Improved Nature-Inspired Algorithms for Numeric Association Rule Mining

Iztok Fister Jr.(✉), Vili Podgorelec, and Iztok Fister

Faculty of Electrical Engineering and Computer Science, University of Maribor,
Koroška cesta 46, 2000 Maribor, Slovenia
iztok.fister1@um.si

Abstract. Nowadays, only a few papers exist dealing with Association Rule Mining with numerical attributes. When we are confronted with solving this problem using nature-inspired algorithms, two issues emerge: How to shrink the values of the upper and lower bounds of attributes properly, and How to define the evaluation function properly? This paper proposes shrinking the interval of attributes using the so-called shrinking coefficient, while the evaluation function is defined as a weighted sum of support, confidence, inclusion and shrink coefficient. The four nature-inspired algorithms were applied on sport datasets generated by a random generator from the web. The results of the experiments revealed that, although there are differences between selecting a specific algorithm, they could be applied to the problem in practice.

Keywords: Association rule mining · Numerical attributes · Nature-inspired algorithms · Optimization

1 Introduction

Association Rule Mining (ARM) is used for discovering the dependence rules between features in a transaction database. On the other hand, Numeric Association Rule Mining (NARM) extends the idea of ARM, and is intended for mining association rules where attributes in a transaction database are represented by numerical values [4]. Usually, traditional algorithms, e.g. Apriori, requires numerical attributes to be discretized before use. Discretization is sometimes trivial, and sometimes does not have a positive influence on the results of mining. On the other hand, many methods exist for ARM that do not require the discretization step before applying the process of mining. Most of the these methods are based on population-based nature-inspired metaheuristics, such as, for example, Differential Evolution or Particle Swarm Optimization. NARM has recently also been featured in some review papers [3,7] which emphasize its importance in the data revolution era.

The objective of this short paper is to extend the paper of Fister et al. [5], where the new algorithm for NARM was proposed, based on Differential Evolution. Indeed, the practical experiments revealed some problems/bottlenecks that can be summarized into two issues:

P. Vasant et al. (Eds.): ICO 2020, AISC 1324, pp. 187–195, 2021.
https://doi.org/10.1007/978-3-030-68154-8_19

- How to shrink the lower and upper borders of numerical attributes?
- How to evaluate the mined rules better?

Each numerical attribute is determined by an interval of feasible values limited by its lower and upper bounds. The broader the interval, the more association rules mined. The narrower the interval, the more specific relations between attributes are discovered. Mined association rules can be evaluated according to several criteria, like support and confidence. However, these cover only one side of the coin. If we would also like to discover the other side, additional measures must be included into the evaluation function. The paper is focused on developing the algorithm for the pure NARM. In line with this, a new evaluation function needs to be proposed. As a result, the main contribution of this paper can be extracted into the following indents:

- the new algorithm is proposed for the pure NARM,
- the new evaluation function is identified,
- the algorithm is applied to a sport dataset consisting of pure numeric attributes.

The remainder of the paper is structured as follows: Sect. 2 highlights background information needed for understanding a subject. In Sect. 3, improved algorithm is presented. Section 4 outlines the experiments as well as presents results. Paper is wrapped up with a conclusion in Sect. 5.

2 Background Information

2.1 Association Rule Mining

This section presents briefly a formal definition of ARM. Let us suppose, a set of objects $O = \{o_1, \ldots, o_m\}$, where m is the number of objects, and transaction set D is given, where each transaction T is a subset of objects $T \subseteq O$. Then, an association rule can be defined as the 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 the association rule [2]:

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

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

where $\mathit{conf}(X \Rightarrow Y) \geq C_{min}$ denotes confidence and $\mathit{supp}(X \Rightarrow Y) \geq S_{min}$ support of association rule $X \Rightarrow Y$. Thus, N in Eq. (3) represents the number of transactions in transaction database D, and $n(.)$ is the number of repetitions of the particular rule $X \Rightarrow Y$ within D. Here, C_{min} denotes minimum confidence and S_{min} minimum support determining that only those association rules with confidence and support higher than C_{min} and S_{min} are taken into consideration, respectively.

2.2 Differential Evolution for ARM Using Mixed Attributes

The basis of our study presents Differential Evolution for ARM using mixed attributes (ARM-DE) proposed by Fister et al. in [5]. Development of this algorithm is divided into three steps:

- domain analysis,
- representation of solution,
- evaluation function definition.

In that study, domain analysis of the observed sport database identified three numerical and eleven categorical attributes. Thus, the former are determined with intervals of feasible values limited by their minimum lower and maximum upper bound values.

The individuals in the population are represented as real-valued vectors, where every numerical attribute consists of corresponding lower and upper bounds. On the other hand, each categorical attribute is identified by the real-value drawn from the interval $[0, 1]$ that is mapped to the corresponding discrete attribute according to dividing the interval into equally-sized non-overlapping sub-intervals. Additionally, the last element in the representation $x_{i,D}$ denotes the so-called cut point, determining which part of the vector belongs to the antecedent and which to the consequent of the mined association rule. The following fitness function was defined for evaluation of solutions:

$$f(\mathbf{x}_i^{(t)}) = \begin{cases} \alpha * \mathit{conf}(\mathbf{x}_i^{(t)}) + \gamma * \mathit{supp}(\mathbf{x}_i^{(t)})/\alpha + \gamma, & \text{if } \mathit{feasible}(\mathbf{x}_i^{(t)}) = \mathit{true}, \\ -1, & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathit{conf}(.)$ is confidence, $\mathit{supp}(.)$ support, α and γ are weights, function $\mathit{feasible}(\mathbf{x}_i)$, denotes if the solution is feasible. The task of the optimization is to find the maximum value of the evaluation function.

3 Improved DE for NARM

The proposed DE for NARM (also NARM-DE) operates on the transaction database using only numerical attributes containing data obtained from a wearable device during sport training. Therefore, the new domain analysis needs to be performed in a first step. The result of this step is illustrated in Table 1, from which it can be seen that domain analysis of this database identified seven numerical attributes characterizing a performance of a realized training session. To each attribute, the corresponding intervals with their minimum lower and maximum upper bounds are assigned in the Table.

Then, the representation of solutions must be adjusted to the new demands. Here, the solutions are represented as real-valued vectors, in the following form:

$$\mathbf{x}_i^{(t)} = \{\underbrace{\langle x_{i,4\pi_j}^{(t)}, x_{i,4\pi_j+1}^{(t)}, x_{i,4\pi_j+2}^{(t)}, x_{i,4\pi_j+3}^{(t)} \rangle}_{At_j^{(t)}}, \ldots, \underbrace{x_{i,4D}^{(t)}}_{Cp_i^{(t)}})\}, \tag{5}$$

Table 1. 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

where elements $x^{(t)}_{i,4\pi_j+k}$, 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:

$$At^{(t)}_{i,j} = \langle x_{i,4\pi_j}, x_{i,3\pi_j+1}, x_{i,4\pi_j+2}, x_{i,4\pi_j+3} \rangle, \tag{6}$$

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.

The threshold value determines the presence or absence of the corresponding numerical feature in the association rule, in other words:

$$At^{(t)}_{\pi_j} = \begin{cases} \begin{cases} [NULL, NULL], & \text{if rand}(0,1) < \mathrm{x^{(t)}_{i,4\pi_j+2}}, \end{cases} \\ \begin{cases} [x^{(t)}_{i,4\pi_j}, x^{(t)}_{i,4\pi_j+1}], & \text{if } \mathrm{x^{(t)}_{i,4\pi_j}} > \mathrm{x^{(t)}_{i,4\pi_j+1}}, \\ [x^{(t)}_{i,4\pi_j+1}, x^{(t)}_{i,4\pi_j}], & \text{otherwise}, \end{cases} \end{cases} \tag{7}$$

where a shrinking coefficient K is expressed as:

$$K = \left(1 - \frac{\left|x^{(t)}_{i,4\pi_j} - x^{(t)}_{i,4\pi_j+1}\right|}{Ub_{\pi_j} - Lb_{\pi_j}}\right), \tag{8}$$

and Lb_{π_j} and Ub_{π_j} denote the corresponding lower and upper bounds. The motivation behind the proposed equation is to shrink the whole interval of the feasible values.

A permutation $\Pi = (\pi_1, \ldots, \pi_D)$ is assigned to each solution $\mathbf{x}^{(t)}_i$, which orders the attributes $At^{(t)}_j$ according to the following equation:

$$x^{(t)}_{i,4\pi_0+3} \geq x^{(t)}_{i,4\pi_j+3} \geq x^{(t)}_{i,4\pi_{D-1}+3}, \quad \text{for } j = 0, \ldots, D-1. \tag{9}$$

Thus, the attributes with the higher value of the fourth element $x^{(t)}_{i,4\pi_j+3}$ are ordered at the start of the permutation, while the attributes with the lower values at the end of the permutation. In this case, each numerical attribute has an equal chance to be selected as an antecedent or consequent of the mined association rule.

The last element in the vector determines the cut point $Cp_i^{(t)}$, expressed as:

$$Cp_i^{(t)} = \left\lfloor x_{i,D}^{(t)} \cdot (D-2) \right\rfloor + 1, \tag{10}$$

where $Cp_i^{(t)} \in [1, D-1]$. In summary, the length of solution vector is the $4 \cdot D + 1$.

The mapping of the solution representation to the corresponding association rule is expressed as follows:

$$\begin{aligned} Ante(X \Rightarrow Y) &= \{o_{\pi_j} | \pi_j < Cp_i^{(t)} \wedge At_{\pi_j}^{(t)} \neq [NULL, NULL]\}, \\ Cons(X \Rightarrow Y) &= \{o_{\pi_j} | \pi_j \geq Cp_i^{(t)} \wedge At_{\pi_j}^{(t)} \neq [NULL, NULL]\}, \end{aligned} \tag{11}$$

where $Ante(X \Rightarrow Y)$ represents a set of objects belonging to antecedent and $Cons(X \Rightarrow Y)$ is a set of objects belonging to consequent of the corresponding association rule. However, the attribute needs to be enabled in order the object to be valid member of the particular set.

Finally, an evaluation function must be defined. As found in the experimental work of Fister et al. [5], however, the main weakness of the ARM-DE was reflected in the fact that the evaluation function consisted of a linear combination of support and confidence measures, which favours expanding the interval of feasible values of numerical variables. Consequently, the expanding caused that the number of mined association rules was increased, and the value of the evaluation function was raised indirectly. On the other hand, the number of categorical attributes was decreased. As a result, a new evaluation function is proposed, as follows in our study:

$$f(\mathbf{x}_i^{(t)}) = \frac{\alpha \cdot supp(\mathbf{x}_i^{(t)}) + \beta \cdot conf(\mathbf{x}_i^{(t)}) + \gamma \cdot inclusion(\mathbf{x}_i^{(t)}) + \delta \cdot K}{\alpha + \beta + \gamma + \delta}, \tag{12}$$

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

$$inclusion(X \Rightarrow Y) = \frac{|Ante(X \Rightarrow Y)| + |Cons(X \Rightarrow Y)|}{m}, \tag{13}$$

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. (12) are set to $\alpha = \beta = \gamma = \delta = 1$ in the study.

Obviously, the task of the NARM-DE algorithm is to maximize the value of the proposed evaluation function.

4 Experiments and Results

The purpose of our experimental work was to show that the improved nature-inspired algorithms for NARM should be applied successfully in practice. In line with this, we focused on the posted issues, such as shrinking the lower and upper bounds of the numerical attributes, and operating of the new evaluation function.

During the experimental work, four different nature-inspired algorithms were employed: Differential Evolution (DE) [6], Particle Swarm Optimization (PSO) [6], Cuckoo Search (CS) [9], and Flower Pollination Algorithm (FPA) [8]. All algorithms used parameter settings as proposed in corresponding literature. In order to make comparative analysis as fair as possible, the number of evaluation function evaluations was fixed as $nFES = 10,000$, while the number of independent runs was set as $nRUN = 5$.

The algorithms solved problems generated by the random sport dataset generator SportyDataGen [1]. The random generator is capable of generating random instances of numerical attributes. The following measures were used for comparing the algorithms: (1) Total number of mined rules, and (2) Average number of antecedents and consequents[1], (3) Average fitness, (4) Average support, (5) Average confidence, (6) Average shrink, and (7) Average inclusion.

The results of NARM are illustrated in Table 2, which presents the mentioned statistical measures for each of the algorithms in the experiments. Let us notice that the best results are depicted in bold case in the Table. As can be seen from Table 2, DE discovered the maximum number of total rules. These rules are of the best average fitness and inclusion. On the other hand, PSO mined rules of the best average support, confidence and shrink. However, CS and FPA achieved the best results according to the average number of antecedents/consequents. In summary, the best results for using in practice were obtained by DE.

Table 2. 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 antecedents/consequents	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

[1] The first number denotes the number of antecedents, while second denotes the number of consequent.

Interestingly, examples of selected solutions that were found by the proposed algorithms are illustrated in Table 3.

Table 3. Examples of solutions found by the proposed algorithms.

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

4.1 Discussion

The results of the mentioned nature-inspired algorithms for NARM showed that selection of the algorithm has a big influence on the quality of the results. Thus, an advantage of the DE algorithm is in the total discovered rules, average fitness and inclusion, while the PSO was better regarding the average support, confidence, and shrink. On the other hand, working with the numerical attributes revealed a lot of issues that need to be considered for the future work. Let us mention only the more important ones:

- How to consider shrinking as the statistical measure? In our results, we considered the shrinking intervals of all attributes, including those that did not arise in the mined rules.
- How to balance the weights of four terms in the proposed evaluation function? In our case, all weights were set to the value of 1.0, which means that all the contributions were weighted equally.
- Is the best mined association rule according to fitness value also the most interesting?
- How to find the balance between shrink and inclusion?

The mentioned issues confirm that the development of the proposed algorithm for pure NARM is far from completion. A lot of researches would be necessary in order to find the proper answers to the mentioned issues.

5 Conclusion

Development of a nature-inspired algorithm for pure NARM demands answers to new issues, such as, for instance: How to shrink the lower and upper bounds of numerical attributes? or How to find the proper evaluation function? The former issue is confronted with the exploration of the search space, while the latter with evaluating the quality of the mined association rules.

This paper proposes usage of the shrinking coefficient, that is determined as a ratio between the difference of the generated upper and lower bounds, and difference of the maximum upper and minimum lower bounds. As an evaluation function, a weighted sum of support, confidence, inclusion, and shrinking coefficient are taken into consideration. However, the weights were set to the same value of 1.0 in our preliminary study. The nature-inspired algorithms for pure NARM were employed to a sample sport dataset generated by the random generator located on the web. Even four nature-inspired algorithms were tested in our comparative study, as follows: DE, PSO, CS, and FPA.

The results of the comparative analysis revealed that, although there are differences between the specific nature-inspired algorithms, these could be applied for solving the problem in practice. On the other hand, a lot of work is necessary in order to find the proper weights for determining the particular contributions of terms in the evaluation function. However, all this work could be a potential direction for the future work.

References

1. Sportydatagen: An online generator of endurance sports activity collections. In: Proceedings of the Central European Conference on Information and Intelligent Systems, Vara ždin, Croatia, 19, 21 September 2018, pp. 171–178 (2018)
2. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proceedings 20th international conference very large data bases, VLDB, vol. 1215, pp. 487–499 (1994)
3. Altay, E.V., Alatas, B.: Performance analysis of multi-objective artificial intelligence optimization algorithms in numerical association rule mining. J. Ambient Intell. Human. Comput. 1–21 (2019)
4. Fister, Jr.I., Fister, I.: A brief overview of swarm intelligence-based algorithms for numerical association rule mining. arXiv preprint arXiv:2010.15524 (2020)
5. Fister Jr., I., Iglesias, A., Galvez, A., Del Ser, J., Osaba, E., Fister, I.: Differential evolution for association rule mining using categorical and numerical attributes. In: Yin, H., Camacho, D., Novais, P., Tallón-Ballesteros, A.J. (eds.) Intelligent Data Engineering and Automated Learning - IDEAL 2018, pp. 79–88. Springer International Publishing, Cham (2018)
6. Storn, R., Price, K.: Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. J. of Global Optimization **11**(4), 341–359 (dec 1997). https://doi.org/10.1023/A:1008202821328
7. Telikani, A., Gandomi, A.H., Shahbahrami, A.: A survey of evolutionary computation for association rule mining. Information Sciences (2020)

8. Yang, X.S.: Flower pollination algorithm for global optimization. In: Durand-Lose, J., Jonoska, N. (eds.) Unconventional Computation and Natural Computation, pp. 240–249. Springer, Heidelberg (2012)
9. Yang, X.S.: Bat algorithm and cuckoo search: a tutorial, pp. 421–434. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-29694-9_17