Improved Nature-inspired algorithms for Numeric Association Rule Mining

Iztok Fister Jr., Vili Podgorelec, Iztok Fister

March 8, 2022

Agenda

1. Contributions
2. ARM
3. NARM-DE
4. Results

Contributions of this study

- A new algorithm is proposed for the pure numerical association rule mining (NARM).
- A new evaluation function is identified.
- The algorithm is applied to a sport dataset consisting of pure numeric attributes.

Association rule mining

An association rule can be defined as implication:

$$X \Rightarrow Y, \tag{1}$$

where $X \subset O$, $Y \subset O$, in $X \cap Y = \emptyset$. The following two measures are defined for evaluating the quality of association rule:

$$conf(X \Rightarrow Y) = \frac{n(X \cup Y)}{n(X)}, \tag{2}$$

$$supp(X \Rightarrow Y) = \frac{n(X \cup Y)}{N}, \tag{3}$$

where $conf(X \Rightarrow Y) \geq C_{min}$ denotes confidence and $supp(X \Rightarrow Y) \geq S_{min}$ support of association rule $X \Rightarrow Y$.

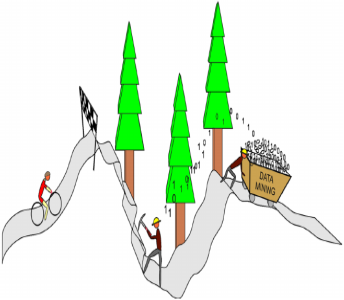
Differential evolution

- Differential Evolution (DE) is an Evolutionary Algorithm.
- DE is appropriate for continuous as well as combinatorial optimization.
- DE is a population-based algorithm that consists of Np real-coded vectors representing the candidate solutions.
- Operators in DE: differential mutation, differential crossover, differential selection.

NARM-DE

Development of NARM-DE algorithm consists of the following stages:

- domain analysis,
- representation of solution and
- definition of a fitness function.



Domain analysis

Table: Domain analysis performed on the sport database.

Attribute	Minimum lower bound	Maximum upper bound
Duration	107.95	142.40
Distance	8.76	85.19
Average_HR	63.00	168.00
Average_ALT	7.23	1779.04
Calories	273.00	2243.00
Ascent	6.0	1884.40
Descent	2.0	1854.20

Representation of solutions

$$x_i^{(t)} = \{\underbrace{\langle x_{i,\mathbf{4}\pi_j}^{(t)}, x_{i,\mathbf{4}\pi_j+\mathbf{1}}^{(t)}, x_{i,\mathbf{4}\pi_j+\mathbf{2}}^{(t)}, x_{i,\mathbf{4}\pi_j+\mathbf{3}}^{(t)}\rangle}_{\boldsymbol{At}_j^{(t)}}, \ldots, \underbrace{x_{i,\mathbf{4}D}^{(t)}}_{\boldsymbol{Cp}_i^{(t)}}\rangle\}, \tag{4}$$

where elements $x_{i,\mathbf{4}\pi_j+k}^{(t)}$, for $i = 1, \ldots, Np$ and $j = 0, \ldots, D-1$ and $k = 0, \ldots, 3$, denote the attributes of features in association rules, t is an iteration counter, and D the number of attributes. Indeed, each numerical attribute is expressed as a quadruple:

$$At_{i,j}^{(\mathbf{t})} = \langle x_{i,\mathbf{4}\pi_j}, x_{i,\mathbf{3}\pi_j+\mathbf{1}}, x_{i,\mathbf{4}\pi_j+\mathbf{2}}, x_{i,\mathbf{4}\pi_j+\mathbf{3}}\rangle, \tag{5}$$

where the first term denotes the lower bound, the second the upper bound, the third the threshold value, and the fourth determines the ordering of the attribute in the permutation.

Definition of the fitness function

Fitness function is expressed as:

$$f(\mathrm{x}_i^{(t)}) = \frac{\alpha \cdot supp(\mathrm{x}_i^{(t)}) + \beta \cdot conf(\mathrm{x}_i^{(t)}) + \gamma \cdot inclusion(\mathrm{x}_i^{(t)}) + \delta \cdot K}{\alpha + \beta + \gamma + \delta}, \tag{6}$$

where $supp(\mathrm{x}_i^{(t)})$ and $conf(\mathrm{x}_i^{(t)})$ represent the support and confidence of the observed association rule, K is the shrink factor, and $inclusion(\mathrm{x}_i^{(t)})$ is defined as follows:

$$inclusion(X \Rightarrow Y) = \frac{|ante(X \Rightarrow Y)| + |cons(X \Rightarrow Y)|}{m}, \tag{7}$$

where $ante(X \Rightarrow Y)$ returns the number of attributes in the antecedent, $cons(X \Rightarrow Y)$ is the number of attributes in consequence of the particular association rule, and m is the total number of attributes. Weights in Eq. (6) are set to $\alpha = \beta = \gamma = \delta = 1$ in the study.

Example results

Table: Number of rules found using different algorithms.

Algorithm	DE	PSO	CS	FPA
Total rules	**241,455**	212,352	74,069	146,659
Average number of Ant/Cons	5/2	5/2	**4/3**	**3/4**
Average Fitness	**0.7506**	0.6519	0.1165	0.1858
Average Support	0.8242	**0.8729**	0.1413	0.1181
Average Confidence	0.9637	**0.9812**	0.5575	0.4106
Average Shrink	0.2639	**0.1703**	0.2048	0.2771
Average Inclusion	**0.9722**	0.8370	0.3260	0.4454

Selected solutions

Table: Examples of solutions found by the proposed algorithms.

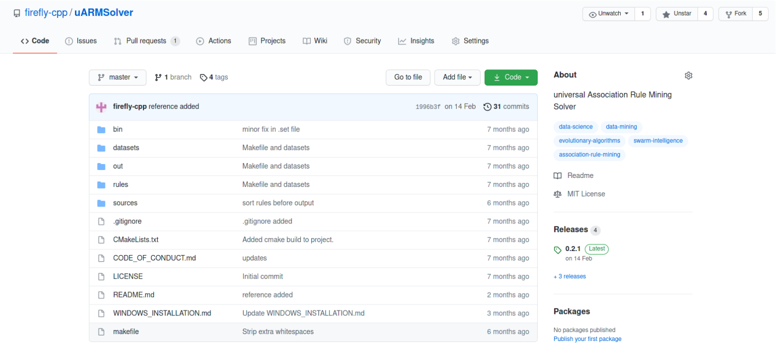
Antecedent	Cut	Consequent
CAL[350.83, 1247.60]$\wedge$ ALT(NO)$\wedge$ DUR[107.95, 142.40]$\wedge$ DESC[312.42, 1409.12]	$\Longrightarrow$	DIST[14.04, 85.19]$\wedge$ ASC[259.34, 1884.40]
ALT[338.67, 589.31]$\wedge$ DUR[131.29, 135.80]$\wedge$ AVHR[63.0, 125.48]$\wedge$ CAL[273.0, 1498.70]$\wedge$ ASC[859.13, 1445.86]	$\Longrightarrow$	DIST[8.76, 85.19]$\wedge$ DESC[774.99, 1258.80]
ALT[7.22, 1134.88]$\wedge$ CAL[440.82, 1966.86]$\wedge$ ASC[6.0, 1503.78]$\wedge$ AVHR[86.16, 158.17]$\wedge$ DIST[17.43, 69.82]	$\Longrightarrow$	DESC[2.0, 1598.18]$\wedge$ DUR[107.95, 142.4]

Conclusion 1/2

- ARM with numerical attributes is a very challenging problem.
- Our proposed solution is based on DE for mining a combination of numeric attributes.
- There are still many future challenges:
  - testing proposed approach on bigger transaction databases,
  - identification of proper weights.

Conclusion 2/2

Software is available on GitHub.



Questions

Please send all questions on:

iztok.fister1@um.si