Mathematical and Computational Applications

MDPI

Article

Evolutionary Selection of a Set of Association Rules Considering Biological Constraints Describing the Prevalent Elements in Bacterial Vaginosis

María Concepción Salvador-González [1], Juana Canul-Reich [1,*], Rafael Rivera-López [2], Efrén Mezura-Montes [3] and Erick de la Cruz-Hernandez [4]

[1] DACyTI, Universidad Juárez Autónoma de Tabasco, Cunduacán 86690, Tabasco, Mexico; mcsalvadorg@gmail.com
[2] DSC, Tecnológico Nacional de México, Instituto Tecnológico de Veracruz, Veracruz 91897, Veracruz, Mexico; rafael.rl@veracruz.tecnm.mx
[3] IIIA, Universidad Veracruzana, Xalapa 91097, Veracruz, Mexico; emezura@uv.mx
[4] DAMC, Universidad Juárez Autónoma de Tabasco, Comalcalco 86658, Tabasco, Mexico; erick.delacruz@ujat.mx
* Correspondence: juana.canul@ujat.mx

Abstract: Bacterial Vaginosis is a common disease and recurring public health problem. Additionally, this infection can trigger other sexually transmitted diseases. In the medical field, not all possible combinations among the pathogens of a possible case of Bacterial Vaginosis are known to allow a diagnosis at the onset of the disease. It is important to contribute to this line of research, so this study uses a dataset with information from sexually active women between 18 and 50 years old, including 17 numerical attributes of microorganisms and bacteria with positive and negative results for BV. These values were semantically categorized for the Apriori algorithm to create the association rules, using support, confidence, and lift as statistical metrics to evaluate the quality of the rules, and incorporate those results in the objective function of the DE algorithm. To guide the evolutionary process we also incorporated the knowledge of a human expert represented as a set of biologically meaningful constraints. Thus, we were able to compare the performance of the rand/1/bin and best/1/bin versions from Differential Evolution to analyze the results of 30 independent executions. Therefore the experimental results allowed a reduced subset of biologically meaningful association rules by their executions, dimension, and DE version to be selected.

Keywords: differential evolution; association rules; bacterial vaginosis

Citation: Salvador-González, M.C.; Canul-Reich, J.; Rivera-López, R.; Mezura-Montes, E.; de la Cruz-Hernandez, E. Evolutionary Selection of a Set of Association Rules Considering Biological Constraints Describing the Prevalent Elements in Bacterial Vaginosis. *Math. Comput. Appl.* **2023**, *28*, 75. https://doi.org/10.3390/mca28030075

Academic Editor: Suchuan Dong

Received: 16 March 2023
Revised: 27 May 2023
Accepted: 12 June 2023
Published: 14 June 2023

1. Introduction

Bacterial Vaginosis (BV) is a common disturbance of the balance of vaginal flora; about 25% of women of childbearing age suffer BV [1]. It is a disease that can be asymptomatic, but symptoms such as discharge, bad vaginal odor, and increased PH can also occur. It can also increase the risk of contracting other infections such as Neisseria gonorrhea, Chlamydia trachomatis, Herpes type 2, and papillomavirus infection, among other sexually transmitted diseases, in addition to being a recurring disease [2,3]. Diseases such as BV and those sexually transmitted can lead to contracting more severe illnesses, including cervical cancer, as has been demonstrated by studies that evaluate the 16S rRNA sequencing to measure the diversity of the vaginal microbiota in women with different BV, human papillomavirus (HPV), and cervical intraepithelial neoplasia (CIN) status [4]. This study compares the microbiota composition of several women to gain insight into a marker of vaginal dysbiosis. The authors use logistic regression to identify risk factors for CIN, such as age, gestational and childbirth history, contraceptive methods, number of sexual partners, BV status, HPV infection status, and condom use. The results show that BV and HR-HPV infection are risk factors for CIN.

Bacterial Vaginosis is a public health problem. The literature mentions that, in healthy vaginal microbiota, lactobacilli are predominating. Otherwise, when lactobacilli are replaced by several bacteria, such as Gardnerella vaginalis and Atopobium vaginae, among others, there exists an imbalance in the vaginal flora that, in most cases, corresponds to a BV. The Nugent score and molecular biology are commonly used for BV diagnosis in the medical area. However, there is no certainty about the causes of this disease. Neither are all the possible pathogen combinations that can cause BV known, because it has a high recurrence rate, and this is essential to identify and treat this disease appropriately [5].

On the other hand, association rules (AR) are one of the four primary data mining tasks [6]. AR algorithms try to find relationships and frequent patterns between data. Quality metrics such as support, confidence, lift, hyper lift, and Fisher's exact test, among others, are used to identify the best patterns [7].

Association Rule Mining (ARM) has been combined with other methods, such as Differential Evolution, in optimization problems with a single objective. In the case of numerical data, they are discretized by grouping into consecutive intervals [8].

Various DE-based methods exist for ARM, such as the DE for ARM using numerical and categorical attributes (ARM-DE), where categorical attributes are discretized into numerical values such as 0 or 1 and encoded in a real-valued parameter vector [9]. Another approach is the Numerical Association Rule Minning (NiaARM), available in its free version with Python libraries [10]. Moreover, other authors have proposed the DE for mining a significant fuzzy association rules (DESigFAR) algorithm that uses fuzzy intervals to discretize the attributes. The authors evaluate each candidate rule using statistical tests and compare their proposal against one genetic algorithm [11]. Although several approaches have been proposed, the use of DE-based algorithms to reduce association rules previously generated by the Apriori algorithm, applied to discover patterns leading to VB, has yet to be studied.

For the reasons mentioned above, the interest of this study is to reduce the number of association rules for contributing to the identification of possible combinations between pathogens of possible bacterial vaginosis with Association Rule Mining and Evolutionary Computation techniques using an adaptation of the Differential Evolution (DE) algorithm to find biologically meaningful association rules from a set of association rules derived from the Apriori algorithm. The use of DE is proposed to decrease the number of rules that were previously generated with Apriori [12]. Another advantage of DE is that it allows the application of biological constraints, thus we claim that the set of rules obtained with Differential Evolution is smaller than that obtained with Apriori alone, and still meets the biological significance required for the diagnosis of diagnostic BV.

2. Materials and Methods

The data used for this study are a dataset with 17 numerical attributes with medical information from 201 sexually active women aged 18 to 50 years who underwent their routine annual gynecological examination at the Metabolic and Infectious Diseases Research Laboratory of the Universidad Juárez Autónoma de Tabasco, and who gave their written consent. The study was designed according to international standards for responsible publication of (COPE) and registered (protocol No. UJAT–20160006) and approved by the Institutional Review Board of the Universidad Juarez Autonoma de Tabasco [13]. We considered 186 records with a positive and negative diagnosis for bacterial vaginosis only. The numerical attributes of integer type used are the density of Lactobacillus crispatus, gasseri, jensenii, and iners. In addition to microorganisms mainly related to BV, Megasphaera type 1, Atopobium vaginae, and Gardnerella vaginalis.

Association rule mining is responsible for discovering interesting patterns within a dataset and is one of the most important knowledge-discovery techniques [14]. An association rule has the form $X \Rightarrow Y$, where X in the rule is called the antecedent, and Y is called consequent [15]. To measure the quality of the association rules, quality metrics are used. The interest of this research is to find association patterns between the pathogens that

cause bacterial vaginosis. According to other authors [16–18], metrics such as Confidence and Lift calculate their values according to the relationship between the antecedent and consequent of a rule.

The quality metrics of our interest are described below [19]:

- Support: It is the number of times the element appears.
- Confidence: It is based on the support of frequent itemsets to generate significant rules according to the value of the confidence that one wants to look for.
- Lift: Calculate the number of times the antecedent and consequent occur together.

Other metrics that were evaluated in this work are the following: [19]:

- Fisher Exact Test: Each rule represents a one-sided Fisher's exact statistical test and the correction is used for multiple comparisons.
- Hyperlift: It is a more robust metric than the lift metric. It is used at low counts and its false positives are less frequent.

The Apriori algorithm is one of the most effective methods for discovering valid, novel, and meaningful rules among data and stands out for its simplicity. However, its results exponentially grow when making associations, generating many rules [15].

The Apriori Algorithm consists of three repetitive cycles where k is the length of the pattern generated in the previous step, i are the generations, $Ck+1$ is the cycle that generates the candidate patterns that join the patterns in Fk, the cycle continues with the pruning and validation of patterns for all the database transactions in T until the set of frequent k-patterns Fk in one iteration is empty. The Algorithm 1 shows the pseudocode of the Apriori Algorithm [20].

Algorithm 1 Apriori pseudocode

Require: n

Generate frequent 1-patterns and 2-patterns using specialized counting methods and denote by $F1$ and $F2$;
$k := 2$;
while Fk is not empty do **do**
  Generate $Ck+1$ by using joins on Fk;
  Prune $Ck+1$ with $Apriori$ subset pruning trick;
  Generate $Fk+1$ by counting candidates in $Ck+1$ with respect to T at support s;
  $k := k+1$;
end while
return $\cup_{i=1}^{k} F_i$;

On the other hand, the DE process begins with the random creation of the initial population. The values of each individual in the population must fit within the pre-established limits of the search space. Then, for each individual, three vectors are combined using the mutation and crossover operators to create a new candidate solution. By comparing current with new individuals, one new population is built. The parameters used by the DE algorithm are the population size (NP), crossover rate (CR), mutation factor (F), and also the bounds of the search space [21].

The DE algorithm simulates natural evolution using vectors. Starting from a target vector $\vec{x}_{i,g}$, the search direction is calculated according to the difference of the vectors $\vec{x}_{r1,g}$ and $\vec{x}_{r2,g}$ chosen at random within the population, and its scale factor F is calculated and added to the base vector $\vec{x}_{r0,g}$ and its result is the mutated vector. The mutated vector is recombined by a binomial crossover defined by the parameter CR. Finally, a binomial cross-type is used. The pseudocode of the DE/rand/1/bin version can be found in Algorithm 2. The difference between the DE/rand/1/bin version and DE/best/1/bin is that in the latter the base vector is the best vector of the current population. The pseudocode of the DE/best/1/bin version is depicted in Algorithm 3 [22].

Algorithm 2 DE/rand/1/bin pseudocode

Require: $g = 0$
Create a random initial population $\vec{x}_{i,g} \forall i, i = 1, \ldots, NP$
Evaluate $f(\vec{x}_{i,g}) \forall i, i = 1, \ldots, NP$
for $g = 1$ to $MAXG$ **do**
  for $i = 1$ to NP **do**
    Select randomly $r_0 \neq r_1 \neq r_2 \neq i$
    $j_{rand} = randint[1, n]$
    for j=1 to n **do**
      if $rand_j[0, 1] < CR \vee j = j_{rand}$ **then**
        $u_{j,i,g+1} = x_{j,r_0,g} + F(x_{j,r_1,g} - x_{j,r_2,g})$
      else
        $u_{j,i,g+1} = x_{j,i,g}$
      end if
    end for
    if $(f(\vec{u}_{i,g+1}) \leq f(\vec{x}_{i,g})$ **then**
      $\vec{x}_{i,g+1} = \vec{u}_{i,g+1}$
    else
      $\vec{x}_{i,g+1} = \vec{x}_{i,g}$
    end if
  end for
  $g = g + 1$
end for

The main objective of the proposed approach, named the Apriori rules reduction by Differential Evolution (AR2DE) approach, is to apply the DE algorithm to select the most important set of association rules generated by the Apriori algorithm. The individuals in the population encode the original association rules using an integer-valued vector (Figure 1). Several AR metrics are included in the fitness function to identify their biological significance to c.

Algorithm 3 DE/best/1/bin pseudocode

Require: $g = 0$
Create a random initial population $\vec{x}_{i,g} \forall i, i = 1,..., NP$
Evaluate $f(\vec{x}_{i,g}) \forall i, i = 1,..., NP$
for $g = 1$ to $MAXGEN$ **do**
  for $i = 1$ to NP **do**
    Select randomly $r_0 \neq r_1 \neq r_2 \neq i$
    $j_{rand} = randint[1, n]$
    for j=1 to n **do**
      if $(rand_j[0, 1] < CR$ or $j = j_{rand})$ **then**
        $u_{j,i,g+1} = x_{j,best,g} + F(x_{j,r_1,g} - x_{j,r_2,g})$
      else
        $u_{j,i,g+1} = x_{j,i,g}$
      end if
    end for
    if $(f(\vec{u}_{i,g+1}) \leq f(\vec{x}_{i,g})$ **then**
      $\vec{x}_{i,g+1} = \vec{u}_{i,g+1}$
    else
      $\vec{x}_{i,g+1} = \vec{x}_{i,g}$
    end if
  end for
  $g = g + 1$
end for

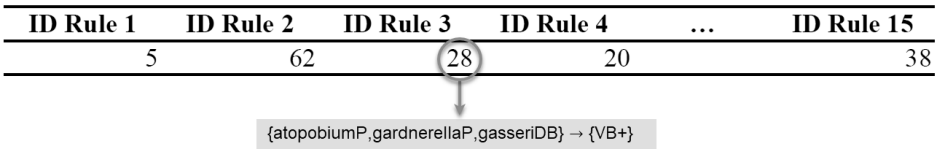

Figure 1. Encoding scheme from 1 at 15 association rules. The integers represent the ID of each association rule.

3. Results

3.1. Experimental Study

In the first part of the experimental study, of the data set consisting of 184 records of positive and negative BV cases, the attributes were discretized according to their numerical value and cataloged into linguistic concepts according to Table 1 to obtain the transactions used by the Apriori algorithm to generate the association rules, resulting in 5248 association rules.

Table 1. Antecedents.

Antecedent	Range	Classification	Type
Age	1	menoredad	Under 30 years old
	2	mayoredad	Over 30 years old
Cristpatus	1	crispatusDB	Low density
	2	crispatusDA	High density
Gasseri	1	gasseriDB	Low density
	2	gasseriDA	High density
Iners	1	inersDB	Low density
	2	inersDA	High density
Jensenii	1	jenseniDB	Low density
	2	jenseniDA	High density
Megasphaera	1	megasphaeraP	Positive
	2	megasphaeraN	Negative
Atopobium	1	atopobiumP	Positive
	2	atopobiumN	Negative
Gardnerella	1	gardnerellaP	Positive
	2	gardnerellaN	Negative

Antecedents itemset values used in the experimental study.

Below, the cases of interest in this research are the rules that have BV+, after applying the filter 91 rules are evaluated in the DE process to reduce according to their biological significance.

3.2. Analysis of Evaluation Metrics

The next part of the experiment was the analysis of the quality metrics, which evaluate the association rules generated by the Apriori algorithm. Since this study focused on the rules that had as a consequent VB+, which represents one element as a consequent, the metrics Fishers Exact Test (Figure 2), Hyperlift (Figure 3), Lift (Figure 4), and Confidence (Figure 5) were evaluated using scatter plots that allow us to visualize their range of data, maximum, and minimum values. The comparison between the graphs shows that the lowest value range is for the Fishers Exact Test metric (Figure 2), followed by Confidence. In the study that metrics are used in DE the very low value ranges do not favor the evolutionary process.

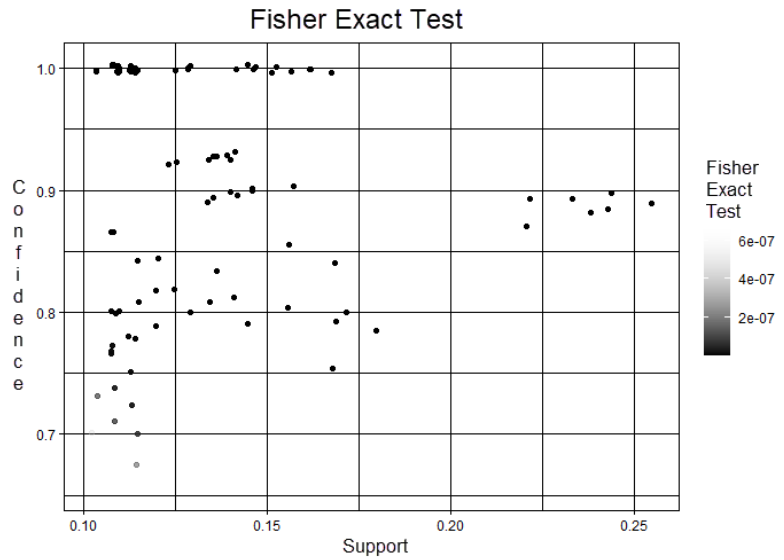

Figure 2. Fisher exact test metric scatter plot.

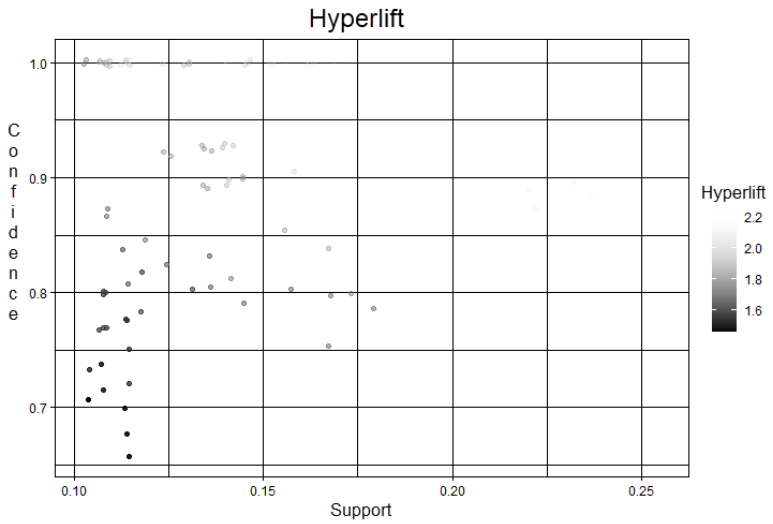

Figure 3. Hyperlift metric scatter plot.

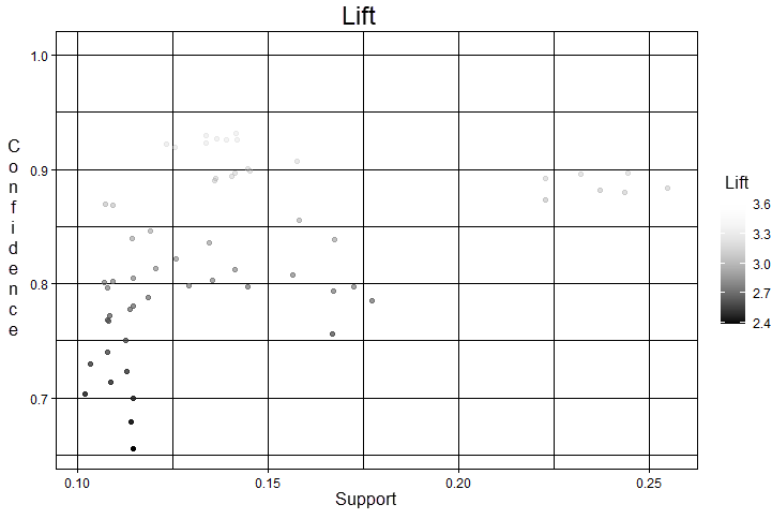

Figure 4. Lift metric scatter plot.

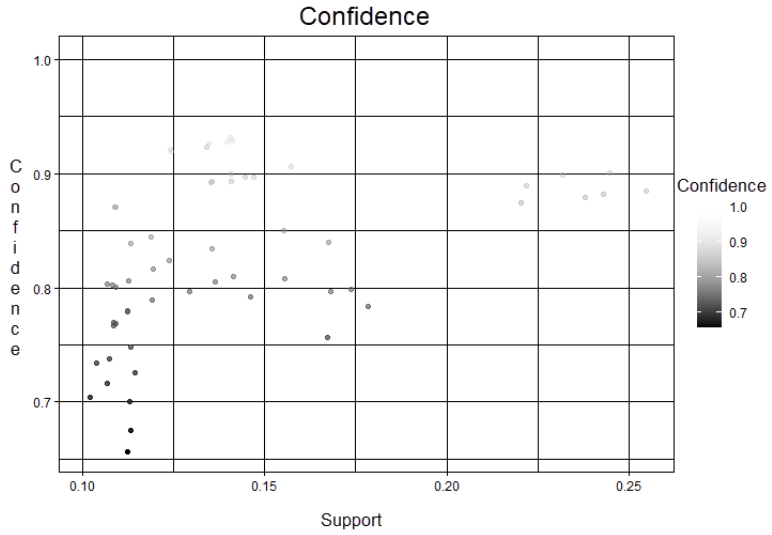

Figure 5. Confidence metric scatter plot.

The scatter plots also show that the highest ranges of values were for the Lift (Figure 4) and Hyperlift metrics (Figure 3).

The metrics that show the best correlation strength between their values according to the scatter plots are Lift (Figure 4) and Confidence (Figure 5). The main metrics according to the literature [16–18] are confidence and lift so lift represented the best option among the four evaluated metrics, and lift, having the highest range of values, favors the evolutionary process in DE.

The result of the analysis of the metrics reveals that lift and confidence are the best-evaluated metrics for this study in addition to evaluating the frequency of positive bacteria and the presence of lactobacillus iners as factors of biological significance in the association rules.

3.3. Implementation of the Differential Evolution (DE) Algorithm

The three main elements in the DE algorithm are the individuals' encoding scheme, the fitness function, and the variation operators.

1. **Encoding scheme:** An individual of the population is a subset of R association rules each identified with an ID number. Figure 1 shows an example of this codification from 1 at 15 rules by ID rule.
   In this work, the value of R is set to 1 to 15 since in [18] authors obtained five rules with a biological significance which were determined by a human expert, so the 15 tests ensure the algorithm will find this minimal set of rules.
2. **Fitness function:** Each j-th individual in the population is evaluated to define the fitness value. In this work, the fitness function $f(x_j)$ is the sum of the S metrics of the association rules encoded on the individual as follows:

$$f(x_j) = \sum_{u=1}^{R} \sum_{w=1}^{S} m_{u,w} \tag{1}$$

   where R is the number of association rules, S is the number of metrics involved to define the solution quality, and $m_{u,w}$ is the w-th metric computed for the u-th rule.
   Since metrics are parameters that allow us to know the quality of attributes quantitatively, support and confidence are normally used [23]. The metrics used in the objective function and described in Section 2 are support, confidence, and lift. In addition, the frequency of positive bacteria in the rules, and the occurrences of high values of lactobacillus iners are included to define the biological significance of the

association rules [13] in this sense higher results from the addition of the metrics have a higher significance.

3. **Variation operators:** Differential mutation and crossover operators are defined to create feasible offspring.

    - **Mutation:** Three randomly chosen individuals of the current population (x^{r_1}, x^{r_2} and x^{r_3}), are different from each other and also different from the target vector, these individuals are linearly combined to yield a *mutated vector* v^i using a user-specified scale factor F to control the differential variation, as follows:

    $$v^i = \lfloor x^{r_1} + \mathrm{F}(x^{r_2} - x^{r_3}) \rceil, \tag{2}$$

    Equation (2) is related to the DE/rand/1/bin variant defined in [24]. Another commonly used variant is known as DE/best/1/bin, where the best individual in the population $\mathbf{x}^{best}$ is combined with two randomly chosen individuals of the current population, as follows:

    $$v^i = \lfloor x^{best} + \mathrm{F}(x^{r_1} - x^{r_2}) \rceil, \tag{3}$$

    - **Crossover:** The mutated vector is recombined with the target vector to build the trial vector u^i. For each $j \in \{1, \ldots, |x^i|\}$, either x^i_j or v^i_j is selected based on a comparison between a uniformly distributed random number $r \in [0, 1]$ and the crossover rate CR. The recombination operator also uses a randomly chosen index $l \in \{1, \ldots, |x^i|\}$ to ensure that u^i acquires at least one value from v^i, as follows:

    $$u^i_j = \begin{cases} v^i_j & \text{if } r \leq \mathrm{CR} \text{ or } j = l, \\ x^i_j & \text{otherwise.} \end{cases} \tag{4}$$

In the Equations (2) and (3), $\lfloor w \rceil$ symbol denotes that the w value is rounded to the nearest integer since the encoding scheme defined for this work indicates that the parameter values are only integers. If a parameter value of a mutated vector is outside its range, it is replaced with a random value between 1 and 91.

3.4. Algorithm Parameters

It is well known that the performance of the Differential Evolution algorithm is affected by the values of its parameters: F (Scale factor), CR (Crossover rate), and NP (Population Size) [25]. The parameter values used in this work are shown in the Table 2 and are based on those commonly used in the existing literature [24]. Since this experimental study is a work in progress, no parameter-tuning process has been carried out.

Table 2. Parameters values.

Parameter	Value
F (Scale factor)	0.7
CR (Crossover rate)	0.5
NP (Population size)	30
MAXGEN (Number of generations)	30
li (lower limit)	1
ls (upper limit)	91

Parameters used in DE/rand/1 and DE/best/1 versions.

4. Discussion

In this work, 30 independent runs were made for the rand/1/bin and best/1/bin versions and 15 tests were made with each version by changing the value of the individual's dimension from 1 to 15.

Table 3 shows the results of 30 independent runs with the two DE variants included in this study (rand/1/bin and best/1/bin). The best results for both versions were when D = 15, the best fitness value in the rand/1/bin version is 96.4201 on test number 14, and for the best/1/bin version is 95.6184 on test 6.

Table 3. Results of 30 independent runs for each DE variant.

Test	Rand/1/Bin	Best/1/Bin	Test	Rand/1/Bin	Best/1/Bin
1	95.5457	94.9080	16	93.2444	92.7573
2	93.4321	94.3232	17	93.1759	93.8795
3	95.0492	95.3183	18	92.96241	94.3069
4	92.6088	93.2915	19	93.1063	95.0708
5	93.1151	94.2030	20	94.2296	93.1564
6	93.0887	**95.6184**	21	95.2363	92.8379
7	92.8634	95.4351	22	93.4080	92.9451
8	94.0896	94.9295	23	96.0235	93.0003
9	95.2359	92.6744	24	<u>94.0122</u>	94.0149
10	92.5891	94.1667	25	92.6781	93.1910
11	94.4798	94.3319	26	94.1964	93.6972
12	93.7070	95.5788	27	92.9309	92.5566
13	94.7591	93.9418	28	94.1645	93.9121
14	**96.4201**	<u>94.1041</u>	29	94.5672	94.0574
15	93.7882	94.4250	30	94.4028	94.2564

The best fitness values are highlighted in bold, and the median value of each variant is underlined.

The statistical comparison for each variant is shown in Table 4, and Figure 6 depicts the convergence plot of the run reaching the median value of the two variants. When comparing the results of the two variants using the Wilcoxon signed-rank exact test by the function wilcox.test from R, V = 163 and the *p*-value = 0.1579 indicated the data in each group are significant correlated.

Table 4. Statistical values.

Statistical Measure	Rand/1/Bin	Best/1/Bin
Best value	96.4201	95.6184
Mean	93.9703	94.0296
Median	93.9002	94.0808
Standard deviation	1.0428	0.8942
Worst value	92.5891	92.5566
Best test number	14	6
Median test number	24	14

Statistical measure for rand/1/bin and best/1/bin.

According to the statistical results, the best value is obtained with the rand/1/bin variant. However, the results obtained in the independent runs and the behavior of the convergence graph show that the best/1/bin variant had better performance in selecting the association rules.

The best individuals of each variant were taken for the 15 tests and decoded to their corresponding association rule according to their ID. Repeated rules were removed and a count of occurrences in both groups of rules was made to know the most frequent ones as shown in the table. Table 5 shows the rules encoded by the best individuals of each variant. Likewise, most of the rules comply with the biological significance requirement of having at least two bacteria present [13]. The biological significance of the items adds weight to rules that carry bacteria positivity, concurrently with showing low-density levels of lactobacillus iners. This result is concordant with clinical findings observed in women with bacterial vaginosis [26].

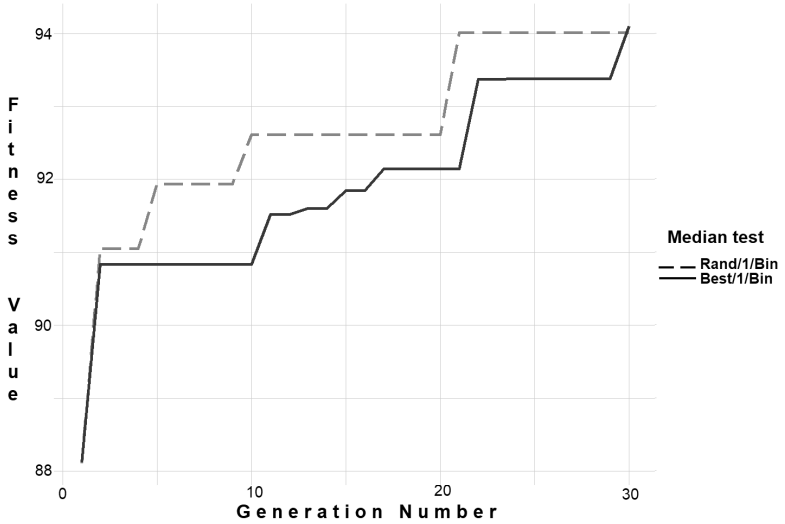

Figure 6. Convergence plot for the median values of the two DE variants.

Table 5. Set of best association rules.

ID	Association Rule
ine 1	{atopobiumP,megasphaeraP} → {VB+}
3	{jenseniiDB,megasphaeraP} → {VB+}
5	{atopobiumP,gardnerellaP} → {VB+}
10	{gardnerellaP,gasseriDB} → {VB+}
14	{atopobiumP,inersDB} → {VB+}
15	{atopobiumP,crispatusDB} → {VB+}
19	{atopobiumP,crispatusDB,megasphaeraP} → {VB+}
20	{atopobiumP,jenseniiDB,megasphaeraP} → {VB+}
21	{atopobiumP,gasseriDB,megasphaeraP} → {VB+}
22	{crispatusDB,jenseniiDB,megasphaeraP} → {VB+}
25	{atopobiumP,crispatusDB,gardnerellaP} → {VB+}
26	{atopobiumP,gardnerellaP,megasphaeraN} → {VB+}
27	{atopobiumP,gardnerellaP,jenseniiDB} → {VB+}
28	{atopobiumP,gardnerellaP,gasseriDB} → {VB+}
37	{atopobiumP,inersDA,jenseniiDB} → {VB+}
42	{atopobiumP,jenseniiDB,mayoredad} → {VB+}
46	{atopobiumP,gasseriDB,inersDB} → {VB+}
53	{atopobiumP,crispatusDB,jenseniiDB,megasphaeraP} → {VB+}
54	{atopobiumP,crispatusDB,gasseriDB,megasphaeraP} → {VB+}
55	{atopobiumP,gasseriDB,jenseniiDB,megasphaeraP} → {VB+}
57	{atopobiumP,crispatusDB,gardnerellaP,megasphaeraN} → {VB+}
58	{atopobiumP,crispatusDB,gardnerellaP,jenseniiDB} → {VB+}
59	{atopobiumP,crispatusDB,gardnerellaP,gasseriDB} → {VB+}
60	{atopobiumP,gardnerellaP,jenseniiDB,megasphaeraN} → {VB+}
61	{atopobiumP,gardnerellaP,gasseriDB,megasphaeraN} → {VB+}
62	{atopobiumP,gardnerellaP,gasseriDB,jenseniiDB} → {VB+}
65	{crispatusDB,gardnerellaP,gasseriDB,jenseniiDB} → {VB+}
72	{atopobiumP,crispatusDB,gasseriDB,mayoredad} → {VB+}
74	{atopobiumP,crispatusDB,inersDB,jenseniiDB} → {VB+}
75	{atopobiumP,crispatusDB,gasseriDB,inersDB} → {VB+}
81	{atopobiumP,crispatusDB,gasseriDB,jenseniiDB,megasphaeraP} → {VB+}
82	{atopobiumP,crispatusDB,gardnerellaP,jenseniiDB,megasphaeraN} → {VB+}
83	{atopobiumP,crispatusDB,gardnerellaP,gasseriDB,megasphaeraN} → {VB+}
84	{atopobiumP,crispatusDB,gardnerellaP,gasseriDB,jenseniiDB} → {VB+}
85	{atopobiumP,gardnerellaP,gasseriDB,jenseniiDB,megasphaeraN} → {VB+}
91	{atopobiumP,crispatusDB,gardnerellaP,gasseriDB,jenseniiDB,megasphaeraN} → {VB+}

Set of the best association rules selected from the two DE variants of the tests with dimensions from 1 to 15.

The 5 most frequent rules of the 15 tests for the rand/1/bin variant by ID are 1, 58, 21, 62, and 19. For the variant, the best/1/bin by ID are 19, 58, 62, and 83. The details for both variants are shown in Figure 7.

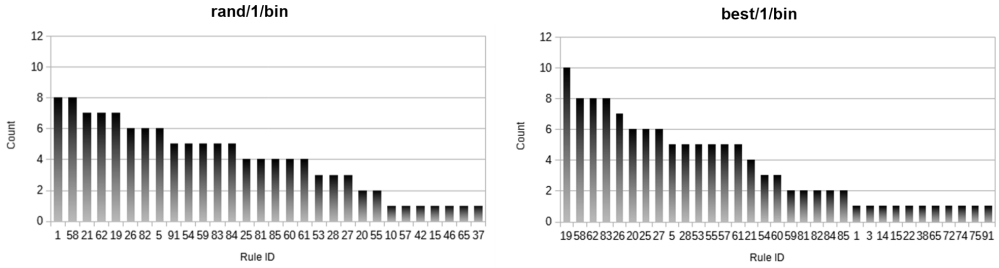

Figure 7. Detail of the frequency of rules per ID for rand/1/bin and best/1/bin variants.

The elements frequently found in the antecedent of the rules of both variants are atopobiumP, crispatusDB, gardnerellaP, jenseniiDB, gasseriDB, megasphaeraP, megasphaeraN, inersDB, inersDA, and mayoredad. The details are shown in Figure 8.

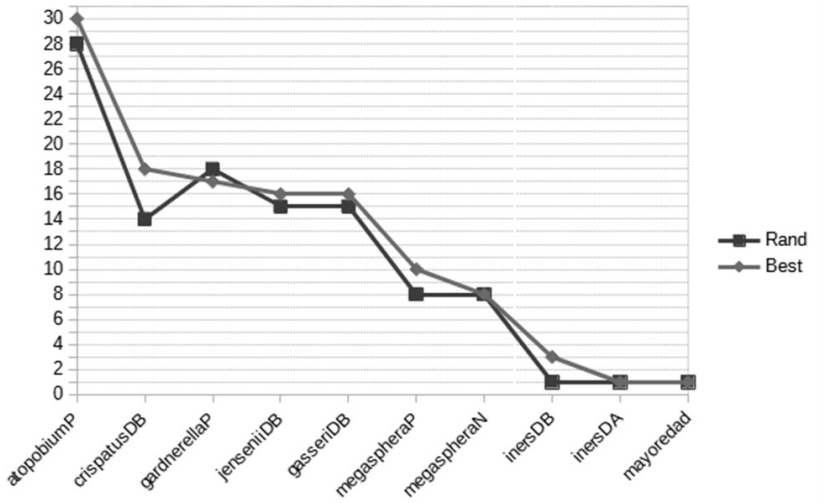

Figure 8. Frequency of antecedent elements for rand/1/bin and best/1/bin variants of the 15 tests.

5. Conclusions

The combination of the Apriori and DE algorithms enables the generation of subsets of rules with biological significance by utilizing a fitness function that incorporates the biological criteria used by experts. The analysis presented in this study demonstrates that the DE/rand/1/bin and DE/best/1/bin algorithms reveal that microorganisms such as Atopobium positive, Gardnerella positive, and L. Crispatus in low density have a greater interaction to present a VB+. The clinical findings coincide with the presence of these microorganisms, which reduce the density of lactobacilli such as L. Crispatus. However, age is not determining factor of a VB+ according to DE algorithms since it is the least frequent antecedent. This study highlights the use of DE algorithms and the integration of biologically significant rules into the objective function.

In that context, the use of DE algorithms and the integration of biological significance rules to the objective function give the expected results, obtaining mostly high-quality association rules. They comply with the requirements of the objective function by having at least two positive bacteria present.

From this perspective, the following three rules were found where there is only one bacterium and one lactobacillus:

- {jenseniiDB,megasphaeraP} → {VB+}

- {gardnerellaP,gasseriDB} → {VB+}
- {atopobiumP,crispatusDB} → {VB+}

In this sense, the validation of the expert indicates that the rules where a bacterium and a lactobacillus are present are those that can be useful for the classification of indeterminate cases, specifically in cases where L. crispatus and iners are not informative. For this reason, they cannot be ruled out and should be validated in other databases and biologically to find out their contribution to the development of the condition.

This approach provides concrete support to experts in identifying relationships that have not been explored or analyzed in the laboratory. The use of computational intelligence approaches in this field of study can be considered highly beneficial for designing new strategies to identify diseases and improve patient health. In future work, it is very important to continue with the validation of the rules by an expert and to carry out tests with a more robust dataset to integrate indeterminate cases, and other rules of biological significance to add penalties to the objective function. It is also proposed to create a new individual coding scheme that allows comparison with other evolutionary computation algorithms for association rule mining and includes parameter adjustment of the DE algorithm.

Author Contributions: Conceptualization, J.C.-R. and R.R.-L.; Formal analysis, E.M.-M.; Investigation, M.C.S.-G.; Resources, E.d.l.C.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research receives no funding from any agency.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from [13] and are available with the permission of [13].

Acknowledgments: The first author (CVU 769227) acknowledges support from the National Council of Science and Technology (CONACYT) of Mexico through a scholarship to pursue graduate studies at the Universidad Juárez Autónoma de Tabasco.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ARM	Association Rule Mining
DE	Differential Evolution
BV+	Bacterial Vaginosis Positive

References

1. Noormohammadi, M.; Eslamian, G.; Kazemi, S.N.; Rashidkhani, B. Association between dietary patterns and bacterial vaginosis: A case–control study. *Sci. Rep.* **2022**, *12*, 12199. [CrossRef] [PubMed]
2. Coudray, M.S.; Madhivanan, P. Bacterial vaginosis—A brief synopsis of the literature. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **2020**, *245*, 143–148. [CrossRef] [PubMed]
3. Onywera, H.; Anejo-Okopi, J.; Mwapagha, L.M.; Okendo, J.; Williamson, A.L. Predictive functional analysis reveals inferred features unique to cervicovaginal microbiota of African women with bacterial vaginosis and high-risk human papillomavirus infection. *PLoS ONE* **2021**, *16*, e0253218. [CrossRef] [PubMed]
4. Xu, X.; Zhang, Y.; Yu, L.; Shi, X.; Min, M.; Xiong, L.; Pan, J.; Liu, P.; Wu, G.; Gao, G. A cross-sectional analysis about bacterial vaginosis, high-risk human papillomavirus infection, and cervical intraepithelial neoplasia in Chinese women. *Sci. Rep.* **2022**, *12*, 6609. [CrossRef] [PubMed]
5. Abou Chacra, L.; Fenollar, F.; Diop, K. Bacterial vaginosis: What do we currently know? *Front. Cell. Infect. Microbiol.* **2022**, *11*, 1393. [CrossRef] [PubMed]
6. Dhaenens, C.; Jourdan, L. Metaheuristics for data mining: Survey and opportunities for big data. *Ann. Oper. Res.* **2022**, *314*, 117–140. [CrossRef]
7. Telikani, A.; Gandomi, A.H.; Shahbahrami, A. A survey of evolutionary computation for association rule mining. *Inf. Sci.* **2020**, *524*, 318–352. [CrossRef]
8. Varol Altay, E.; Alatas, B. Performance analysis of multi-objective artificial intelligence optimization algorithms in numerical association rule mining. *J. Ambient Intell. Humaniz. Comput.* **2020**, *11*, 3449–3469. [CrossRef]

9. Fister, I.; Iglesias, A.; Galvez, A.; Del Ser, J.; Osaba, E.; Fister, I. Differential evolution for association rule mining using categorical and numerical attributes. In Proceedings of the 19th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2018), Madrid, Spain, 21–23 November 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 79–88.
10. Stupan, Ž.; Fister, I. NiaARM: A minimalistic framework for Numerical Association Rule Mining. *J. Open Source Softw.* **2022**, *7*, 4448. [CrossRef]
11. Zhang, A.; Shi, W. Mining significant fuzzy association rules with differential evolution algorithm. *Appl. Soft Comput.* **2020**, *97*, 105518. [CrossRef]
12. SuryaNarayana, G.; Kolli, K.; Ansari, M.D.; Gunjan, V.K. A traditional analysis for efficient data mining with integrated association mining into regression techniques. In *Proceedings of the 3rd International Conference on Communications and Cyber Physical Engineering (ICCCE 2020)*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 1393–1404.
13. Sanchez-Garcia, E.K.; Contreras-Paredes, A.; Martinez-Abundis, E.; Garcia-Chan, D.; Lizano, M.; de la Cruz Hernandez, E. Molecular epidemiology of bacterial vaginosis and its association with genital micro-organisms in asymptomatic women. *J. Med. Microbiol.* **2019**, *68*, 1373–1382. [CrossRef] [PubMed]
14. Lin, H.K.; Hsieh, C.H.; Wei, N.C.; Peng, Y.C. Association rules mining in R for product performance management in industry 4.0. *Procedia CIRP* **2019**, *83*, 699–704. [CrossRef]
15. Agrawal, R.; Srikant, R. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994), Santiago, Chile, 12–15 September 1994; Volume 1215, pp. 487–499.
16. Shigetoh, H.; Nishi, Y.; Osumi, M.; Morioka, S. Combined abnormal muscle activity and pain-related factors affect disability in patients with chronic low back pain: An association rule analysis. *PLoS ONE* **2020**, *15*, e0244111. [CrossRef] [PubMed]
17. Olow, A.K.; van't Veer, L.; Wolf, D.M. Toward developing a metastatic breast cancer treatment strategy that incorporates history of response to previous treatments. *BMC Cancer* **2021**, *21*, 212. [CrossRef] [PubMed]
18. de la Cruz Ruiz, F.; Canul-Reich, J. Reglas de asociación para el estudio de la vaginosis bacteriana. *Komputer Sapiens* **2022**, *II*, 26–30.
19. Hahsler, M. A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules. Available online: https://mhahsler.github.io/arules/docs/measures (accessed on 15 March 2023).
20. Aggarwal, C.C.; Bhuiyan, M.A.; Hasan, M.A. *Frequent Pattern Mining Algorithms: A Survey*; Springer: Berlin/Heidelberg, Germany, 2014.
21. Storn, R.; Price, K. Differential Evolution—A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. Glob. Optim.* **1997**, *11*, 341–359. [CrossRef]
22. Mezura-Montes, E.; Miranda-Varela, M.E.; del Carmen Gómez-Ramón, R. Differential evolution in constrained numerical optimization: An empirical study. *Inf. Sci.* **2010**, *180*, 4223–4262. [CrossRef]
23. Bramer, M. *Principles of Data Mining*; Springer: Berlin/Heidelberg, Germany, 2007; Volume 180, pp. 205–209.
24. Price, K.; Storn, R.M.; Lampinen, J.A. *Differential Evolution: A Practical Approach to Global Optimization*; Springer: Berlin/Heidelberg, Germany, 2006. [CrossRef]
25. Das, S.; Suganthan, P.N. Differential Evolution: A Survey of the State-of-the-Art. *IEEE Trans. Evol. Comput.* **2011**, *15*, 4–31. [CrossRef]
26. Zariffard, M.R.; Saifuddin, M.; Sha, B.E.; Spear, G.T. Detection of bacterial vaginosis-related organisms by real-time PCR for Lactobacilli, Gardnerella vaginalis and Mycoplasma hominis. *FEMS Immunol. Med. Microbiol.* **2002**, *34*, 277–281. [CrossRef] [PubMed]