ISSN 1870-4069

Three Metrics for Building Association Rules Using Differential Evolution for Bacterial Vaginosis

Freddy García-Fuentes[1], Juana Canul-Reich[1], Rafael Rivera-López[2], Efrén Mezura-Montes[3], Erick De-La-Cruz-Hernández[4]

[1] Universidad Juárez Autónoma de Tabasco, DACYTI, Mexico

[2] Instituto Tecnológico de Veracruz, Mexico

[3] Universidad Veracruzana, IIIA, Mexico

[4] Universidad Juárez Autónoma de Tabasco, DAMC, Mexico

211H18004@alumno.ujat.mx,juana.canul@ujat.mx, rafael.rl@veracruz.tecnm.mx, emezura@uv.mx, erick.delacruz@ujat.mx

Abstract. This paper proposes a Differential-Evolution-based approach to building association rules describing the relationship between the elements triggering bacterial vaginosis. The differential evolution algorithm uses a population of real-valued vectors as candidate solutions representing an association rule, which is evaluated on the clinical dataset using a fitness function. This function includes three metrics to measure the quality rule to generalize the dataset. The dataset has binary attributes representing the absence or presence of the bacteria. The resulting rules indicate that this approach can build rules with biological significance.

Keywords: Bacterial vaginosis, differential evolution, association rules.

1 Introduction

Bacterial vaginosis (BV) is a vaginal infection characterized by a bacterial disorder due to a change in the vaginal flora. The anaerobic pathogens that increase concentration by 10 to 100 times are species of *Prevotella and Peptostreptococcus, Gardnerella vaginalis, Mobiluncus spp, Bacteroides spp, Peptostreptococcus spp, Urea-plasma urealyticum, and Mycoplasma hominis* [4]. BV is a silent health problem since the symptoms can go unnoticed, causing severe consequences such as premature delivery, post-abortion infection, pelvic inflammatory disease, and sexually transmitted diseases [8].

Data Mining is an exciting Artificial Intelligence area that allows for information analysis and knowledge discovery. Data Mining techniques such as association rule mining (AR) allow identifying correlation, association, and frequent patterns from a dataset [6]. In particular, the Apriori algorithm [14] is one of the most widely used algorithms for pattern discovery, using frequent itemsets to generate association rules [3]. A disadvantage of the Apriori algorithm is the combinatorial exploitation of the rules produced, so applying techniques to obtain a reduced set of high-quality rules is essential [7]. To address this problem, we propose using a Differential-Evolution-based approach to create meaningful association rules of high quality. The Differential Evolution (DE) algorithm is an effective method used in numerical optimization, which both explores and exploits a set of solutions in a continuous space based on an intelligent and robust combination scheme [1].

The rest of this manuscript is organized as follows: Section 2 describes the dataset and the methods used in this work. The elements used to implement the proposed method are detailed in Section 3, and Section 5 shows the preliminary results. Finally, conclusions and future work are described in Section 6.

2 Materials and Methods

2.1 Data

The dataset used in this work includes information from Mexican sexually active women between 18 and 50 years old. They underwent a gynecological examination at the Metabolic and Infectious Diseases Research Laboratory of the Universidad Juárez Autónoma de Tabasco [12]. The dataset comprises 201 observations and 11 attributes that led to a BV presence or absence diagnosis. Attributes used describes four lactoballici (*crispatus*, *gasseri*, *iners*, and *jensenni*) and seven bacteria (*atopobium*, *garnerella vaginalis*, *megasphaera*, *mycoplasma hominis*, *ureaplasma parvum*, *ureaplasma urealuytum*, and *mycoplasma genitalium*). Fifty-one positive and 134 negative vaginosis cases exist, and 16 indeterminate cases.

2.2 Association Rules

Association rule mining (AR) is a data mining technique that identifies associations between attributes from a dataset [2]. Association rule is formalized as the implication $if \ldots then \ldots$ in the form $A \Rightarrow B$, where A is the antecedent and B the consequent. To select an AR as a candidate, it must meet some quality measure. The measures most commonly used are Support, Confidence, and Lift. They are defined as follows:

$$\text{Support}(A \Longrightarrow B) = P(A \cup B), \tag{1}$$

$$\text{Confidence}(A \Longrightarrow B) = P(A|B), \tag{2}$$

$$\text{Lift}(A \Longrightarrow B) = \frac{\text{Confidence}(A \Longrightarrow B)}{\text{Support}(B)}. \tag{3}$$

2.3 Differential Evolution Algorithm

Differential Evolution (DE) is an efficient evolutionary algorithm for solving optimization problems in continuous spaces [13]. DE encodes candidate solutions through real-valued vectors and applies a difference vector to disrupt a population of these solutions. First, a population of candidate solutions is randomly created, then applying the DE evolutionary process that builds a new population using mutation, crossover, and selection operators at each iteration. Instead of implementing traditional crossover and mutation operators, DE applies a linear combination of several candidate solutions selected randomly to produce a new solution. Finally, DE returns the best candidate solution in the current population when the stop condition is fulfilled.

If for each $j \in 1, \ldots, |x^i|$, x_j^{min} and x_j^{max} are the minimum and the maximum values of the j-th parameter, respectively, the j-th value of x^i in the initial population is calculated as follows:

$$x_j^i = x_j^{min} + r(x_j^{max} - x_j^{min}), \tag{4}$$

where $r \in [0, 1]$ is a uniformly distributed random number.

Furthermore, the mutation, crossover, and selection operators are defined as follows:

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

$$v^i = x^{r_1} + \mathrm{F}\big(x^{r_2} - x^{r_3}\big). \tag{5}$$

  Eq. 5 is related with the DE/rand/1 variant defined in [9].
- **Crossover:** The mutated vector is recombined with the target vector to build the trial vector u^i. For each $j \in \big\{1, \ldots, |x^i|\big\}$, either x_j^i or v_j^i 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 gets at least one value from v^i, as follows:

$$u_j^i = \begin{cases} v_j^i & \text{if } r \leq \text{CR or } j = l, \\ x_j^i & \text{otherwise.} \end{cases} \tag{6}$$

- **Selection:** A one-to-one tournament is applied to determine which vector, between x^i and u^i, is selected as a member of the new population.

An advantage of DE is that it uses a few control parameters: a crossover rate Cr, a mutation scale factor F, and a population size NP.

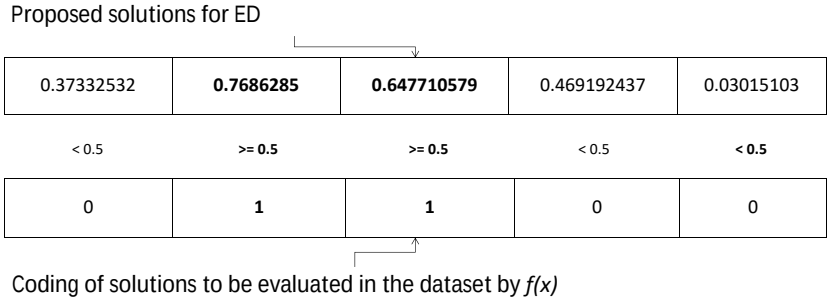

Fig. 1. Coding the candidate solution proposed by DE.

3 Implementation

DE evolves a population of randomly generated real-valued vectors. The evolutionary process is guided by a fitness function $f(x)$ determining the quality value of each individual in the population. In this work $f(x)$ is computed using several metrics searching for a maximum correlation between dataset attributes. The two crucial elements determining the success of the evolutionary process to create representative association rules to identify positive BV cases are defined as follows:

3.1 DE/rand/1/bin Version Implemented

DE/rand/1/bin is a classic DE variant, where *rand* indicates that base vectors are randomly chosen, 1 indicates that only one vector difference is used to form the mutated population, and the term *bin* (from binomial distribution) points out that uniform crossover is employed during the formation of the trial population.

3.2 Solution Encoding Scheme

Since the dataset's attributes are categorical, a scheme is required where the vector of real numbers can represent the selection or not of some attribute. The threshold-based scheme is the traditional approach to represent the selected attributes from a dataset: If the i-th parameter value is greater than 0.5, then the i-th attribute is chosen to build an association rule; otherwise, this attribute is discarded [10]. This scheme is used in this work, and Fig. 1 shows an example of this mapping scheme.

3.3 Fitness Function Definition

In research health, qualitative or dichotomous attributes are frequently used. In this research, we propose studying associations between bacteria that trigger

BV+. There are different statistical models to analyze qualitative data based on contingency tables. Thus, the objective function is made up of three statistical tests: Chi-squared, Yates and Fisher. These metrics are defined as follows:

- **Chi-squared Test:**

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{7}$$

  where O_i is the observed value (the number of cases observed in a cell contingency table). E_i is the expected value, the number of expected cases in each cell of the contingency table.
- **Yates's formula:**

$$X^2(Yates) = \frac{n\left(\mid ad - bc \mid -\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}. \tag{8}$$

- **Fisher's test:**

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}. \tag{9}$$

  In Eqs. 8 and 9, a, b, c, d are the expected frequencies in a 2 x 2 contingency table. $(a+b)$ is the sum of the two frequencies in row 1 and $(c+d)$ the sum of the frequencies in row 2. Similarly $(a+c)$ is the sum of the frequencies in column 1 and $(b+d)$ the sum of the frequencies in column 2. The total sum of all frequencies is given by n [15].

The fitness function is defined as follows:

$$f(x) = \begin{cases} p & \text{if expected frecuency} < 3 \\ X^2(Yates) & \text{if expected frecuency} > 3 \text{ and} < 5 \\ X^2 & \text{otherwise.} \end{cases} \tag{10}$$

3.4 Contingency Table

The contingency table plays an important role in the process of discovering the association between the attributes to be analyzed. The contingency table in Fig. 2, is composed of rows and columns, in the cells are recorded the absolute or relative frequencies where it is possible to analyze the correlation between two variables from the calculations that can be performed. From the set of solutions that integrate the initial population proposed by the evolutionary algorithm, they are coded in a binary format described in 3.2 , to select the combination of attributes of the study data set and placed in the contingency table.

To determine the existence of a correlation between the selected attributes, the fitness function *f(x)* is applied to the contingency table, and the results obtained will determine the existence of correlation at the 95% confidence level. The fitness function will guide the evolutionary process to determine the quality of each individual, for our case study, the fitness function is composed of three statistical tests under these restrictions:

Variable of study

VB+	1	1	0	1	1	1	1	1	0	0
Bacteria	1	1	0	0	0	0	1	1	1	1

	B	B'	
A	n(AB)	n(AB')	n(A)
A'	n(A´B)	n(A'B')	n(A')
	n(B)	n(B')	N

→

	Vaginosis		
Bacteria	VB+	VB-	Total
Present	4	2	6
Absent	3	1	4
Total	7	3	10

Fig. 2. The content of each cell is the number of frequencies that simultaneously satisfy the two conditions as $n(AB)$ [5] and N is total number of cases.

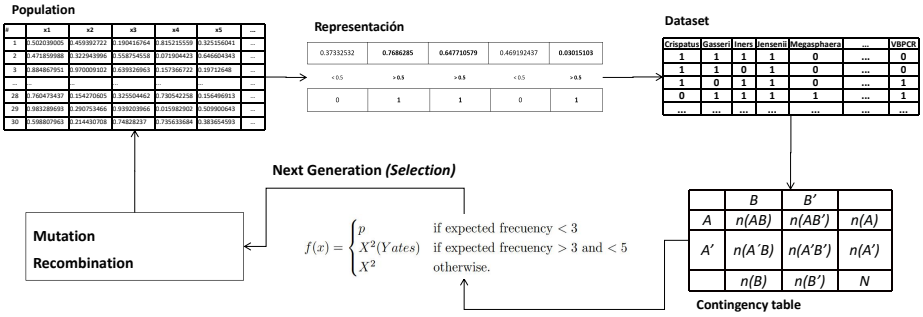

Fig. 3. Experimental design to build Association Rules using Differential Evolution.

– If 25% of the expected frequencies is >3 and <5, the Yates test is applied.
– If 25% of the expected frequencies is < 3, Fisher's exact text is used.

After calculating the existence of dependence between attributes, the strength of the association is measured with Pearson's *Phi* correlation function. If the result is close to one, it indicates a strong association, while if the results are close to zero, the association is very poor or non-existent.

4 Experimental Design

The following diagram details the steps to create rules with differential evolution.

The initial population is defined with random numbers within the continuous space. The solutions composing the population are evaluated with the objective function *f(x)* composed of the three metrics. To evaluate the solution, it is coded to a binary format to select the combination of attributes of the dataset corresponding to the coding. Subsequently, the result of the selected set of attributes from the dataset is placed in a contingency table and the function *f(x)* is applied under the criteria described in Section 3.4.

Table 1. Chi-squared test result.

No.	Rule	$f(x)$	Phi
1	Jensenni, Megasphaera, Atopobium $\Longrightarrow$ VB+	1.71×10^{-17}	0.625
2	Gasseri, Megasphaera, Atopobium $\Longrightarrow$ VB+	4.86×10^{-19}	0.655
3	Iner, Gardnerella $\Longrightarrow$ VB+	6.1×10^{-9}	0.427

Table 2. Yates's correction test result.

No.	Rule	$f(x)$	Phi
1	Iners, Jensenii, Megasphaera $\Longrightarrow$ VB+	7.32×10^{-10}	0.473
2	Gasseri, Iners, Jensenii, Atopobium, Gardnerella $\Longrightarrow$ VB+	2.01×10^{-9}	0.463
3	Gasseri, Iners, MH $\Longrightarrow$ VB+	1.39×10^{-4}	0.303

Table 3. Fisher's exact test result.

No.	Rule	$f(x)$	Phi
1	Crispatus, Gardnerella, MH, MG $\Longrightarrow$ VB+	5.0×10^{-6}	0.366
2	Gasseri, Iners, Megasphaera, Atopobium, MG, UP $\Longrightarrow$ VB+	2.1×10^{-5}	0.344
3	Crispatus, Gasseri, Iners, Megasphaera, MG, UP $\Longrightarrow$ VB+	8.8×10^{-5}	0.321

Finally, the solution with the best fitness is passed to the next generation, this process will continue until the stop condition is completed.

5 Results

The results obtained by minimizing the fitness function *f(x)* described in Section 3.3 are summarized below. *f(x)* evaluates the results under the p-value criterion, the null hypothesis is either approved or rejected with a confidence level of 95%. The null hypothesis H_0 is rejected if the p-value is less than 0.05 meaning that there is an association. Under this criterion, we minimize the fitness function *f(x)* and measure the strength of the association with the Pearson correlation coefficient Phi, of the analyzing variables.

The results are detalled in Table 1, 2 y 3. In this tables, the rules with the highest statistical value are depicted. The DE parameters are adjusted according to the recommendation suggested in the existing literature [13], as follows:

1. $F \in [0.5, 1.0]$
2. $CR \in [0.8, 1.0]$

and a population of 30 individuals.

The evolutionary process is stopped when 30 generations are reached.

In the results of the Chi-square test in Table 1, rule one indicates that *Jensenni, Megasphaera and Atopobium* bacteria trigger Bacterial Vaginosis with association strength Phi = 0.625. In the results of the Yates test in Table 2, rule one formed by bacteria *Iners, Jensenii, Megasphaera* indicated positive

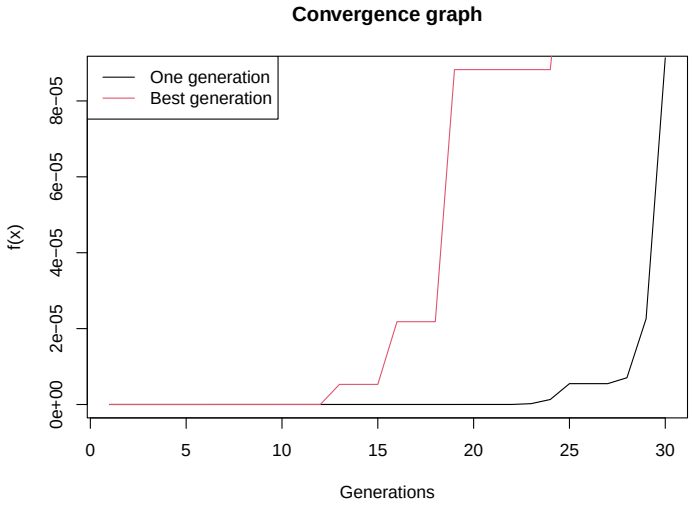

Fig. 4. Convergence of the evolution in the solutions of the fitness function.

for bacterial vaginosis with an association Phi= 0.473. Finally, in Table 3 the Fisher's test results for rule one *Crispatus, Gardnerella, Mycoplasma Hominis and Mycoplasma Genitalium* indicated positive for Baginosis with an association Phi = 0.366.

From the results, the rules that may be involved in the development of the infection are summarized under the criterion of highest statistical significance. The explanation of the cause of infection is so far complicated by the large number of bacteria that coexist in the vagina of women.

Furthermore, Figure 4 shows the fitness function convergence behavior to the best individual in the population.

6 Conclusion

BV is a health problem and should be treated early to avoid future risks in women. In this work, we implemented the differential evolution algorithm as a tool to discover strongly related association rules and avoid generating low-quality rules. The proposed approach is used to avoid the combinatorial explosion present when the Apriori algorithm is applied.

In this analysis, several tests were performed by modifying the parameters according to the limits suggested by the experts. The observed results allow knowing the bacteria associated with bacterial vaginosis with statistical values. This is an on-going research, and currently a function with a set of biological constraints is being implemented to guide the evolutionary search towards the

optimal result. Other bacteria causing the Bacterial Vaginosis infection will also be expected to be discovered.

References

1. Coello, C.: Introducción a la computación evolutiva (Notas de curso). CINVESTAV-IPN, México, DF (2004)
2. Ceglar, A., Roddick, J.: Association mining. ACM Computing Surveys. 38, 5 (2006)
3. Dongre, J., Prajapati, G.L., Tokekar, S.V.: The role of apriori algorithm for finding the association rules in data mining. In: ICICT 2014, pp. 657—660 (2014)
4. García, P.: Vaginosis bacteriana. Revista Peruana De Ginecología Y Obstetricia, 53, 167–171 (2007)
5. Geng, L., Hamilton, H.: Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR), 38, 9 (2006)
6. Hernández Orallo, J., et al.: Introducción a la Minería de Datos. Biblioteca Hernán Malo González (2004)
7. Kotsiantis, S., Kanellopoulos, D.: Association rules mining: A recent overview. GESTS Int. Trans. on Computer Science & Engineering, 32, 71–82 (2006)
8. Pérez-Gómez, J.F., Canul-Reich, J., Hernández-Torruco, J., Hernández- Ocaña, B.: Predictor selection for bacterial vaginosis diagnosis using decision tree and relief algorithms. Applied Sciences 10(9), 3291 (2020)
9. Price, K., Storn, R., Lampinen, J.: Differential evolution: a practical approach to global optimization Springer (2006)
10. Rivera-López, R., Mezura-Montes, E., Canul-Reich, J., Cruz-Chávez, M. A.: A permutational-based differential evolution algorithm for feature subset selection. Pattern Recognition Letters, 133, 86–93 (2020)
11. Saldaña, M.: La prueba chi-cuadrado o ji-cuadrado (2). Revista Enfermería Del Trabajo. 1, 31–38 (2011)
12. Sanchez-Garcia, E., et al.: Molecular epidemiology of bacterial vaginosis and its association with genital micro-organisms in asymptomatic women. Journal Of Medical Microbiology, 68, 1373–1382 (2019)
13. Storn, R., Price, K.: Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal Of Global Optimization, 11, 341–359 (1997)
14. Wu, X., et al.: Top 10 algorithms in data mining. Knowledge & Information Systems, 14, 1–37 (2008)
15. Zar, J.: A fast and efficient algorithm for the Fisher exact test. Behavior Research Methods, Instruments, & Computers, 19, 413–414 (1987)