ISSN 1870-4069

Selection of a Fixed-Length Set of Biologically-Constrained Association Rules for Bacterial Vaginosis Diagnosis

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

[1] Universidad Juárez Autónoma de Tabasco,
División Académica de Ciencias y Tecnologías de la Información,
Mexico

[2] Instituto Tecnológico de Veracruz,
División en Ciencias de la Salud,
Mexico

[3] Universidad Veracruzana,
Instituto de Investigaciones en Inteligencia Artificial,
Mexico

[4] Universidad Juárez Autónoma de Tabasco,
División Académica Multidisciplinaria de Comalcalco,
Mexico

mcsalvadorg@gmail.com, {juana.canul, erick.delacruz}@ujat.mx,
rrivera@itver.edu.mx, emezura@uv.mx

Abstract. This paper describes a Differential-Evolution-based approach for selecting a reduced subset of association rules previously generated by the Apriori algorithm. The selected rules are those with biological significance for the diagnosis of Bacterial Vaginosis. We use integer-based vectors as population individuals of the evolutionary algorithm and a combination of various rule metrics to define the fitness function. The experimental results indicate that the DE/best/1/bin variant performs better than the DE/rand/1/bin variant and that the approach reaches the expected results.

Keywords: Differential evolution, association rules, bacterial vaginosis.

1 Introduction

Bacterial Vaginosis is the most common of the vaginal diseases in women of reproductive age. It is associated with several severe health conditions such as preterm delivery, post-abortion infection, pelvic inflammatory disease, and sexually transmitted diseases [10]. As in other fields of knowledge, machine learning techniques have been used to detect this condition [2].

On the other hand, Association Rule Mining is an important topic in data mining used to identify the relationships strongly associated among itemsets in a dataset [15]. In

Table 1. Antecedent itemset values used in the experimental study.

Dataset feature	Values	Description
Cristpatus	1	crispatusA
Gasseri	2	crispatusB
	1	gasseriA
Iners	2	gasseriB
	1	inersA
Jensenii	2	inersB
	1	jenseniA
Megasphaera	2	jenseniB
	1	megasphaeraP
Atopobium	2	megasphaeraN
	1	atopobiumP
Gardnerella	2	atopobiumN
	1	gardnerellaP
	2	gardnerellaN

ID of rule 1	ID of rule 2	ID of rule 3	ID of rule 4	...	ID of rule N
34	125	200	6		29

Fig. 1. Encoding scheme to select N association rules.

the related specialized literature, we found that several computational techniques, such as Simulated Annealing [6], Genetic Programming [9], Differential Evolution [14], and Genetic Algorithms [8], have been applied to generate and optimize Association Rules for a wide range of real applications. In particular, the Differential Evolution algorithm has proven its effectiveness in optimizing machine learning models.

To the best of our knowledge, no study has been found in the existing literature that applies Association Rule Mining and Differential Evolution to select biologically meaningful rules for the diagnosis of bacterial vaginosis infection. This work addresses the adaptation of the Differential Evolution algorithm to determine association rules using biological constraints in cases of Bacterial Vaginosis Positive (BV+).

2 Materials and Methods

For this study, a dataset with 17 features with medical information of 201 sexually active women aged 18 to 50 who underwent their annual gynecological inspection routine at the Laboratory of Research in Metabolic and Infectious Diseases, Universidad Juarez Autonoma de Tabasco is used [12]. According to our interest, we considered the records with a positive result for bacterial vaginosis only.

After this selection, 51 records remained, with the variables representing the Crispatus, Gasseri, Inners, and Jensenii lactobacillus, and the Megasphaera, Atopobium, and Gardnerella bacteria. An association rule has the form $X \rightarrow Y$, where X is the rule's antecedent, and Y is its consequent [1].

The metrics most commonly used for the validation of the obtained rules are the following [7]:

Table 2. Parameters values.

Parameter	Value	Parameter	Value
F (Scale factor)	0.9	CR (Crossover rate)	0.5
NP (Population size)	20	MAX_GEN (Number of generations)	30

– **Support:** The frequency count of a rule.

– **Confidence:** The probability that the elements in the consequent are in the antecedent.

– **Coverage:** The frequency with which the rule antecedent appears.

– **Lift:** It compares the expected frequency of a rule with the expected frequency at random.

– **Confidence-boost:** The relationship between the confidence of rules that have the same consequent but different elements in the antecedent.

As part of association rule mining, the dataset is processed for the Apriori algorithm, one of the most widely used algorithms for pattern discovery using frequent itemsets to generate association rules [5]. 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. 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. 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.

Since the information in the dataset is not numerical, the DE algorithm must be adapted to generate optimized results. We encode the values with integer-valued vectors, implying that the algorithm's operators must be modified to create only feasible solutions. Another critical element is the definition of the objective function, which must correctly guide the evolutionary process. In the present work, metrics used with association rules should be considered, as well as those defining the biological significance levels for the problem under study.

3 Experimental Study

The experimental study includes three stages. First, the meaning of the values that each attribute can take on are defined as indicated in Table 1. In total, there are 51 records in the dataset.

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	36.8267	37.6078	16	37.1960	36.8431
2	37.9019	37.6666	17	37.7450	38.0588
3	37.1372	37.4919	18	37.4509	38.1372
4	36.9411	37.5294	19	37.1372	36.8235
5	37.2549	37.5686	20	36.9215	36.5098
6	36.5882	37.6666	21	37.3333	36.7058
7	36.7254	37.0980	22	37.0588	36.6862
8	37.5686	37.1568	23	38.0588	36.6862
9	36.6470	**38.3333**	24	**38.8039**	36.9019
10	37.5490	37.1764	25	36.8627	37.0000
11	36.5098	37.8431	26	37.7647	37.2941
12	36.6666	37.7450	27	38.1372	37.1372
13	36.8039	37.5098	28	37.4117	36.6470
14	37.3921	37.3921	29	37.0588	37.0000
15	36.8039	37.3921	30	38.2549	36.9019

Next, the Apriori algorithm[5] is applied and 332 association rules are generated, all for cases of BV+. Each rule ends up with one or more features in the antecedent part, and the value of BV+ is set as the consequent since these are the cases of interest in this work. Finally, the DE algorithm is used to find a reduced set of association rules, based on their biological significance.

3.1 Implementation of The Differential Evolution (DE) Algorithm

Three elements are first defined to implement the DE algorithm: the individuals' encoding scheme, the fitness function, and the variation operators.

1. **Encoding scheme:** An individual of the population is a subset of N association rules each identified with an ID number. Fig. 1 shows an example of this codification. In this work, the value of N is set to 6 since in [4] authors obtained five rules with biological significance which were determined by a human expert, so $N = 6$ rules ensures the algorithm will find this minimal set of rules.
2. **Fitness function:** Each i-th individual in the population is evaluated to define the fitness value. In this work, the fitness function $f(x_i)$ is the sum of the M metrics of the association rules encoded on the individual as follows:

$$f(x_i) = \sum_{j=1}^{N} \sum_{k=1}^{M} m_{j,k}, \tag{1}$$

where N is the number of desired association rules, M is the number of metrics involved to define the solution quality, and $m_{j,k}$ is the k-th metric computed for the j-th rule.

[5] In this work, the arules R package is used to create the association rules (cran.r-project.org/web/packages/arules/index.html).

Table 4. Statistical values of the experimental study.

Statistical measure	**rand/1/bin**	**best/1/bin**
Best value	38.8039	38.3333
Mean	37.2849	37.2836
Median	37.1666	37.2352
Standard deviation	0.5575	0.4773
Worst value	36.5098	36.5098
Best test number	24	9
Median test number	16	26

Seven metrics are used in the fitness function: support, confidence, coverage, lift, confidence boost, frequency of positive bacteria in the rules, and the occurrences of high values of lactobacillus iners. The first five metrics are previously described in Section 2.

The other two metrics are used to determine the presence of some bacteria, and lactobacillus [4]. These metrics are included to define the biological significance of the association rules in this sense the higher results of the addition of the metrics have higher significance.

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

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

   Eq. 2 is related with the DE/rand/1 variant defined in [11]. Other commonly used variant is known as DE/best/1, where the best individual in the population $\mathbf{x}^{best}$ is combined with two random chosen individuals of the current population, as follows:

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

   – **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^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 gets at least one value from v^i, as follows:

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

In the Eqs. 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 332.

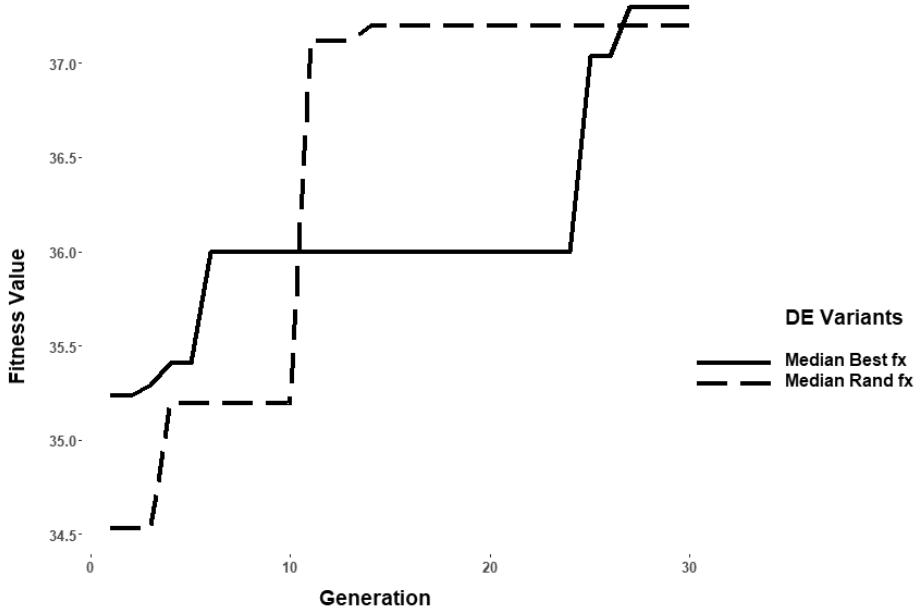

Fig. 2. Convergence plot for the median values of the two DE variants.

3.2 Algorithm Parameters

It's well known that the performance of the DE algorithm is affected by the values of its parameters: F, CR, and NP [3]. The parameter values used in this work are based on those commonly used in the existing literature [11]. Since this experimental study is a work in progress, no parameter tuning process has been carried out.

4 Results

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 fitness value for the rand/1/bin version is 38.8039 on test number 24 and for the best/1/bin version is 38.3333 on test 9. The best fitness values are highlighted in bold, and the best median value of each variant is underlined.

The statistics comparison for each variant is shown in Table 4, and Fig. 2 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 statistical test, the calculated p-value is 0.4065, indicating that the two variants have the same behavior.

Table 5 shows the rules encoded by the best individuals of each variant. 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.

Table 5. Reduced set of association rules selected by the two DE variants.

ID	Association rule
Variant: best/1/bin	
306	{atopobiumP, crispatusA, gardnerellaP, gasseriA, jenseniA} → {VB+}
139	{atopobiumP, gardnerellaP, inersA, megasphaeraP} → {VB+}
224	{atopobiumP, crispatusA, gardnerellaP, megasphaeraP} → {VB+}
328	{atopobiumP, crispatusA, gardnerellaP, gasseriA, inersA, jenseniA} → {VB+}
268	{atopobiumP, crispatusA, gardnerellaP, inersA, megasphaeraP} → {VB+}
210	{atopobiumP, inersA, jenseniA, megasphaeraP} → {VB+}
Variant: rand/1/bin	
212	{atopobiumP, gasseriA, inersA, megasphaeraP} → {VB+}
209	{atopobiumP, gardnerellaP, gasseriA, inersA, jenseniA} → {VB+}
103	{atopobiumP, gardnerellaP, inersA} → {VB+}
124	{atopobiumP, gardnerellaP, jenseniA} → {VB+}
245	{atopobiumP, gardnerellaP, inersA, jenseniA, megasphaeraP} → {VB+}
296	{atopobiumP, crispatusA, gardnerellaP, inersA, jenseniA} → {VB+}

Likewise, all resulting rules comply with the biological significance requirement of having at least two bacteria present [12]. Biological significance adds weight to rules that carry bacteria and, at the same time, show high levels of lactobacillus iners.

5 Conclusions and Future Work

The experimental results shown have been validated by an expert biologist, who observed that multiple combinations of present bacteria (indicated with the letter P) and absent lactobacillus (indicated with the letter A) could lead to the disease appearance in the resulting rules.

Thus, the algorithm's behavior using the coding scheme and the fitness function lead to rules with biological significance. Furthermore, our results show that using DE to select association rules created with Apriori is a promising approach to identifying a high-quality and compact rule set for BV diagnosis.

In future work, it is crucial to continue with the validation of the rules by a human expert to corroborate their feasibility. Another point is to add penalties in the fitness function for antecedent itemsets unlikely to occur when there exists a positive consequent.

Additionally, new encoding schemes will be studied so that the number of selected rules is not previously defined. In this sense, it is also proposed to test with other DE variants and try different techniques for the algorithm-parameter-tuning to improve the algorithm performance.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules Between Sets of Items in Large Databases. In: Proceedings of the International Conference on Management of Data, pp. 207–216 (1993). DOI: 10.1145/170035.170072.
2. Baker, Y. S., Agrawal, R., Foster, J. A., Beck, D., Dozier, G.: Applying Machine Learning Techniques in Detecting Bacterial Vaginosis. In: The International Conference on Machine Learning and Computing, vol. 1, pp. 241–246 (2014). DOI: 10.1109/ICMLC.2014.7009123.
3. Das, S., Suganthan, P. N.: Differential evolution: A Survey of The State-of-the-art. In: IEEE Transactions on Evolutionary Computation, vol. 15, pp. 4–31 (2011). DOI: 10.1109/TEVC.2010.2059031.
4. De la Cruz, F., Canul-Reich, J.: Reglas de asociacion para estudiar patrones bacterianos ´ involucrados en el desarrollo de vaginosis bacteriana. Komputer Sapiens, vol. 14, no. 2 (2022)
5. Dongre, J., Prajapati, G. L., Tokekar, S. V.: The Role of Apriori Algorithm for Finding the Association Rules in Data Mining. In: International Conference on Information and Computer Technologies, pp. 657–660 (2014). DOI: 10.1109/ICICICT.2014.6781357.
6. Guo, H., Li, Y., Liu, X., Li, Y., Sun, H.: An Enhanced Self-adaptive Differential Evolution Based on Simulated Annealing for Rule Extraction and its Application in Recognizing Oil Rservoir. Applied Intelligence, vol. 44, no. 2, pp. 414–436 (2016). DOI: 10.1007/s10489-015-0702-x.
7. Hahsler, M.: A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules. Southern Methodist University (2015)
8. Leske, M., Bottacini, F., Afli, H., Andrade, B. G. N.: BiGAMi: Bio-objective Genetic Algorithm Fitness Function for Feature Selection on Microbiome Datasets. Methods and Protocols, vol. 5, no. 3 (2022). DOI: 10.3390/mps5030042.
9. Luna-Romera, J. M., Reyes, O., del Jesús-Díaz, M. J., Soto, S. V.: Reglas de asociacionnen datos multi-instancia mediante programación genética gramatical. In: Congreso de la Asociacion Española de Inteligencia Artificial: Avances en Inteligencia Artificial, pp. 815–820 (2018)
10. 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, vol. 10, no. 9, pp. 3291 (2020). DOI: 10.3390/app10093291.
11. Price, K., Storn, R. M., Lampinen, J. A.: Differential evolution: A Practical Approach to Global Optimization (2006). DOI: 10.1007/3-540-31306-0.
12. 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. Journal of Medical Microbiology, vol. 68, no. 9, pp. 1373–1382 (2019). DOI: 10.1099/jmm.0.001044
13. Storn, R., Price, K.: Differential Evolution – a Simple and Efficient Heuristic for Global Optimization Over Continuous Spaces. Global Optimization, vol. 11, no. 4, pp. 341–359 (1997). DOI: 10.1023/A:1008202821328.
14. Wang, C., Liu, Y., Zhang, Q., Guo, H., Liang, X., Chen, Y., Xu, M., Wei, Y.: Association Rule Mining Based Parameter Adaptive Strategy for Differential Evolution Algorithms. Expert Systems with Applications, vol. 123, pp. 54–69 (2019). DOI: 10.1016/j.eswa.2019.01.035.
15. Zhang, C., Zhang, S.: Association Rule Mining: Models and Algorithms (2002). DOI: 10.1007/ 3-540-46027-6.