-wise and in two out of five folds F1-score-wise.
- Overall, it also achieved the best means in accuracy and F1-score, with the smallest standard deviation (which means that the results are more robust).
- Also, the average ranks are the best for both accuracy and F1-score.

Conclusions

- **EvoFS** consistently chooses a single feature – the *J vector angle* as most important, while *existence of the diphasic derivation of T wave in channel DI* was the next most frequently chosen feature.
- **DynFS** extended this, with *R wave average width in V4 channel* and *S' wave in DII channel*.
- The results show that the periodic reductions of the genotype result in superiority, compared to similar optimization techniques, without the genotype reductions implemented.
- The idea of shrinking the search space proves to be promising.