International Journal of Data Warehousing and Mining
Volume 18 • Issue 1

Association Rule Mining Based on Hybrid Whale Optimization Algorithm

Zhiwei Ye, Hubei University of Technology, China & Fujian Provincial Key Laboratory of Data Intensive Computing, China & Key Laboratory of Intelligent Computing and Information Processing, China*

Wenhui Cai, Hubei University of Technology, China

Mingwei Wang, Hubei University of Technology, China

Aixin Zhang, Hubei University of Technology, China

Wen Zhou, Hubei University of Technology, China

Na Deng, Hubei University of Technology, China

Zimei Wei, Hubei University of Technology, China

Daxin Zhu, Quanzhou Normal University, China & Fujian Provincial Key Laboratory of Data Intensive Computing, China & Key Laboratory of Intelligent Computing and Information Processing, China

ABSTRACT

Association rule mining (ARM) is one of the most significant and active research areas in data mining. Recently, whale optimization algorithm (WOA) has been successfully applied in the field of data mining; however, it easily falls into the local optimum. Therefore, an improved WOA-based adaptive parameter strategy and Levy flight mechanism (LWOA) is applied to mine association rules. Meanwhile, a hybrid strategy that blends two algorithms to balance the exploration and exploitation phases is put forward, that is, grey wolf optimization algorithm (GWO), artificial bee colony algorithm (ABC), and cuckoo search algorithm (CS) are devoted to improving the convergence of LWOA. The approach performs a global search and finds the association rules sets by modeling the rule mining task as a multi-objective problem that simultaneously meets support, confidence, lift, and certain factor, which is examined on multiple data sets. Experimental results verify that the proposed method has better mining performance compared to other algorithms involved in the paper.

KEYWORDS

Association Rule Mining, Data Mining, Hybrid Strategy, Levy Flight, Whale Optimization Algorithm

INTRODUCTION

Data mining is a general and persuasive technique for extracting valuable knowledge from data sources (Telikani et al., 2020). Association Rule Mining (ARM) is one of the most ordinary and crucial tasks in data mining, which can find the association between the data by mining the frequent itemsets (Baró et al., 2020). ARM has high practical value in real life because it contributes to summarizing laws from big data, which has been extensively applied in the fields of healthcare, recommender systems, market analysis, and transportation (Das et al., 2021). For example, Anand Hareendran and Vinod Chandra (2017) extracted the features and their relationships from liver transplantation data using ARM and designed high precise mining model. Viktoratos et al. (2018) proposed a novel approach by

DOI: 10.4018/IJDWM.308817 *Corresponding Author

combining community-based knowledge with ARM to relieve the cold start problem in recommender systems. Wen et al. (2019) designed a hybrid temporal ARM to predict traffic congestion, and Hong et al. (2020) employed ARM to discover the contributory crash-risk factors of vehicle-involved crashes in the transportation field. The primary purpose of ARM is to find rules that satisfy the predefined minimum support and confidence levels from a given database. Apriori algorithm and FP-growth algorithm are the most typical ARM algorithms. The principle of these algorithms requires scanning the entire data, selecting frequent itemsets by meeting the minimum support threshold, and then obtaining association rules based on the minimum confidence level (Chiclana et al., 2018). Traditional algorithms have low efficiency and high memory overhead for data processing in massive data sets. One of the most effective approaches is combining the optimizer technique with ARM, and previous studies showed that the meta-heuristic method based on population was an effective way to solve the optimization problem (Shu et al., 2022).

Therefore, many swarm-inspired algorithms have been proposed for mining a good subset of association rules. For example, Sarath and Ravi (2013) proposed a binary particle swarm optimization (BPSO) algorithm to generate association rules by constructing a combinatorial global optimizer problem. Kuo et al. (2019) used Pareto-based PSO to optimize the goals, including comprehensibility, confidence, and surprisingness to discover valuable and interesting association rules from numerical valued data sets. Jyoti and Sharma (2020) focused on applying rule mining techniques based on ant colony optimization (ACO) in data classification. Mlakar et al. (2017) designed a modified single-objective binary cuckoo search (MBCS-ARM) that included novel representations of individuals, which helped deal with high dimension problems with an increasing number of attributes. Heraguemi et al. (2015) proposed a bat-based algorithm for ARM (BAT-ARM) and then designed a novel multi-objective bat algorithm for ARM (MOB-ARM) and applied it (Heraguemi et al., 2018). However, most relevant algorithms for controlling association rules are often computationally expensive and possibly generate many irrelevant rules (Kumari et al., 2019).

Because of good optimization performance, Whale Optimization Algorithm (WOA) has been widely employed in data mining. For instance, Kotary et al. (2021) proposed many-objective WOA to handle robust distributed clustering in a wireless sensor network. Too et al. (2021) introduced a new variant of WOA (SBWOA) based on spatial bounding strategy to play the role of finding the potential features from the high-dimensional feature space. Sheikhi (2021) proposed a fake news detection system based on content features and the WOA-Xgbtree algorithm, which had achieved good results in news classification. Wu et al. (2019) proposed a memetic fuzzy whale optimization (MFWO) algorithm to solve the data clustering problem. Samantaray and Sahoo (2021) designed a hybrid model combining a support vector machine and whale optimization algorithm (SVM-WOA), which more effectively predicted suspended sediment concentration than the SVM-PSO model. Yin (2020) et al. proposed an improved WOA based on the chaos theory and logistic mapping technique for brain tumor classification. As for ARM, Sharmila and Vijayarani (2021) used fuzzy logic and WOA for frequent item identification and association rule generation. Heraguemi et al. (2021) proposed an updated WOA for mining association rules (WO-ARM), improving running time and memory usage. ÇELİK (2019) proposed a rule discovery tool for classification using WOA and proved that WOA was an appropriate candidate for classification processes. The above research showed that WOA could effectively mine association rules because of its fast convergence, strong searching ability, simple structure, and easy implementation. However, WOA was limited by the update strategy of population individual position, which was susceptible to fall into the local optimal solution.

To obtain better performance, an improved WOA with adaptive parameter strategy and Levy Flight (LWOA) is introduced to obtain better solutions in the whole search process. In addition, a new hybrid strategy based on LWOA is designed for optimizer and solution search problems, whose aims are to provide promising candidate solutions for LWOA. The basic idea is to guide the position update of LWOA by comparing the optimal solutions between the two algorithms. However, taking a befitting algorithm to hybridize with LWOA is significant for further enhancing the search ability of whales

in this study. Artificial Bee Colony Algorithm (ABC) had a good balance between intensification and diversity (Akay et al., 2021). Danish et al. (2019) proposed a global ABC (GABCS) for data clustering. Sharma et al. (2015) used ABC to generate high-quality association rules for searching frequent itemsets from large data sets. Cuckoo Search Algorithm (CS) had superior performance in exploring solution space to find the global optimal solution (Cuong-Le et al., 2021). Mohammed and Duaimi (2018) proposed an improved discrete CS for ARM (DCS-ARM). A method to enhance the performance of the multiclass support vector machine (MSVM) classifier using the modified cuckoo search (MCS) was proposed in Mehedi et al., (2021). Abdulgader and Kaur (2019) used Grey Wolf Optimization algorithm (GWO) to evolve Mamdani fuzzy rules and proved that the GWO was better than PSO in time and classification accuracy. For these reasons, LWOA is hybridized with the above three algorithms (ABC, CS, and GWO) respectively in the paper. The main contributions of this paper are concluded as follows:

- A parameter adjustment strategy is adopted to help whales explore the search space adaptively, which can improve the exploratory behavior in the early search stage and exploitation ability in the later search stage.
- Levy Flight is embedded in the update process, which can optimize whale diversification and avoid local minima.
- A hybrid strategy is presented to balance the exploration and exploitation phases. ABC, CS and GWO are hybridized with LWOA respectively to improve the optimization capability.
- A multi-objective function is designed for dealing with the ARM problem taking into account different metrics such as support, confidence, lift, and certain factor.

The rest of the paper is organized as follows: The second section briefly gives a fundamental overview of the association rules and WOA. The related technologies of LWOA are introduced in the third section. The fourth section explains the specific workflow of mining association rules using a hybrid strategy based on LWOA. The next section presents the experimental part and comparative studies. Finally, conclusions and future work are drawn in the sixth section.

BACKGROUND

Association Rule Mining

Let $I = \left\{I_1, I_2, ..., I_k\right\}$ be an itemset. Let D be a set of database transactions where each transaction T is a nonempty itemset such that $T \subseteq I$. An association rule is an expression of the form $A \rightarrow B$, where $A, B \subseteq I, A \cap B = \varnothing$. A is called the antecedent and B is called the consequent of the rule.

Support: The support of association rule $A \rightarrow B$ is calculated by dividing the occurrence frequency $P\left(A \cup B\right)$ by the total amount of transactions $\left|D\right|$ in the transactional database, as in Equation (1).

$$Sup\left(A \rightarrow B\right) = \frac{P\left(A \cup B\right)}{\left|D\right|} \tag{1}$$

Confidence: $Conf\left(A \rightarrow B\right)$ is the proportion of transactions in D containing A that also contain B, and the rule confidence is defined in Equation (2).

$$Conf\left(A \to B\right) = \frac{P\left(A \cup B\right)}{P\left(A\right)} \tag{2}$$

Lift: The lift metric represents the ratio between the rule's confidence value and the rule's expected confidence value. The lift metric is expressed as Equation (3).

$$Lift\left(A \to B\right) = \frac{Conf\left(A \to B\right)}{Sup\left(B\right)} \tag{3}$$

If $Lift\left(A \to B\right) < 1$, then the occurrence of A is negatively correlated with the occurrence of B. If $Lift\left(A \to B\right) > 1$, then A and B are positively correlated. If $Lift\left(A \to B\right) = 1$, then A and B are independent and there is no correlation between them.

Certain factor (CF): CF is used to evaluate the probability of consequent occurrence when the antecedent has occurred in association rules. In this metric, which can take values in the range of (-1,1), values higher than 0 represent positive correlation, values less than 0 represent negative correlation, and 0 represents independence. The CF metric is determined as follows by Equation (4).

$$CF = \begin{cases} \dfrac{Conf\left(A \to B\right) - Sup\left(B\right)}{1 - Sup\left(B\right)}, Conf\left(A \to B\right) > Sup\left(B\right) \\ \dfrac{Conf\left(A \to B\right) - Sup\left(B\right)}{Sup\left(B\right)}, Conf\left(A \to B\right) < Sup\left(B\right) \\ 0, Conf\left(A \to B\right) = Sup\left(B\right) \end{cases} \tag{4}$$

THE OVERVIEW OF WHALE OPTIMIZATION ALGORITHM

WOA is proposed according to the behavior of whale hunting prey. The algorithm consists of three stages: encircling the prey, shrinking the bubble net, and searching the prey (Mirjalili & Lewis, 2016).

Encircling the Prey

Humpback whales can recognize the position of the prey, regard it as the best candidate solution and surround it. The mathematical model for this behavior is shown in Equations (5-6).

$$\vec{X}^{t+1} = \vec{X}^{t}_{best} - \vec{A} \cdot \vec{D} \tag{5}$$

$$\vec{D} = \left| \vec{C} \cdot \vec{X}^{t}_{best} - \vec{X}^{t} \right| \tag{6}$$

Where $\vec{X}^{t}_{best}$ indicates the optimal solution in generation t. $\vec{D}$ represents the distance vector from the search agent to the prey. $\vec{A}$ and $\vec{C}$ are coefficient vectors, which are calculated according to Equations (7-8) respectively.

$$\vec{A} = 2 \cdot \vec{a} \cdot \vec{r} - \vec{a} \tag{7}$$

$$\vec{C} = 2 \cdot \vec{r} \tag{8}$$

Where $\vec{r}$ is a random vector between 0 and 1. In the iteration process, the update equation of parameter $\vec{a}$ is defined in Equation (9).

$$\vec{a} = 2 - 2 \cdot \frac{\vec{t}}{T} \tag{9}$$

Where $\vec{t}$ represents the current number of iteration, and T is the maximum number of iteration, $\vec{a}$ is decreased from 2 to 0 linearly.

Bubble Mesh Shrinkage

The humpback whales move towards the target prey in a spiral uprising. This stage is expressed as Equations (10-11).

$$\vec{X}^{t+1} = \vec{D} \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X}^{t}_{best} \tag{10}$$

$$\vec{D} = \left| \vec{X}^{t}_{best} - \vec{X}^{t} \right| \tag{11}$$

Where b is a constant and $l \in \left[0,1\right]$. In WOA, the possibility of a whale updating its position in a shrink encircling and spiral uprising is 0.5, respectively. The mathematical model is shown in Equation (12).

$$\vec{X}^{t+1} = \begin{cases} \vec{X}^{t}_{p} - \vec{A} \cdot \vec{D} & p < 0.5 \\ \vec{D} \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X}^{t}_{p} & p \geq 0.5 \end{cases} \tag{12}$$

Where p is a random number between 0 and 1.

Searching Prey

In this stage, the humpback whales update their position to a random whale. The mathematical model of this process is described as Equation (13).

$$\vec{X}^{t+1} = \vec{X}^{t}_{rand} - \vec{A} \cdot \left| \vec{C} \cdot \vec{X}^{t}_{rand} - \vec{X}^{t} \right| \tag{13}$$

Where $\vec{A}$ is a random number between $\left[-\vec{a}, \vec{a}\right]$. The whale approaches the best solution when $\left|\vec{A}\right| < 1$. A whale is randomly selected to update the location of the search agents when $\left|\vec{A}\right| \geq 1$.

In summary, the pseudo-code of the WOA is described in Table 1.

Table 1. Pseudo-Code of WOA

Algorithm 1. Pseudo-Code of WOA
1: Input population and maximum number of iterations
2: Initialize the parameters(a , A , C , l and p) of WOA
3: Initialize the whales' population randomly
4: Calculate the fitness of each whale
5: Choose the best whale to be X_{best}
6: While(t <maximum number of iterations)
7: for each whale
8: if(p <0.5)
9: if($\left\|A\right\|$ <1)
10: Update the position of the current whale by the Equation (5)
11: else if($\left\|A\right\|$ ³1)
12: Select a random whale(X^t_{rand})
13: Update the position of the current whale by the Equation (13)
14: End if
15: else if(p ³0.5)
16: Update the position of the current whale by the Equation (10)
17: End if
18: End for
19: Calculate the fitness of each whale
20: Update X_{best} if there is a better solution
21: Update the parameters(a , A , C , l and p) of WOA
22: $t = t + 1$
23: End while
24: Return X_{best}

THE IMPROVED WHALE OPTIMIZATION ALGORITHM

For WOA, the coefficient vector $\vec{A}$ determines whether the search is global or local, and the value of vector $\vec{A}$ is closely related to the convergence factor a . This section defines the adaptive parameter strategy in Equation (14). The design can expand population diversity in the initial stage and later accelerate the convergence factor a .

$$a = 2 \cdot (\frac{1}{\varepsilon})^{\frac{t}{T}} \cdot e^{-w \cdot \frac{T-t}{T}} \tag{14}$$

Where t is the current iteration number, T is the maximum iteration number, ε and w are adjustment coefficients. It takes $\varepsilon = 5$ and $w = 0.01$ in the paper.

Levy flight mechanism can be utilized to enhance the global search ability of WOA. Therefore, Levy flight mechanism is introduced into the WOA based on nonlinear strategy. The position update Equation (13) of the LWOA is replaced by Equation (15).

$$\vec{X}^{t+1} = \vec{X}^{t}_{rand} + r \oplus Levy\left(s\right) \tag{15}$$

Where r defines a random number between 0 and 1, $\oplus$ represents entry wise multiplication. The Levy Flight provides the distribution a random walk as in Equation (16):

$$Levy\left(s\right) \sim \left|s\right|^{-1-\beta}, 0 < \beta \leq 2 \tag{16}$$

Where s is the step length of Levy flight, and it can be calculated by Mantega's algorithm as in Equation (17).

$$s = \frac{\mu}{\left|v\right|^{1/\beta}}, \mu \sim N\left(0, \sigma_{\mu}^{2}\right), v \sim N\left(0, \sigma_{v}^{2}\right) \tag{17}$$

$$\sigma_{\mu} = \left\{\frac{\Gamma\left(1+\beta\right)\cdot \sin(\pi\beta \,/\, 2)}{\Gamma\left[\left(1+\beta\right)/2\right]\cdot \beta \cdot 2^{(\beta-1)/2}}\right\}^{\frac{1}{\beta}}, \sigma_{v} = 1 \tag{18}$$

Where $\beta = 1.5$, μ and v are both normal distributions. Γ is the standard Gamma function. In brief, the pseudo-code of the LWOA is described in Table 2.

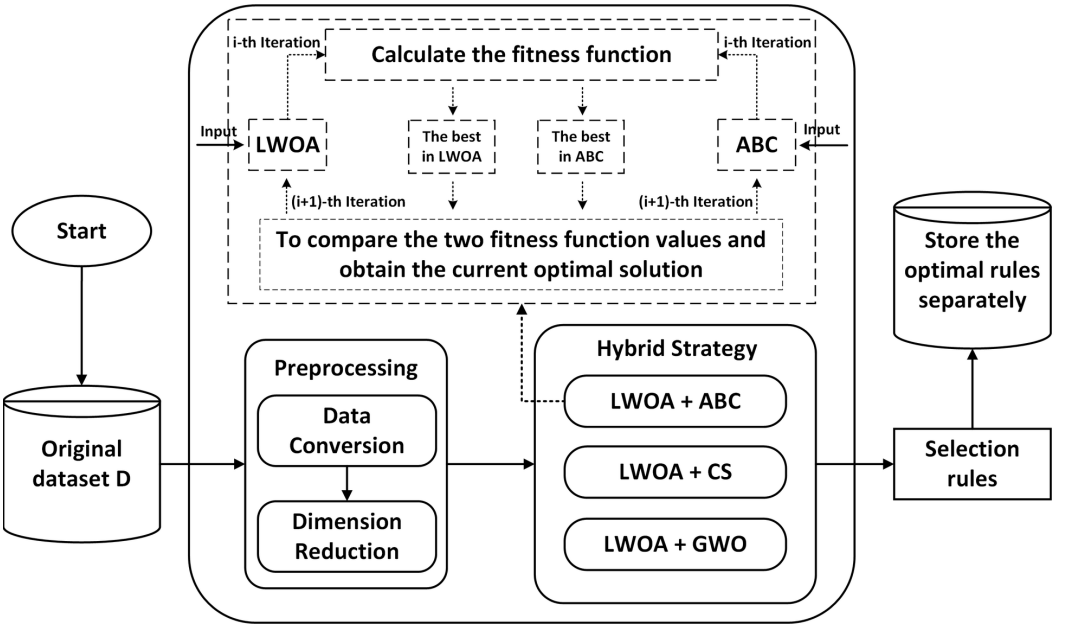
Table 2. Pseudo-Code of LWOA

Algorithm 2. Pseudo-Code of LWOA
1: Input population and maximum number of iterations **2: Initialize** the parameters(a , A , C , l , p , ε and w) of LWOA **3: Initialize** the whales' population randomly **4:** Calculate the fitness of each whale **5:** Choose the best whale to be X_{best} **6: While**(t <maximum number of iterations) **7: for each** whale **8: if**(p < 0.5) **9: if**($\left\|A\right\|$ < 1) **10:** Update the position of the current whale by the Equation (5) 11: else if($\left\|A\right\|$ ³ 1) **12:** Select a random whale(X^t_{rand}) **13:** Update the position of the current whale by the Equation (15) **14: End if** 15: else if(p ³ 0.5) **16:** Update the position of the current whale by the Equation (10) **17: End if** **18: End for** **19:** Calculate the fitness of each whale **20:** Update X_{best} if there is a better solution **21:** Update the parameter a by the Equation (14) **22:** Update the parameters(A , C , l and p) of LWOA **23:** $t = t + 1$ **24: End while** **25: Return** X_{best}

THE HYBRID STRATEGY FOR ASSOCIATION RULES MINING

Because of the complex nature of ARM, a powerful method that can explore the entire search space and exploit local areas is required to apply. The hybridization of different search strategies is a solution that can combine the advantages of some algorithms to balance the exploration and exploitation phases. In order to enhance the ability of LWOA to find association rules, ABC, CS, and GWO are selected to improve the convergence of the LWOA. The hybrid algorithms LWOA-ABC, LWOA-CS, and LWOA-GWO are applied to mine association rules. The following subsections focus on the working principle of mining association rules based on three hybrid algorithms, also shown in Figure 1.

Figure 1. Overall model flow chart



PREPROCESSING

This section introduces two stages of data preprocessing: data conversion and dimension reduction. The first stage is to speed up the calculation of both the support and the confidence of an itemset without repeatedly scanning the database, while the purpose of dimension reduction is to simplify the problem and reduce the loss of computing resources.

DATA CONVERSION

Data are transformed to binary values for computation and storage. As shown in Figure 2, it is a small database with five transactions of T1 to T5 and six items of I1 to I6. Each transaction is a nonempty set of items. If this item exists, it is denoted as 1. Otherwise, it is denoted as 0. Consider the example T1; it contains items I1 and I2. Therefore, the corresponding bit string of B1 is 110000.

Figure 2. Binary transformation of discrete data

Original Data

T1	I1	I2		
T2	I1	I3	I4	I5
T3	I1	I3	I4	I6
T4	I1	I2	I3	I4
T5	I1	I2	I3	I6

Transform Binary

Transformed Data

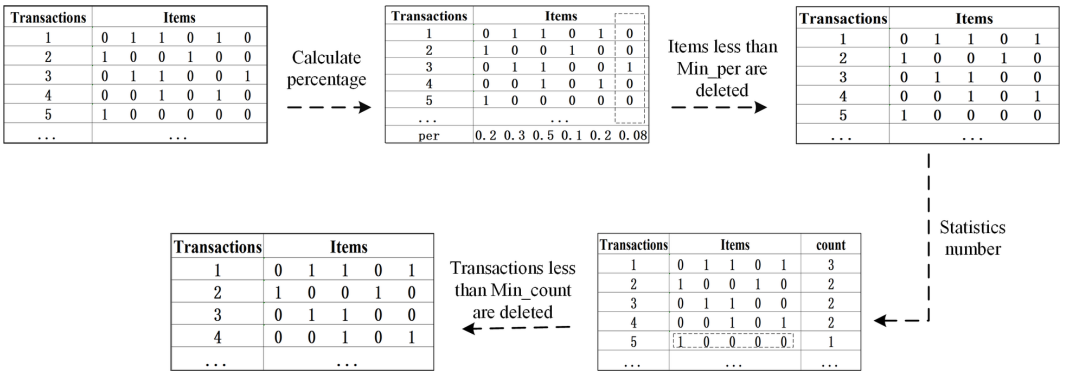
	I1	**I2**	**I3**	**I4**	**I5**	**I6**
B1	1	1	0	0	0	0
B2	1	0	1	1	1	0
B3	1	0	1	1	0	1
B4	1	1	1	1	0	0
B5	1	1	1	0	0	1

DIMENSION REDUCTION

In the era of rapid information development, extracting practical features helps simplify the complexity of the problem in the face of massive data. Therefore, feature selection is crucial. In order to improve the computational efficiency and extract more high-quality rules, the original datasets are dimensionally reduced.

As shown in Figure 3, firstly, by calculating the proportion of each item in all transactions, the items that are less than the specified threshold (that is Min_per) are removed. Then the total number of items for each transaction is counted, and the transactions that are less than the specified threshold (that is Min_count) are also removed, thus obtaining the data with reduced dimensions.

Figure 3. Dimension reduction process



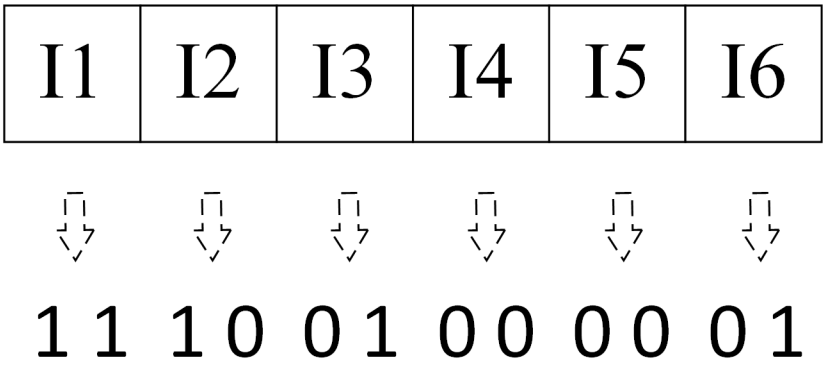
RULE ENCODING

Generally, data can be divided into the continuous type and discrete type. In ARM, continuous data will be transformed into discrete data in a limited continuous space, and the transaction records in the discrete data can be converted into 0 or 1 for storage, which is convenient for subsequent calculations. In the paper, a sigmoid function common in biology that maps continuous values to 0 and 1 is used, as shown in Equation (19).

$$S\left(x\right)=\frac{1}{1+e^{-x}} \tag{19}$$

In the process for ARM, each candidate solution represents an association rule. There are two commonly used coding methods, that is, Pittsburgh Approach and Michigan Approach. In the Pittsburgh coding method, each particle represents a set of rules. After an association rule is generated, the quality of this rule needs to be evaluated. This coding method avoids the repetition of rules but increases the calculation amount. In the Michigan coding method, a particle only represents a rule, and the rule already contains the distinction between the rule before and after, so the calculation is more straightforward and faster, and the coding is easier to operate. On this basis, all algorithms in the paper use Michigan encoding, and a rule representation generated according to the data set in Figure 2 is shown in Figure 4.

Figure 4. The representation scheme of rules



In Figure 4, the code of I1 is 11, indicating that I1 exists in a rule and is the antecedent of the rule. The encoding of I2 is 10, indicating that this item also exists in this rule and is the afterword of the rule. The code of item I3 is 01, indicating that item I3 does not exist in this rule. Therefore, the association rule can be interpreted as I1 → I2.

THE FITNESS FUNCTION

The selection of the fitness functions is crucial to solving the optimization problem. In ARM, confidence and support are commonly used as evaluation metrics. Lift and CF are introduced to verify the reliability of association rules in the paper. By referring to the weighting method in multi-objective, the fitness function proposed in the paper is shown in Equation (20).

$$f\left(x\right) = \frac{Sup\left(A \to B\right) + Conf\left(A \to B\right)}{\min Sup + \min Conf} * Lift\left(A \to B\right) + CF \tag{20}$$

Where $\min Sup$ represents the minimum support threshold, and $\min Conf$ represents the minimum confidence threshold.

THE PROPOSED HYBRID STRATEGY

LWOA with adaptive strategy and levy flight mechanism has a better ability to jump out of the local optimum than standard WOA. However, there is still a risk of falling into the local optima when a single LWOA algorithm is utilized for tackling complex ARM problems. Therefore, a solution combining different strategies to help whales move to promising positions is proposed.

ABC algorithm performs a good balance between diversity and intensification by the foraging process including the employed bees' phase, onlookers' phase and scouts' phase. However, the exploitation ability of ABC mainly depends on the employed bees and onlookers. The employed bee and onlookers stages are realized by changing the single parameter of the food source (old solution), which will make similar solutions gather in a small range. Therefore, ABC has a fast convergence speed, and its combination with LWOA could make up for the slow convergence of LWOA.

CS is an algorithm that primarily utilizes random movement to find promising solutions, and this typical random walk is called levy flight. Levy flight with infinite mean and variance can help

CS effectively explore the search space, which also makes CS have excellent performance in finding the global optimal solution. Comparatively, LWOA may be more difficult to find the global optimal solution than CS, because levy flight is only embedded into the searching prey phase of LWOA. Hence, CS is hybridized with LWOA in the hope that this would help the whale to find the best solution more efficiently.

In GWO, each wolf is guided by the top three wolves in the population to update their position, which pays more attention to exploitation rather than exploration. LWOA uses levy flight to reposition some solutions around randomly selected whales during optimization, which may be easier to escape from local optimal solutions than GWO. Theoretically, the exploration potential in LWOA is slightly higher than that of GWO, while the GWO performs better in exploiting optimal solutions than LWOA.

In summary, ABC, CS, and GWO have their advantages that can guide whales in different aspects. To better combine the advantages of both algorithms, co-evolution is achieved by comparing and sharing the current optimal solution between two algorithms. The hybrid algorithms LWOA-ABC, LWOA-CS, and LWOA-GWO work in the same way. The broad view of these hybrid algorithms is shown in Figure 5, and the implementation steps are as follows:

Step 1: The initial population is divided into two equal subpopulations. The first subpopulation is assigned to Algorithm 1 (LWOA) and another subpopulation is assigned to Algorithm 2 (ABC, CS or GWO).

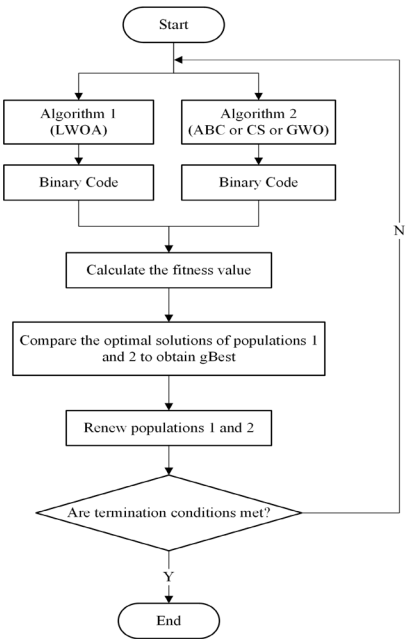
Step 2: All individuals in two subpopulations are encoded into binary values and the corresponding fitness values are calculated.

Step 3: Compare the optimal solutions obtained in algorithms 1 and 2, and select the better solution as the current global optimal solution ($gBest$).

Step 4: Both algorithms use $gBest$ as the best solution and update their populations with their strategies.

Step 5: Repeat steps 3 and 4 until the operating conditions are satisfied and the final solution ($gBest$) is output.

Figure 5. Flow chart of hybrid algorithms



EXPERMENTAL RESULTS AND COMPARISIONS

All the algorithms in the experiment are coded in MyEclipse and run on a Windows platform (i7 processor and 12 GB memory). Each algorithm is run 50 times for a case, and the average results are recorded.

DATASETS

The experimental data are obtained from the "LUCS-KDD Discretised/Normalised" database and are available at "https://cgi.csc.liv.ac.uk/~frans/KDD/Software/LUCS-KDD-DN/DataSets/dataSets.html". The original dataset's rows and columns are reduced using the dimensionality reduction method in the paper, and the final data are shown in Table 3.

Table 3. Datasets descriptions

Dataset	Rows	Columns
Anneal	627	23
Ecoli	313	18
Flare	1208	21
Hepatitis	116	29
Ionosphere	212	41
Led7	3200	14

PARAMETER SELECTION

The Anneal dataset with moderate dimensions is selected to conduct six group of tests on the population sizes and iteration times. Confidence and time are selected as evaluation criteria. These algorithms proposed in the paper were tested with a rule mining algorithm, namely WO-ARM (Heraguemi et al., 2021). The minimum support threshold is set as 0.1, and the minimum confidence threshold is set as 0.4. As shown from the data in Table 4, when the population size is 60 and the iteration time is 1000, most algorithms perform best in confidence. Therefore, to ensure the experiment's fairness, this group of parameters is selected as the parameters of the experiment.

Table 4. Different populations and iteration tests

Populations	Iterations	Evaluation Indicators	WO-ARM	LWOA	LWOA-ABC	LWOA-CS	LWOA-GWO
10	**100**	Conf	0.5428	0.5486	0.6964	0.5149	0.5909
		Time(ms)	27	20	33	39	31
20	**200**	Conf	0.7597	0.7556	**0.832**	0.7602	0.8024
		Time(ms)	88	48	96	129	92
40	**500**	Conf	0.7656	0.7771	0.8147	0.8275	0.7952
		Time(ms)	481	266	646	773	666
60	**1000**	Conf	0.7612	**0.7781**	0.8193	**0.833**	**0.8145**
		Time(ms)	1470	798	1523	2432	1543
80	**1500**	Conf	**0.7817**	0.7703	0.8008	0.8181	0.7995
		Time(ms)	2867	1553	2815	4829	3210
100	**2000**	Conf	0.7653	0.7694	0.8011	0.8129	0.8019
		Time(ms)	4877	2630	4922	8444	5564

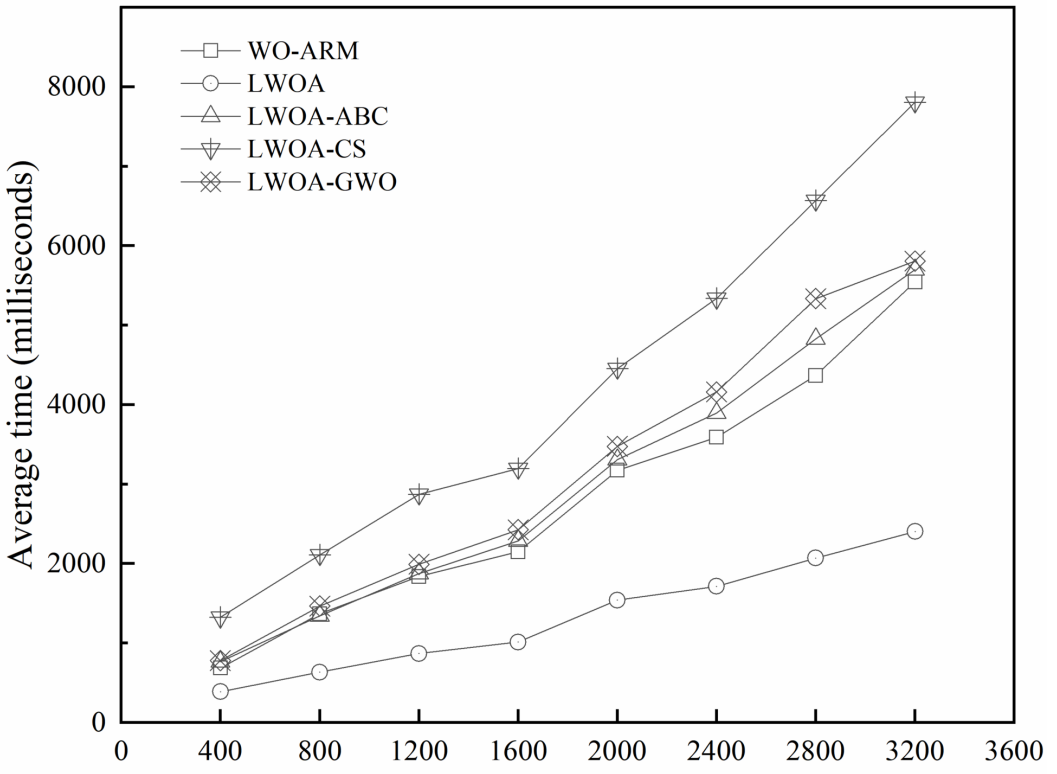
PERFORMANCE COMPARISON

Since the Led7 dataset has the highest number of records, the proposed algorithms are used to perform time statistics under the different numbers of records for this dataset. The results are shown in Table 5 and Figure 6. It could be seen that LWOA takes the least time and LWOA-CS takes the most time. There is little difference in the runtime of LWOA-GWO, LWOA-ABC, and WO-ARM in any number of records. When the number of data records is four times (i.e., the number of records is 1600), the runtime of WO-ARM, LWOA, and LWOA-CS is increased 3.1 times, 2.6 times, and 2.4 times respectively. When the number of data records is eight times, the running time of WO-ARM, LWOA, and LWOA-CS is increased 8.0 times, 6.1 times, and 5.9 times respectively.

Table 5. Running time of each algorithm with different number of data records (ms)

Number of records	WO-ARM	LWOA	LWOA-ABC	LWOA-CS	LWOA-GWO
400	688	390	767	1325	781
800	1370	636	1345	2112	1468
1200	1838	870	1876	2872	1991
1600	2149	1013	2287	3198	2428
2000	3174	1541	3307	4457	3474
2400	3589	1714	3897	5339	4161
2800	4370	2071	4832	6571	5336
3200	5546	2405	5699	7805	5808

Figure 6. Evolution of the running time according to number of records



Figures 7-9 show the average fitness changing process of 1000 iterations of these algorithms on six data sets. It has been analyzed from figures that the convergence speed of WO-ARM is the fastest and tends to be stable after about the 100th iteration, indicating that the algorithm might be difficult to obtain the global optimal solution. As the number of iterations increases, the average fitness of LWOA and the other three hybrid algorithms keeps improving and is significantly higher than that of WO-ARM, proving the effectiveness of these algorithms. LWOA-CS and LWOA-ABC perform well, and it seems that LWOA-CS has a better exploration ability.

Figure 7. The changing process of average fitness on Anneal and Ecoli datasets

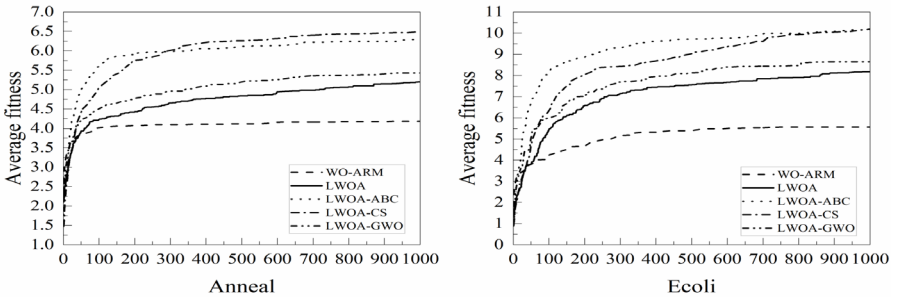


Figure 8. The changing process of average fitness on Flare and Hepatitis datasets

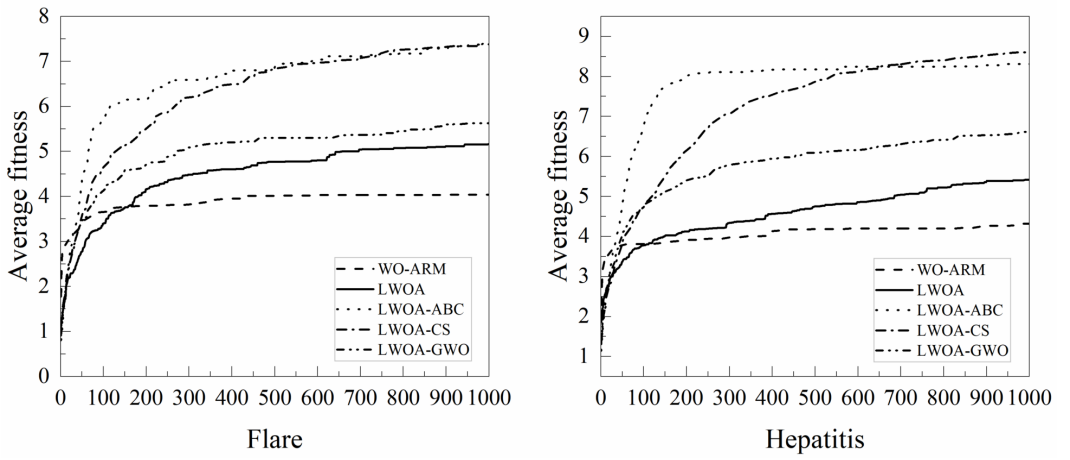
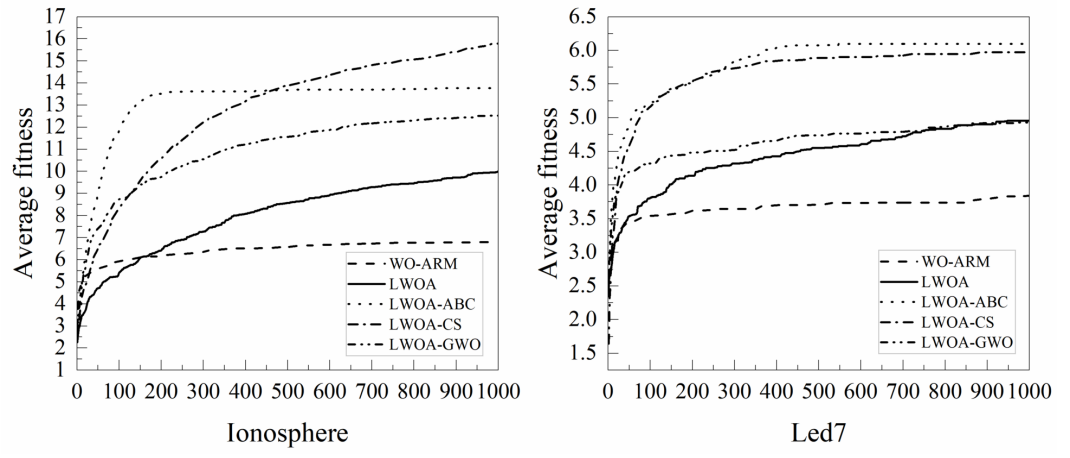


Figure 9. The changing process of average fitness on Ionosphere and Led7 datasets



COMPREHENSIVE EVALUATION

In this section, the average confidence, average support, average lift, average CF, average running time, and the average number of mining rules are tested on six benchmark datasets. Detailed statistics are given in Table 6.

Table 6. Detailed data of each algorithm under different metrics

Dataset	Algorithm	Conf	Sup	Time(ms)	Num	Lift	CF
Anneal	WO-ARM	0.7612	**0.3908**	1470	62	1.164	0.2003
	LWOA	0.7781	0.3776	**798**	131	1.2547	0.3136

Table 6 continued on next page

Table 6 continued

Dataset	Algorithm	Conf	Sup	Time(ms)	Num	Lift	CF
	LWOA-ABC	0.8193	0.2791	1523	1381	1.4252	0.5311
	LWOA-CS	**0.833**	0.2723	2432	**3124**	**1.453**	**0.5913**
	LWOA-GWO	0.8145	0.3141	1543	756	1.3529	0.4777
Ecoli	WO-ARM	0.7937	0.335	638	19	1.7055	0.4075
	LWOA	0.7966	**0.3394**	**416**	61	1.8491	0.4838
	LWOA-ABC	0.8164	0.303	596	376	**2.0826**	0.59
	LWOA-CS	**0.8211**	0.3315	1123	**531**	1.9464	**0.5906**
	LWOA-GWO	0.8051	0.3261	648	155	1.9442	0.5472
Flare	WO-ARM	0.7251	0.3149	2001	24	1.2633	0.2443
	LWOA	0.7502	**0.3356**	**1036**	65	1.2669	0.2732
	LWOA-ABC	**0.781**	0.2401	1814	573	**1.5492**	**0.459**
	LWOA-CS	0.7782	0.2579	2687	**853**	1.4879	0.4115
	LWOA-GWO	0.7651	0.281	1893	288	1.3972	0.3683
Hepatitis	WO-ARM	0.6888	0.3754	581	60	1.1588	0.1902
	LWOA	0.7082	**0.3917**	**498**	124	1.1999	0.2531
	LWOA-ABC	0.8046	0.1738	664	1976	2.1316	0.6748
	LWOA-CS	**0.811**	0.1758	1467	**3094**	**2.1539**	**0.6864**
	LWOA-GWO	0.729	0.2784	699	1342	1.5874	0.4493
Ionosphere	WO-ARM	0.6222	**0.2545**	1180	227	2.0063	0.361
	LWOA	0.6376	0.2249	**776**	286	2.2595	0.4154
	LWOA-ABC	0.7801	0.1235	1278	2331	3.9972	0.7151
	LWOA-CS	**0.7968**	0.117	2567	**4726**	**4.3041**	**0.7452**
	LWOA-GWO	0.653	0.1562	1304	2555	3.0477	0.5261
Led7	WO-ARM	0.6465	**0.2604**	5546	29	1.2181	0.2106
	LWOA	0.6444	0.2477	**2405**	91	1.3213	0.2387
	LWOA-ABC	0.6526	0.2035	5699	546	**1.4182**	0.2831
	LWOA-CS	0.656	0.2066	7805	**691**	1.404	0.281
	LWOA-GWO	**0.6635**	0.225	5808	187	1.4119	**0.3058**

Figure 10 indicates the average confidence and support analysis of five algorithms on different datasets, from which it has been analyzed that the average confidence value of proposed algorithms is significantly higher than WO-ARM. Specifically, in the Hepatitis, the average confidence value of rules generated by LWOA-ABC and LWOA-CS is more than 0.8, which is 11%-12% higher than WO-ARM. On the Ionosphere dataset, the average confidence value of LWOA-CS is 16% higher than LWOA and 17% higher than WO-ARM. The confidence level of LWOA-GWO is slightly higher than that of LWOA. On all datasets, the average confidence value of LWOA-GWO is 1.9% higher than LWOA and 3% higher than WO-ARM. It indicates that hybrid algorithms LWOA-CS, LWOA-ABC, and LWOA-GWO can mine association rules of higher quality compared to the remaining two

algorithms. For support, it can be seen from the figure that the average support value of WO-ARM and LWOA is mostly higher than other algorithms. If we take the average results on all datasets, the average support value of rules generated by WO-ARM and LWOA is about 0.32 and 0.32 respectively. Likewise, the support value of LWOA-ABC, LWOA-CS, and LWOA-GWO is approximately 0.22, 0.23, and 0.26 respectively. The average support value of hybrid algorithms proposed in the paper is lower than WO-ARM and LWOA.

Figure 10. Comparison of average confidence and support

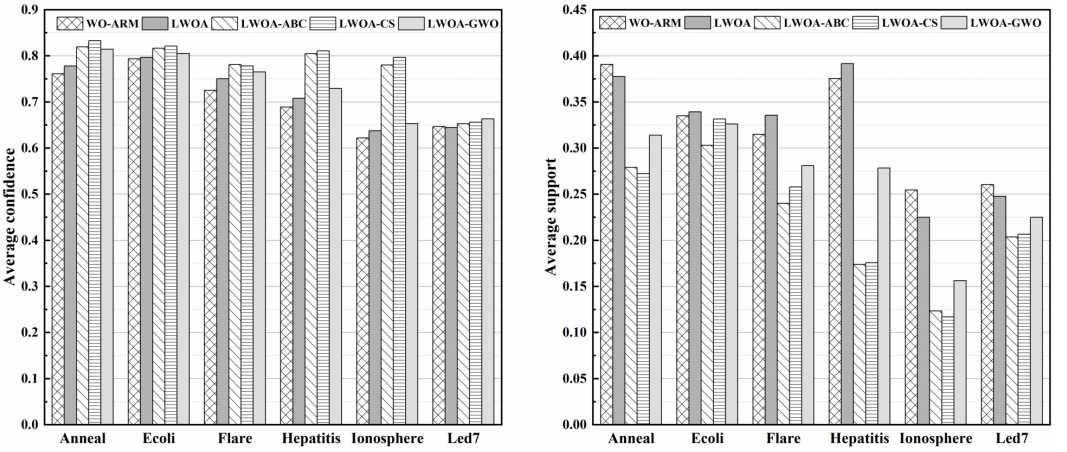
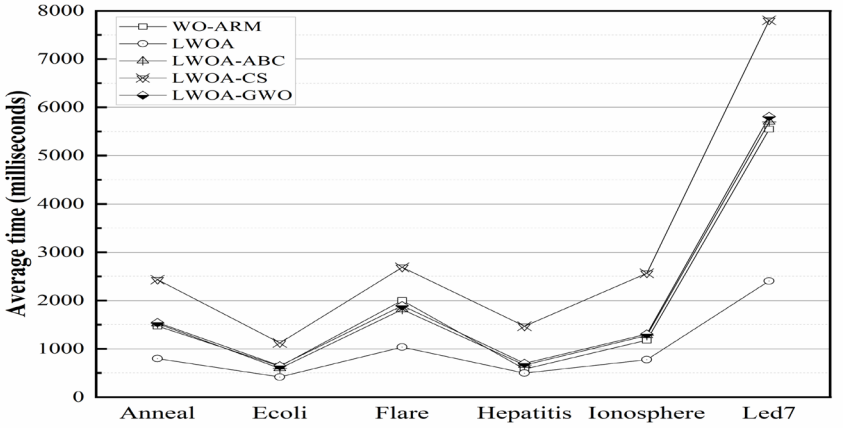


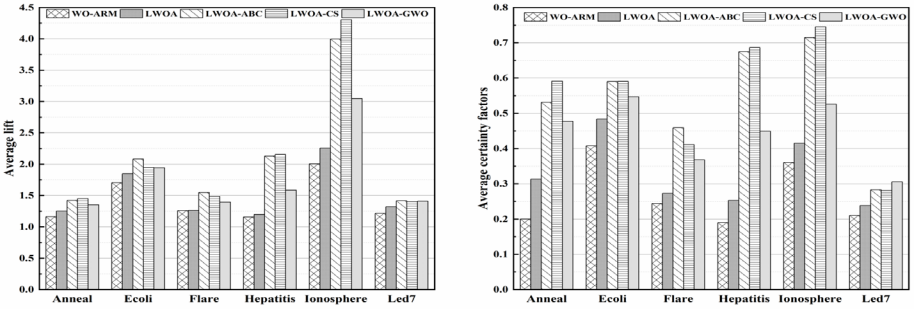
Figure 11 shows the running efficiency of each algorithm, from which it can be seen that LWOA takes the least amount of time and LWOA-CS takes the most amount of time. For instance, in the Led7, the running time of LWOA is 2405 milliseconds, which is equal to 43.3% of WO-ARM. It indicates that LWOA has an improvement in execution time compared to WOA. There is little difference in the running time of LWOA-GWO, LWOA-ABC, and WO-ARM: 5808, 5699, and 5546 milliseconds respectively. Because CS contains a Levy flight strategy and has the largest number of rules, the running time of LWOA-CS is higher than LWOA-ABC and LWOA-GWO.

Figure 11. Comparison of average time



The average lift and the average CF are used to test the relevance of the association rules. It can be seen from Figure 12 that the lift and CF values of all algorithms are more significant than 1.0, and the values obtained by hybrid algorithms are significantly higher than those generated by WO-ARM and LWOA. For the Ionosphere dataset, the lift value of LWOA-CS is 2.3% higher than WO-ARM and 2% higher than LWOA. The CF value of LWOA-ABC, LWOA-CS, and LWOA-GWO is 0.7151, 0.7425, and 0.5261 respectively. It can be inferred that hybrid algorithms proposed in the paper have better performance in both lift and CF, which means that the mined association rules have strong relevance.

Figure 12. Comparison of average lift and certainty factor



On analysis of the above tables and figures, it could be inferred that the algorithms proposed in the paper can obtain strong association rules with high confidence while losing as little support as possible. The best performer on all datasets is LWOA-CS, followed by LWOA-ABC and LWOA-GWO, which have much better optimization capability than WO-ARM. From Figures 7-9, it can be concluded that LWOA-CS has a stronger global search ability compared to the LWOA, which is consistent with the original intention of the hybrid hypothesis proposed in the paper. LWOA-ABC has the fastest convergence speed, especially in the Ionosphere, which proves that ABC can approach the optimal solution at the early stage and guide LWOA to exploit it. The iteration curves of LWOA-GWO and LWOA have similar trends with the increase of iteration times, but LWOA-GWO has been able to fetch better quality solutions, which indicates LWOA-GWO has a stronger exploration and exploitation. As a result, the hybrid algorithms proposed in the paper are effective and robust for mining association rules.

CONCLUSION

Data mining is one of the most essential techniques for discovering knowledge from data and transactions. ARM is an important data mining method used to spot the hidden patterns in data. A new hybrid strategy based on WOA is proposed in the paper for mining association rules, which is based on the idea that hybrid algorithms spend less time mining high-quality association rules without looking for frequent itemsets. Experiments on six benchmark data sets prove that the hybrid algorithms are better than standard WOA in most evaluation metrics. In terms of lift and CF, LWOA-ABC and LWOA-CS are apparently higher than the related algorithm WO-ARM. Moreover, the average confidence of hybrid algorithms is 2%-8% higher than other algorithms in all datasets, among which LWOA-CS performs best. Generally, the proposed association rule mining methods are able to keep an excellent balance between the efficiency and quality of mining rules, which makes them more proper to deal with the practical work. In the future, there is interest to improve the three position-update strategies

in WOA for the multi-objective association rule mining problem. In addition, these hybrid methods can also be effectively used in other data mining techniques, such as classification and clustering.

CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

FUNDING AGENCY

This work is funded by Fujian Provincial Key Laboratory of Data Intensive Computing and Key Laboratory of Intelligent Computing and Information Processing, Fujian No.BD201801.

ACKNOWLEDGMENT

REFERENCES

Abdulgader, M., & Kaur, D. (2019). Evolving Mamdani fuzzy rules using swarm algorithms for accurate data classification. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 175907–175916. doi:10.1109/ACCESS.2019.2957735

Akay, B., Karaboga, D., Gorkemli, B., & Kaya, E. (2021). A survey on the Artificial Bee Colony algorithm variants for binary, integer and mixed integer programming problems. *Applied Soft Computing*, *106*, 107351. doi:10.1016/j.asoc.2021.107351

Anand Hareendran, S., & Vinod Chandra, S. S. (2017). Association rule mining in healthcare analytics. In *International Conference on Data Mining and Big Data* (pp. 31-39). Springer. doi:10.1007/978-3-319-61845-6_4

Baró, G. B., Martínez-Trinidad, J. F., Rosas, R. M. V., Ochoa, J. A. C., González, A. Y. R., & Cortés, M. S. L. (2020). A PSO-based algorithm for mining association rules using a guided exploration strategy. *Pattern Recognition Letters*, *138*, 8–15. doi:10.1016/j.patrec.2020.05.006

ÇELİK, U. (2019). WOA-miner: Classification rule discovery using whale optimization algorithm. *Cumhuriyet Science Journal*, *40*(1), 186–196.

Chiclana, F., Kumar, R., Mittal, M., Khari, M., Chatterjee, J. M., & Baik, S. W. (2018). ARM–AMO: An efficient association rule mining algorithm based on animal migration optimization. *Knowledge-Based Systems*, *154*, 68–80. doi:10.1016/j.knosys.2018.04.038

Cuong-Le, T., Minh, H. L., Khatir, S., Wahab, M. A., Tran, M. T., & Mirjalili, S. (2021). A novel version of Cuckoo search algorithm for solving optimization problems. *Expert Systems with Applications*, *186*, 115669. doi:10.1016/j.eswa.2021.115669

Danish, Z., Shah, H., Tairan, N., Gazali, R., & Badshah, A. (2019). Global artificial bee colony search algorithm for data clustering. *International Journal of Swarm Intelligence Research*, *10*(2), 48–59. doi:10.4018/IJSIR.2019040104

Das, S., Tamakloe, R., Zubaidi, H., Obaid, I., & Alnedawi, A. (2021). Fatal pedestrian crashes at intersections: Trend mining using association rules. *Accident; Analysis and Prevention*, *160*, 106306. doi:10.1016/j.aap.2021.106306 PMID:34303494

Heraguemi, K., Kadri, H., & Zabi, A. (2021). Whale optimization algorithm for solving association rule mining issue. *International Journal of Computing and Digital Systems*, *10*(1), 333–342. doi:10.12785/ijcds/100133

Heraguemi, K. E., Kamel, N., & Drias, H. (2015). Association rule mining based on bat algorithm. *Journal of Computational and Theoretical Nanoscience*, *12*(7), 1195–1200. doi:10.1166/jctn.2015.3873

Heraguemi, K. E., Kamel, N., & Drias, H. (2018). Multi-objective bat algorithm for mining numerical association rules. *International Journal of Bio-inspired Computation*, *11*(4), 239–248. doi:10.1504/IJBIC.2018.092797

Hong, J., Tamakloe, R., & Park, D. (2020). Application of association rules mining algorithm for hazardous materials transportation crashes on expressway. *Accident; Analysis and Prevention*, *142*, 105497. doi:10.1016/j.aap.2020.105497 PMID:32442668

Jyoti, B., & Sharma, A. K. (2020). A perspective view of mining rules with ant colony optimization technique: Ant-miner. In *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)* (pp. 804-808). IEEE.

Kotary, D. K., Nanda, S. J., & Gupta, R. (2021). A many-objective whale optimization algorithm to perform robust distributed clustering in wireless sensor network. *Applied Soft Computing*, *110*, 107650. doi:10.1016/j.asoc.2021.107650

Kumari, P. L., Sanjeevi, S. G., & Rao, T. M. (2019). Mining top-k regular high-utility itemsets in transactional databases. *International Journal of Data Warehousing and Mining*, *15*(1), 58–79. doi:10.4018/IJDWM.2019010104

Kuo, R. J., Gosumolo, M., & Zulvia, F. E. (2019). Multi-objective particle swarm optimization algorithm using adaptive archive grid for numerical association rule mining. *Neural Computing & Applications*, *31*(8), 3559–3572. doi:10.1007/s00521-017-3278-z

Mehedi, I. M., Ahmadipour, M., Salam, Z., Ridha, H. M., Bassi, H., Rawa, M. J. H., Ajour, M., Abusorrah, A., & Abdullah, M. P. (2021). Optimal feature selection using modified cuckoo search for classification of power quality disturbances. *Applied Soft Computing*, *113*, 107897. doi:10.1016/j.asoc.2021.107897

Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in Engineering Software*, *95*, 51–67. doi:10.1016/j.advengsoft.2016.01.008

Mlakar, U., Zorman, M., Fister, I. Jr, & Fister, I. (2017). Modified binary cuckoo search for association rule mining. *Journal of Intelligent & Fuzzy Systems*, *32*(6), 4319–4330. doi:10.3233/JIFS-16963

Mohammed, R. A., & Duaimi, M. G. (2018). Association rules mining using cuckoo search algorithm. *International Journal of Data Mining. Modelling and Management*, *10*(1), 73–88.

Samantaray, S., & Sahoo, A. (2021). Prediction of suspended sediment concentration using hybrid SVM-WOA approaches. *Geocarto International*, 1–27. doi:10.1080/10106049.2021.1920638

Sarath, K. N. V. D., & Ravi, V. (2013). Association rule mining using binary particle swarm optimization. *Engineering Applications of Artificial Intelligence*, *26*(8), 1832–1840. doi:10.1016/j.engappai.2013.06.003

Sharma, P., Tiwari, S., & Gupta, M. (2015). Association rules optimization using artificial bee colony algorithm with mutation. *International Journal of Computers and Applications*, *116*(13).

Sharmila, S., & Vijayarani, S. (2021). Association rule mining using fuzzy logic and whale optimization algorithm. *Soft Computing*, *25*(2), 1431–1446. doi:10.1007/s00500-020-05229-4

Sheikhi, S. (2021). An effective fake news detection method using WOA-xgbTree algorithm and content-based features. *Applied Soft Computing*, *109*, 107559. doi:10.1016/j.asoc.2021.107559

Shu, Z., Ye, Z., Zong, X., Liu, S., Zhang, D., Wang, C., & Wang, M. (2022). A modified hybrid rice optimization algorithm for solving 0-1 knapsack problem. *Applied Intelligence*, *52*(5), 5751–5769. doi:10.1007/s10489-021-02717-4

Telikani, A., Gandomi, A. H., & Shahbahrami, A. (2020). A survey of evolutionary computation for association rule mining. *Information Sciences*, *524*, 318–352. doi:10.1016/j.ins.2020.02.073

Too, J., Mafarja, M., & Mirjalili, S. (2021). Spatial bound whale optimization algorithm: An efficient high-dimensional feature selection approach. *Neural Computing & Applications*, *33*(23), 16229–16250. doi:10.1007/s00521-021-06224-y

Viktoratos, I., Tsadiras, A., & Bassiliades, N. (2018). Combining community-based knowledge with association rule mining to alleviate the cold start problem in context-aware recommender systems. *Expert Systems with Applications*, *101*, 78–90. doi:10.1016/j.eswa.2018.01.044

Wen, F., Zhang, G., Sun, L., Wang, X., & Xu, X. (2019). A hybrid temporal association rules mining method for traffic congestion prediction. *Computers & Industrial Engineering*, *130*, 779–787. doi:10.1016/j.cie.2019.03.020

Wu, Z. X., Huang, K. W., Chen, J. L., & Yang, C. S. (2019). A memetic fuzzy whale optimization algorithm for data clustering. In *2019 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1446-1452). IEEE. doi:10.1109/CEC.2019.8790044

Yin, B., Wang, C., & Abza, F. (2020). New brain tumor classification method based on an improved version of whale optimization algorithm. *Biomedical Signal Processing and Control*, *56*, 101728. doi:10.1016/j.bspc.2019.101728