A novel approach to discover numerical association based on the Coronavirus Optimization Algorithm

C. Segarra-Martín
Department of Engineering
International University of Menéndez Pelayo
ES-28040, Madrid, Spain
100000206@alumnos.uimp.es

M. Martínez-Ballesteros
Department of Computer Science
University of Seville
ES-41012, Seville, Spain
mariamartinez@us.es

A. Troncoso
Data Science and Big Data Lab
Pablo de Olavide University
ES-41013, Seville, Spain
atrolor@upo.es

F. Martínez-Álvarez
Data Science and Big Data Lab
Pablo de Olavide University
ES-41013, Seville, Spain
fmaralv@upo.es

ABSTRACT

The disease caused by the SARS-CoV-2 (COVID-19) has affected millions of people around the world since its detection in 2019. This pandemic inspired the development of the Coronavirus Optimization Algorithm (CVOA), a bio-inspired metaheuristic that was originally used to adjust deep learning models for time series forecasting, by means of a binary codification. In this paper, a integer codification for the CVOA individual is introduced and used for optimizing a novel approach for numerical association rules mining. As an application case, the prediction of earthquakes of large magnitude has been addressed. This kind of events are rare and, therefore, they can be characterized by rules with very high interest or lift and low support. Thus, the algorithm has been applied to the extraction of rules meeting specific criteria in an earthquake data set, provided by the National Geographic Institute of Spain. The results show CVOA as a promising tool for numerical association rules mining, obtaining rules with useful and meaningful information for predicting the occurrence of large earthquakes.

CCS CONCEPTS

• **Information Systems** → *Association Rules*; • **Mathematics of computing** → *Optimization with randomized search heuristics*; • **Applied computing** → *Forecasting*;

KEYWORDS

Bioinspired metaheuristic, numerical association rules, time series, earthquake magnitude prediction.

ACM Reference Format:
C. Segarra-Martín, M. Martínez-Ballesteros, A. Troncoso, and F. Martínez-Álvarez. 2022. A novel approach to discover numerical association based on the Coronavirus Optimization Algorithm. In *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), April 25–29, 2022, Virtual Event*. ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3477314.3507343

SAC '22, April 25–29, 2022, Virtual Event

https://doi.org/10.1145/3477314.3507343

1 INTRODUCTION

Bio-inspired algorithms are metaheuristics that mimic nature to solve optimization problems. In biology, examples of optimization can be found by looking at natural phenomena that manage to find optimal strategies for the high complexity of life, while maintaining diversity and adaptation to the environment [3]. Viruses spread by infecting people, who may infect new individuals, die or recover after passing the disease. The immune system, along with the development of vaccines, serves to fight the virus and reduce its impact.

Metaheuristics must deal with large search spaces, even infinite for continuous cases, and they must find suboptimal solutions in reasonable execution times [4]. The rapid spread of the SARS-CoV-2 virus has inspired the development of the coronavirus optimization algorithm, known as Coronavirus Optimization Algorithm (CVOA), based on the COVID-19 spread model [12].

The CVOA algorithm can be combined with other artificial intelligence techniques, such as association rule mining, which is an important research area in data mining. It is aimed at extracting interesting correlations, frequent patterns or associations between sets of items in transactional databases or other data repositories [6]. Discovering rules for attributes with numerical values is a challenge, since the most popular methods for association rule mining cannot be applied to numerical data without prior discretization, leading to increased computational cost and loss of information [16]. Thus, this paper presents an adapted version of CVOA that uses an integer codification for numerical association rules mining (NARM).

Earthquake magnitude prediction is a problem of utmost relevance. The occurrence of earthquakes of large magnitude is particularly difficult to predict because these events occur rarely in the nature and exhibit no apparent correlation with the past. Such phenomena can be characterized by numerical association rules, as firstly proposed in [14]. Thus, the proposed algorithm has been adapted to find rules with a consequent including earthquakes with

large magnitude, which can be characterized by rules with very high lift and low support. These rules are also used to characterize outliers of temporal data so that these outliers can model any type of natural disaster. Hence, a case study is included in which CVOA is used to obtain association rules in a Spanish earthquake data set. The code, developed in Python, can be found in the Supplementary Material section.

2 RELATED WORKS

Many bio-inspired metaheuristics can be found in the literature. For example, a virus-based optimization algorithm (VOA) was proposed in [7]. It is a population metaheuristic algorithm that mimics the behaviour of viruses attacking a living cell. In contrast to CVOA, it simulates generic viruses and requires a search for suitable parameter values. The method was improved in [8] and the results show that it is a viable solution for continuous optimization.

Several evolutionary algorithms and fuzzy evolutionary algorithms for NARM are presented and compared in [23]. The performance of these metaheuristics has been compared with the classical Apriori algorithm [22]. This algorithm works in two phases, first it extracts the frequent itemsets and then generates the rules from them. A genetic algorithm is proposed in [1] as a strategy to mine association rules directly without the need to generate frequent itemsets previously. A similar approach is presented in [24], where the minimum support of the rules is not necessary. The results showed that the proposed method significantly reduces computational costs and generates interesting association rules.

A problem arises when classical association rule mining methods deals with numerical data since a prior discretization is needed, which involves a higher computational effort and information loss. Techniques based on evolutionary algorithms are proposed in several works [9–11] to find association rules in numerical datasets without a previous discretization of the attributes. Thus, the interval bounds of the attributes are decided in the evaluation process of the solutions at each iteration [17].

The result of the analysis and the comparison of the methods presented in [23] shows that intelligent algorithms seem to be the best alternative for the complex NARM problem. They are more efficient by avoiding pre-processing of data and several metrics computation. It can be concluded that new algorithms can be adapted to obtain better results, which justifies the application of CVOA for association rules.

3 DESCRIPTION OF CVOA FOR INTEGER CODIFICATION

In this section, the proposed individual codification and the optimization process is described, as well as the assigned parameter values. Finally, the implementation of a version with several strains running in parallel is included.

3.1 Individual codification

CVOA is an iterative searching process to find solutions for an optimization problem. In a similar way to other population algorithms, it begins with the initialization of a population of individuals, which is updated in each iteration, by exploring the solution search space,

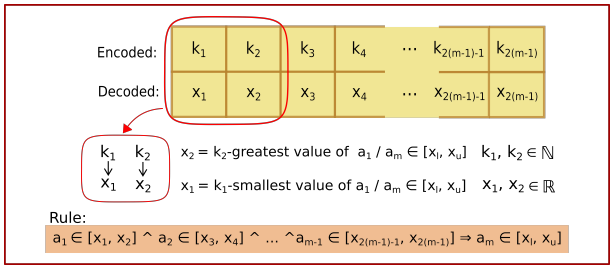

Figure 1: Individual codification for NARM in a data set with the proposed version of CVOA, where $A = \{a_1, \ldots, a_m\}$ is the set of attributes, with values $x \in \mathbb{R}$, and x_l, x_u, the lower and upper interval limits, pre-specified for the attribute a_m, which is the consequent of the rule.

until the stopping criterion is met. Finally, it returns the best solution found. For this particular case, solutions returned by the algorithm represent association rules.

An association rule is an implication in the form $X \Rightarrow Y$, where X is the antecedent and Y the consequent. The rule means *X implies Y* [6]. Finding rules that characterize outliers can be viewed as a special form of association rule mining where conclusions of rules are pre-specified [20]. Given a data set, the attribute whose value is used to consider whether an example is an outlier is set as a consequent.

Extracting rules from a numeric data set presents the following problem. Numeric attributes are usually defined on a wide range of different values. In order to apply data mining techniques, the attribute domains are divided into intervals. This is known as discretization [21]. As this process entails an additional computational cost due to the selection of the suitable intervals, and possible loss of information, in CVOA the amplitude of the intervals is decided in each iteration of the algorithm.

In the following lines the codification used for the individuals candidates to be solutions of CVOA hybridized with NARM will be described. As already mentioned, the consequent of the rule is fixed and pre-specified by the user, passing it as a parameter. The algorithm optimizes the antecedent of the rule, which consists of a number of intervals equal to the number of attributes of the data set minus the attribute set as consequent. Therefore, the CVOA individual is a collection of elements that encode the antecedent of the rule. Each element encodes the lower or upper limit of an interval. Given a data set with m attributes and n instances which meet the condition of the consequent, that is, the value of the attribute is in the specified interval, the individual consists of a list of 2(m-1) integers k, where $k \in [1, n-1]$. Even positions in the list correspond to the intervals lower limit, and their values are integer numbers k_p in the range $[1, n-1]$. Their corresponding decoded values are the (k_p)-smallest values for each attribute of the n instances. That is, if $k_p = 1$, the lower limit of the interval is the minimum value of the corresponding attribute. Elements in odd positions encode the intervals upper limit, and they can take values k_i between 1 and $n - k_p$. Their decoded values are the (k_i)-greatest values for each attribute of the n instances. In this case, $k_i = 1$ means that the upper limit of the interval is the maximum value. In Figure 1 this codification is shown graphically.

3.1.1 Fitness function selection. In this work, a combination of CVOA with NARM is proposed. Therefore, the objective is to optimize the interest of the mined rules. The measure used to evaluate the rules is lift. Other standard measures are support, to find frequent patterns, and confidence, to assess the quality of the rules. However, confidence does not capture the correlation that exists between the antecedent and consequent of the rule. Moreover, the interestingness measure lift captures that correlation in the sense that it tells us whether the antecedent influences the consequent positively (lift >1) or negatively (lift <1). Therefore, using lift instead of confidence as a criterion for discovering association rules may be more effective. In addition, it is intended that the rules obtained with CVOA characterize outliers in data sets, which are defined by rules with high lift and very low support. For this particular case, such rules identify earthquakes with large magnitude. Please refer to [9] for more information about these measures.

3.1.2 Infection procedure. An individual becomes infected through the mutation of the contagious individual. Individuals are lists that encode a certain number of intervals, so this mutation consists of the infection of these intervals. To determine how many intervals are affected, the trip distance (*TRAVEL_DISTANCE*) is used. This distance must be a number between 1 and the total number of intervals, that is, half the length of the individual. In this way, an interval is chosen at random to be infected, until having modified as many as *TRAVEL_DISTANCE* indicated.

3.1.3 Interval infection. When infecting a single interval, its amplitude (a_i) is decreased or increased in a certain percentage, resulting in a new amplitude (a_f). The percentages 25%, 50% and 75% are considered, and one of them is randomly chosen as follows. A variable $d = |a_f - a_i|$ is defined as the difference between the new and the initial amplitudes. A number $P \in \{0, 1, 2\}$ is generated such that $P = 0 \Rightarrow d = 0.25 \times a_i$; $P = 1 \Rightarrow d = 0.5 \times a_i$ and $P = 2 \Rightarrow d = 0.75 \times a_i$. Since the individuals are actually a codification of the intervals, and consist of lists of integers, the difference must also be an integer. For this reason, the difference d is rounded. The decision whether the amplitude of the interval is increased or decreased is made by randomly generating a number between 0 (the interval is increased) and 1 (the interval is decreased). To increase the interval, the amount $round(d)/2$, rounded down, is subtracted from each element that encodes one extreme of that interval, the number k that is mentioned in the section that explains the encoding. To reduce the interval, the amount $round(d)/2$, rounded up, is subtracted. If the value of k resulting from modifying an extreme of an interval is less than 1, it is assigned 1.

3.2 Parallel strains implementation

The algorithm presented in this work is configured so that a parallelized version can be implemented simulating various strains of SARS-CoV-2. In this way, diversification is increased, that is, it allows covering a larger size of the search space. Please refer to the original CVOA work for further details [12].

3.3 Suggested parameters setup

In this section, the values assigned to CVOA parameters are presented. These values have been selected from different sources such as the World Health Organization (WHO) [19]. It should be noted that each iteration of the algorithm simulates a week. Full description for these parameters can be found in the original CVOA work [12]:

(1) $P\_DIE = 0.06$ has been assigned.
(2) $P\_SUPERSPREADER = 0.1$.
(3) $SUPERSPREADER\_RATE \in [6, 15]$
(4) $P\_REINFECTION = 0.001$.
(5) $P\_ISOLATION = 0.7$, since this value ensures that $R_0 \leq 1$, as shown in the original work
(6) $P\_TRAVEL = 0.1$.
(7) $SOCIAL\_DISTANCING \in [7, 12]$.
(8) $PANDEMIC\_DURATION = 20$.

4 RESULTS: APPLICATION TO EARTHQUAKE MAGNITUDE PREDICTION

The methodology explained in this work has been applied to optimize the interest of rules generated with earthquake data. In this section it is presented, firstly, a description of the data set used and, thereafter, the results obtained and their analysis.

The data set used in this work has been retrieved from the catalogue of Spanish's Geographical Institute, which contains the location and magnitude of Spanish earthquakes. It is the dataset used in [13], however, in this paper we only focus on the prediction of earthquakes of moderate-large magnitude, which is considered a major challenge in the literature. Thus, the objective is to find patterns that precede moderate-large earthquakes, therefore M_a will be set as the consequence of the rules (magnitude of an earthquake occurred after the conditions determined in the antecedent). For this reason, it has been decided to extract rules for $M_a \in [4.4, 6.2]$ and $M_a \in [5.0, 6.0]$. Note that Δt and Δb represent the time elapsed, in years, between the previous and current earthquake and the increment of the b-value [18], while M_p identifies the magnitude of the previous earthquake (the last one reported).

The algorithm has been configured with ten coronavirus strains running in parallel. All of them have the same parameter values, described in Section 3.3. Table 1 shows the five best rules found for $M_a \in [4.4, 6.2]$, after five executions. Regarding the metrics, the confidence is 100%, the lift measure takes a value of 16.47 and the support 1.1×10^{-3}, for all the five rules. According to what has been explained about this metric, a value greater than 1 leads to consider that the rules are interesting and that the antecedent and consequent attributes are positively correlated. Regarding the support of the rules, it varies between 0.11% and 0.34%, low values, but expected since we are dealing with infrequent magnitudes, which represent a minority in the data set. Specifically, earthquakes that satisfy $M_a \in [4.4, 6.2]$ account for 6% of the total records in the data set.

Table 2 show the results for the consequent $M_a \in [5.0, 6.0]$, the same way as above: for ten strains running in parallel. Confidence is also 100% for all rules. Support for all the five rules is 0.11%. But it should be noted that in this case, the percentage of records that satisfy $M_a \in [5.0, 6.0]$ is reduced to 1.6%. The lift measure is equal to 62.357 for all rules, which also indicates a positive correlation.

The total of rules generated for both ranges of M_a have a negative Δb, that is, the b-value decreases in all of them. The works in

Table 1: Antecedent of the five best rules found for the consequent $M_a \in$ [4.4, 6.2] using ten strains.

Antecedent
$\Delta t \in [0.019, 0.029] \wedge \Delta b \in [-0.020, -0.016] \wedge M_p \in [3.6, 3.8]$
$\Delta t \in [0.018, 0.082] \wedge \Delta b \in [-0.055, -0.026] \wedge M_p \in [4.1, 4.4]$
$\Delta t \in [0.029, 0.033] \wedge \Delta b \in [-0.026, -0.016] \wedge M_p \in [3.6, 3.7]$
$\Delta t \in [0.019, 0.082] \wedge \Delta b \in [-0.019, -0.016] \wedge M_p \in [3.6, 4.0]$
$\Delta t \in [0.011, 0.331] \wedge \Delta b \in [-0.039, -0.024] \wedge M_p \in [4.0, 5.3]$

Table 2: Antecedent of the five best rules found for the consequent $M_a \in$ [5.0, 6.0] using ten strains.

Antecedent
$\Delta t \in [0.130, 0.137] \wedge \Delta b \in [-0.019, -0.013] \wedge M_p \in [3.1, 5.2]$
$\Delta t \in [0.001, 0.137] \wedge \Delta b \in [-0.040, -0.036] \wedge M_p \in [3.9, 5.2]$
$\Delta t \in [0.012, 0.080] \wedge \Delta b \in [-0.045, -0.039] \wedge M_p \in [3.8, 3.9]$
$\Delta t \in [0.048, 0.080] \wedge \Delta b \in [-0.045, -0.026] \wedge M_p \in [3.9, 5.2]$
$\Delta t \in [0.080, 0.137] \wedge \Delta b \in [-0.101, -0.017] \wedge M_p \in [4.1, 5.0]$

[2, 5, 15] draw similar conclusions than those obtained by the rules with reference to variations of the b-value.

5 CONCLUSIONS

A new algorithm based on CVOA to mine numerical association rules has been proposed in this work. An integer codification has been described to represent these rules. The results obtained show that CVOA is a good alternative to the classic association rule mining methods, due to its efficiency in optimizing the objective function, without requiring previous steps such as the extraction of frequent patterns or the numerical attributes discretization. The proposed codification allows working with data defined in $\mathbb{R}$, creating and modifying intervals in each iteration. Another advantage of this method is that it does not require a prior discretization of the data, so there is no loss of information derived from this process. The algorithm has been validated by characterizing moderate-large earthquakes in Spain and show that CVOA is a good technique to find interesting rules, managing to optimize the lift and allowing to discover the influence of the b-value and the time elapsed in the occurrence of moderate-large earthquakes.

ACKNOWLEDGEMENTS

The authors would like to thank the Spanish Ministry of Economy and Competitiveness for the support under the project TIN2017-88209-C2-1-R, and to the Junta de Andalucía for the support under projects PY20-00870 and UPO-1380516.

SUPPLEMENTARY MATERIAL

The Python implementation of CVOA with integer codification, to optimize the interest of association rules, can be found in the GitHub repository (https://github.com/cristinasegarra7/CVOA_rules).

REFERENCES

[1] B. Alataş and E. Akin. 2006. An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules. *Soft Computing* 10 (2006), 230–237.
[2] E. Bayrak and Y. Bayrak. 2011. Temporal and Spatial Variations of b-value in the Western Anatolia. In *Proceedings of the Congress of Balkan Geophysical Society*. 1–6.
[3] S. Binitha and S. S. Sathya. 2012. A Survey of Bio inspired Optimization Algorithms. *International Journal of Soft Computing and Engineering* 2 (2012), 137–151.
[4] I. Boussaïd, J. Lepagnot, and P. Siarry. 2013. A survey on optimization metaheuristics. *Information Sciences* 237 (2013), 82–117.
[5] J. Chen and S. Zhu. 2020. Spatial and temporal b-value precursors preceding the 2008 Wenchuan, China, earthquake (Mw = 7.9): implications for earthquake prediction. *Geomatics, Natural Hazards and Risk* 11 (2020), 1196–1211.
[6] S. Kotsiantis and D. Kanellopoulos. 2006. Association Rules Mining: A Recent Overview. *GESTS International Transactions on Computer Science and Engineering* 32 (2006), 71–82.
[7] Y.-C. Liang and J. R. Cuevas-Juárez. 2016. A novel metaheuristic for continuous optimization problems: Virus optimization algorithm. *Engineering Optimization* 48 (2016), 73–93.
[8] Y.-C. Liang and J. R. Cuevas Juárez. 2020. A self-adaptive virus optimization algorithm for continuous optimization problems. *Soft Computing* 24 (2020), 13147–13166.
[9] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, and J. C. Riquelme. 2011. An evolutionary algorithm to discover quantitative association rules in multidimensional time series. *Soft Computing* 15, 10 (2011), 2065.
[10] M. Martínez-Ballesteros, A. Troncoso, F. Martínez-Álvarez, and J. C. Riquelme. 2016. Improving a multi-objective evolutionary algorithm to discover quantitative association rules. *Knowledge and Information Systems* 49, 2 (2016), 481–509.
[11] D. Martín, M. Martínez-Ballesteros, D. García-Gil, J. Alcalá-Fdez, F. Herrera, and J.C. Riquelme-Santos. 2018. MRQAR: A generic MapReduce framework to discover quantitative association rules in big data problems. *Knowledge-Based Systems* 153 (2018), 176–192.
[12] F. Martínez-Álvarez, G. Asencio-Cortés, J. F. Torres, D. Gutiérrez-Avilés, L. Melgar-García, R. Pérez-Chacón, C. Rubio-Escudero, J. C. Riquelme, and A. Troncoso. 2020. Coronavirus Optimization Algorithm: A Bioinspired Metaheuristic Based on the COVID-19 Propagation Model. *Big Data* 8, 4 (2020), 308–322.
[13] F. Martínez-Álvarez, A. Morales-Esteban, A. Troncoso, J. L. de Justo, and C. Rubio-Escudero. 2010. Pattern recognition to forecast seismic time series. *Expert Systems with Applications* 37, 12 (2010), 8333–8342.
[14] F. Martínez-Álvarez, A. Morales-Esteban, A. Troncoso, and J. C. Riquelme. 2011. Computational intelligence techniques for predicting earthquakes. In *Proceedings of the International Conference on Hybrid Artificial Intelligence Systems*. 287–294.
[15] F. Martínez-Álvarez, J. Reyes, A. Morales-Esteban, and C. Rubio-Escudero. 2013. Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula. *Knowledge-Based Systems* 50 (2013), 198–210.
[16] B. Minaei-Bidgoli, R. Barmaki, and M. Nasiri. 2013. Mining numerical association rules via multi-objective genetic algorithms. *Information Sciences* 233 (2013), 15–24.
[17] F. Moleshi, A. Haeri, and F. Martínez-Álvarez. 2020. A novel hybrid GA–PSO framework for mining quantitative association rules. *Soft Computing* 24 (2020), 4645–4666.
[18] A. Morales-Esteban, F. Martínez-Álvarez, A. Troncoso, J. L. Justo, and C. Rubio-Escudero. 2010. Pattern Recognition to Forecast Seismic Time Series. *Expert System with Applications* 37 (2010), 8333–8342.
[19] World Health Organization. 2020. Coronavirus disease 2019 (COVID-19): Situation report 74. Technical report, WHO. https://www.who.int/docs/default-source/coronaviruse/situationreports/20200403-sitrep-74-covid-19-mp.pdf
[20] C. Robardet, B. Crémillleux, and J. F. Boulicaut. 2002. Characterization of unsupervised clusters with the simplest association rules: application for child's meningitis. In *Proceedings of the European Conference on Artificial Intelligence*, P. Lucas, L. Asker, and S. Miksch (Eds.). 61–65.
[21] A. Salleb-Aouissi, C. Vrain, and C. Nortet. 2007. Quantminer: A genetic algorithm for mining quantitative association rules. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 1035–1040.
[22] R. Srikant and R. Agrawal. 1996. Mining quantitative association rules in large relational tables. In *Proceedings of the International Conference on Management of Data*. 1–12.
[23] E. Varol-Altay and B. Alatas. 2020. Intelligent optimization algorithms for the problem of mining numerical association rules. *Physica A: Statistical Mechanics and its Applications* 540 (2020), 123142.
[24] X. Yan, C. Zhang, and S. Zhang. 2009. Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. *Expert Systems with Applications* 36 (2009), 3066–3076.