Notifications

[IJCOPI] Editor Decision

2021-05-11 06:29 PM

Jesus Francisco Pérez-Gómez, Juana Canul-Reich, Erick De La Cruz-Hernandez:

We have reached a decision regarding your submission to International Journal of Combinatorial Optimization Problems and Informatics, "An Enhanced Method for Diagnosis of Bacterial Vaginosis based on Support Vector Machines with Linear Kernel".

Our decision is to: Accept Submission

Please send us as soon as possible the source files of your manuscript. At this time we have only the pdf. We need the *.doc source to start the stage of copyediting.

Special Issue Guest Editors

International Journal of Combinatorial Optimization Problems and Informatics

International Journal of Combinatorial Optimization Problems and Informatics (http://www.ijcopi.org)

ISSN:

E-ISSN: 2007-1558

Temática: Engineering; Combinatorial Optimization Problems

Institución Editora: Editorial Académica Dragón Azteca

Editor (es): Dr. Jorge A. Ruíz Vanoye

Refine Your Search Results

ISSN: 2007-1558

Search

Sort By: Title (A-Z)

Search Results

Found 1 results (Page 1)

Share These Results

Exact Match Found

INTERNATIONAL JOURNAL OF COMBINATORIAL OPTIMIZATION PROBLEMS AND INFORMATICS

Publisher: INT JOURNAL COMBINATORIAL OPTIMIZATION PROBLEMS & INFORMATICS , ALTAIR 14 COL LOMAS JIUTEPEC, JIUTEPEC, MEXICO, MORELES, 62550

ISSN / eISSN: 2007-1558

Web of Science Core Collection: Emerging Sources Citation Index

Share This Journal

View profile page

** Requires free login.*

Filters

Clear All

Web of Science Coverage

Open Access

Category

Country / Region

Language

Frequency

Journal Citation Reports

Actualizado en
Junio 17, 2021

International Journal of Combinatorial Optimization Problems and Informatics

Búsqueda

Introduzca una o más palabras | Todos los índices | En la Revista | Búsqueda

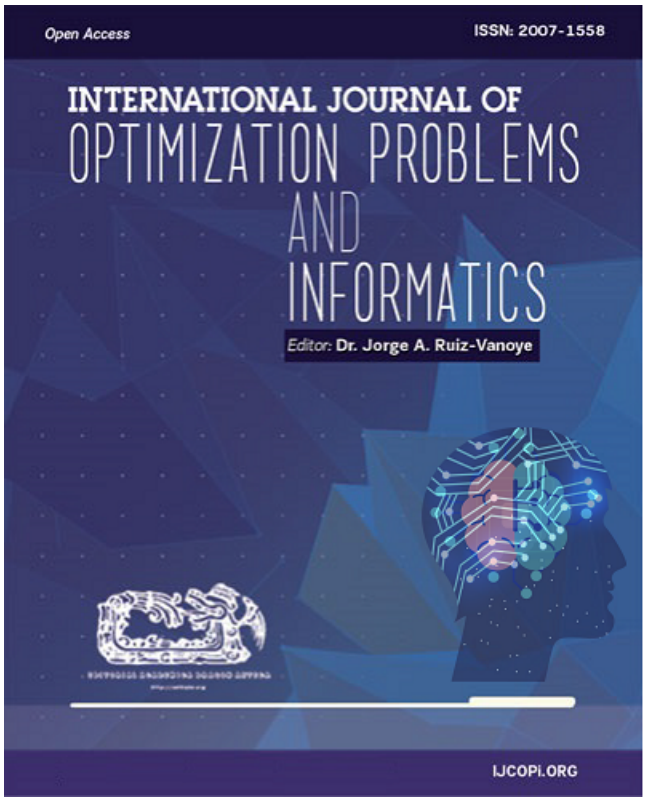
Publicación de
EDITADA (EDITorial Academica Dragon Azteca, S. de R.L. de C.V. or Aztec Dragon Academic Publishing)
versión On-line ISSN 2007-1558

Misión

El objetivo de la revista IJCOPI es proveer información en áreas relacionadas con la Optimización Combinatoria de Problemas (COP□s) y la informática para crear una colección de documentos, reportes técnicos, noticias, tutoriales, reseñas de libros, casos de estudio, entre otros, dirigidos a estudiantes e investigadores del área. Se utiliza un proceso de revisión por pares que asegura la validez y relevancia de los contenidos. La revista IJCOPI es única en este campo y constituye una fuente indispensable de guía y conocimiento a estudiantes e investigadores en la academia y la industria relacionados con la optimización combinatoria de problemas y la informática. IJCOPI es una revista de acceso abierto que permite el depósito de los artículos publicados en ella en repositorios temáticos o institucionales (Green journal). Publicación cuatrimestral.

Indexing/abstracting

Indexing and abstracting:



Published: 2021-09-11

Articles

Editorial: A Brief Panorama of Artificial Intelligence in Mexico

Osslan Osiris Vergara Villegas, Manuel Nandayapa, Juan Humberto Sossa Azuela, Félix Agustín Castro Espinoza

PDF

A Comparative Study of Dendrite Neural Networks for Pattern Classification

98-108

PDF

An Enhanced Method for Diagnosis of Bacterial Vaginosis based on Support Vector Machines with Linear Kernel

Jesus Francisco Pérez-Gómez, Juana Canul-Reich, Erick De La Cruz-Hernandez

109-121

PDF

Information

For Readers

For Authors

For Librarians

An Enhanced Method for Diagnosis of Bacterial Vaginosis based on Support Vector Machines with Linear Kernel

Jesus Francisco Pérez-Gómez
Universidad Juarez Autonoma de Tabasco. Division Académica de Ciencias y Tecnologías de la Información

Juana Canul-Reich
Autonomous Juárez University of Tabasco.

Erick De La Cruz-Hernandez
Autonomous Juárez University of Tabasco

Keywords: Classification, bacterial vaginosis diagnosis, kernel-based method, reduction dimension, svm, feature selection

Abstract

Bacterial Vaginosis (BV) is a pathological condition that causes complications in women's health. Efforts to characterize it have failed to reveal a BV etiology. In this work, the Support Vector Machine (SVM) is used as base classifier in three different scenarios to identify between classes of VB. The first scenario uses the entire feature set in the dataset. The second scenario uses two sub-datasets created with the features in two rankings obtained from previous work. The third scenario uses one feature at a time to create classifiers. Performance measures in each are given. The dataset used is a real vaginal microbiology test of 201 women from Tabasco, Mexico. Results show SVM surprisingly obtained 100% accuracy in a classifier made of a single feature. This research is a first effort to lay the groundwork for computer-based BV diagnosis as advice.

PDF

Published
2021-09-11

How to Cite
Pérez-Gómez, J. F., Canul-Reich, J., & De La Cruz-Hernandez, E. (2021). An Enhanced Method for Diagnosis of Bacterial Vaginosis based on Support Vector Machines with Linear Kernel. *International Journal of Combinatorial Optimization Problems and Informatics*, *12*(3), 109-121. Retrieved from

Information
For Readers
For Authors
For Librarians

International Journal of Combinatorial Optimization Problems and Informatics, 12(3), Sep-Dec 2021, 109-121. ISSN: 2007-1558.

www.editada.org

An Enhanced Method for Diagnosis of Bacterial Vaginosis based on Support Vector Machines with Linear Kernel.

Jesús Francisco Pérez-Gómez[1], *Juana Canul-Reich*[1*], *Erick De La Cruz-Hernández*[2]
[1]Universidad Juárez Autónoma de Tabasco, División Académica de Ciencias y Tecnologías de la Información. Carretera Cunduacán-Jalpa KM. 1 Col. La Esmeralda CP. 86690. Cunduacán, Tabasco, México.
[2]Universidad Juárez Autónoma de Tabasco, División Académica Multidisciplinaria de Comalcalco. Ranchería Sur 4ta. sección. Comalcalco, Tabasco, México. Comalcalco, Tabasco, México.
*Corresponding author: juana.canul@ujat.mx

Abstract. Bacterial Vaginosis (BV) is a pathological condition that causes complications in women's health. Efforts to characterize microorganisms associated to BV etiology have failed. In this work, the Support Vector Machine (SVM) is used as base classifier in three scenarios to identify between classes of BV. The first scenario uses the entire feature set in the dataset. The second scenario uses two sub-datasets created with the features in two rankings obtained from previous work. The third scenario uses one feature at a time to create classifiers. Performance measures in each are given. The dataset used is a real vaginal microbiology test of 201 women from Tabasco, Mexico. Results show that SVM surprisingly obtained 100% accuracy in a classifier made of a single feature. This research is a first effort to lay the groundwork for computer-based BV diagnosis as advice.
Keywords: classification, feature selection, bacterial vaginosis diagnosis, kernel-based method, reduction dimension, SVM.

Article Info
Received Jan 2, 2021
Accepted May 13, 2021

1 Introduction

Bacterial vaginosis (BV) is a dysbiosis commonly detected in sexually active women. During this condition, clinical manifestations such as abnormal vaginal discharge in color (gray or green) and a fishy smell can be observed [1]. Women with this infection have a 60% higher risk of contracting Human Immunodeficiency Virus (HIV) and the chances of transmitting HIV to uninfected partner are increased by 30% [2]. A large percentage of patients are asymptomatic, which further complicates the diagnosis [3].

Efforts to characterize BV using microscopic assays, microbiological culture, and sequenced-based methods have all failed to reveal an etiology that can be consistently documented in all women with BV [4]. Among the classical procedures for the diagnosis of BV are the Amsel criteria and the Nugent score [3,5]. Another technique most recently employed for the diagnosis of BV is real-time Polymerase Chain Reaction, also known as Quantitative PCR (qPCR) [1,5]. Moreover, most of these procedures are unreliable [6], and others offer poor information due to the complex polymicrobial nature of the BV [7].

In this work, experiments with the use of Support Vector Machine (SVM) to diagnose BV under three scenarios are reported. The first scenario uses SVM with the entire set of features in a BV dataset. The second scenario consists of experiments for evaluating SVM on two sub-datasets created with selected features based on two feature rankings obtained from previous work [8]. In essence, these sub-datasets are made of the fifteen features identified as most relevant according to different feature selection methods investigated in [8]. The third scenario consists of experiments with SVM for BV diagnosis when using individually one feature at a time to create predictive models. The 10-fold Cross-Validation (10FCV)

technique was used as a validation scheme in the experiments of this work as similarly was used in [8][9][10]. Results obtained from this work show that SVM is 100% accurate using one feature only to identify the BV. The dataset used in all experiments consists of molecular diagnosis of bacterial vaginosis. It contains 201 instances and 57 features. The samples correspond to women from Comalcalco, Tabasco, México, and they were obtained and analyzed at the Research Laboratory in Infectious and Metabolic Diseases of the Comalcalco Multidisciplinary Academic Division [1].

This document is organized as follows: Section 2 describes some works related to methods and techniques from the machine learning area focused on the study of bacterial vaginosis. Section 3 details the dataset used and the machine learning methods implemented in this research. Section 4 explains in detail the experimental phases of the investigation. Section 5 shows the results obtained in the experiments developed. Finally, Section 6 provides the general conclusions of this work.

2 Related work

In this section, some studies related to methods and techniques from the machine learning area applied to bacterial vaginosis are described. Some of those works have motivated the use of the algorithms proposed in the experimental development of this research.

In the work of Pérez-Gómez's [11] classification algorithms such as SVM and Logistic Regression (LR) and feature selection methods such as decision trees and Relief were used to determine the combination of techniques with the highest predictive values in the bacterial vaginosis diagnosis. In that paper, the dataset used consists of clinical and biological information about vaginal microorganisms of patients from two universities in Baltimore and Atlanta, United States of America [12]. The experiments consisted of 30 runs of the algorithms under a cross-validation scheme. Performance measures such as accuracy, balanced accuracy, sensitivity, and specificity were obtained for comparison purposes. Based on the obtained results was determined that the SVM classifier algorithm with the use of only the 15 most relevant features identified by decision trees obtained up to 100% in all predictive values calculated. Even, in experiments with the use of the entire set of features, SVM obtained an accuracy of more than 95%.

In the research of Yolanda Baker [13], it was proposed to find the most relevant features of BV, and applying some classification methods to diagnose it. In the experiments implemented with the WEKA software twenty feature selection methods and nine classification methods were applied. Measures such as accuracy, recall, and the number of features reduced were some of the performance metrics reported in this research. The dataset contains 1601 instances and 418 features. It consists of a combination of clinical information and Amsel criteria test results [14]. Through the experiments performed, it was found that the Functional Trees as a classification method and WrapperSubSetEval as a feature selection method obtained accuracies over 97%.

In the paper of Beck and Foster [10], the Random Forest (RF) and LR were implemented to diagnose the BV. Rank criteria such as purity increase in the node from RF and the mean ratio and standard deviation from the LR process were used to evaluate the feature relevancy and create the feature rankings. The feature selector method called RELIEF was used to create a third feature ranking. A table with the most relevant features of BV was obtained from the feature selection methods, and a sub-dataset with the top fifteen features was created. Later, this sub-dataset was used to perform additional experiments. Based on the results, features such as *Aeroccocus*, *Atopobium*, *Dialister*, *Eggerthella*, and *Gardnerella* were categorized as the most relevant for BV. It was also determined that the RF as classifier obtained the highest performance in most of the experiments performed and that only a few features are necessary to create models with predictive values above 95% of accuracy.

3 Materials and methods

This section details the dataset, methods, and techniques used in the experimental design. The methods described below were implemented in the R programming-language [15] using the R-Studio [16] environment version 1.2.5001.

3.1 Bacterial vaginosis dataset

The dataset [1] used in the experiments resulted from a study of molecular diagnosis of bacterial vaginosis. It is integrated by cervical samples of 201 gynecological tests. The microorganisms in the cervical samples were determined by qPCR technique (VBPCR). The samples and microbiological analysis were performed at the infectious disease and metabolic investigation laboratory from the Comalcalco Multidisciplinary Academic Division[1], Tabasco, México.

The dataset contained missing data and data that did not correspond to information related to bacterial vaginosis, so it was pre-processed. In this work, data preprocessing was similarly implemented as in [8]. According to the dataset providers, this process does not reduce the relevant information about bacterial vaginosis diagnosis. This process is detailed below:

1. The instances in the dataset with null values were eliminated.
2. The features in the dataset with null values were eliminated.
3. The original dataset contains three principal classes: positive, negative, and undefined. The third class - named undefined- was eliminated with the aim that the classification methods implemented in this paper identify between healthy or sick patients -positive or negative-.

Finally, the preprocessed dataset contains 173 instances (125 to BV- and 48 to BV+) and 34 features. A summary of the features in the dataset is shown in Table 1. The dataset can be provided to interested upon request to the corresponding author.

Table 1. Feature set of bacterial vaginosis dataset [1] after the preprocessing phase.

Features	Values
VBPCR	Class label: 1=positive, 2=negative
EDADENA, EDAD30	Pacient age
Citolog, CitologiaOrd, CitologiaBICAT	Normal, ordinary or abnormal citology
Crispatus, L. Gasseri, L. Iners, L. Jensenii, CripatusCq, GasseriCq, JenseniiCq, InersCq, Megasphaera Phylotype1, Atopobium vaginae, Gardnerella vaginalis., CT, NG, MH, UP, UU	Microorganisms analyzed by qPCR.
BVNumero	Pathogens number
BVCombination	Pathogens combination
HSV1/2	Herpes type 1 and 2
RMY0911ELSY	Related with HPV positivity
ELSY, HPV, HPVgenotypes, SingleHPVComplete, MultipleHPVComplete, LRIHPVComplete, PHRHPVComplete, HRHPVComplete	Related with HPV

[1] http://www.ujat.mx/damc

3.2 Best predictors of bacterial vaginosis

In the paper of Pérez-Gómez [8], the most relevant features in the BV dataset were obtained through five feature selection methods. Based on these methods, two combined feature rankings were obtained. The first one was calculated with the scaled and averaged feature relevancy across the five methods implemented. The second one was calculated based on frequency analysis. For a more detailed description see [8]. The 15 most relevant features in both two rankings are the base to create the two sub-datasets used in Scenario 2 experiments. Both two feature rankings are shown in Table 2.

Table 2. Fifteen most relevant features of bacterial vaginosis dataset obtained in [8].

Feature ranking 1	Feature ranking 2
Atopobium	BVNumero
BVCombination	Atopobium
BVNumero	GardnerellaVaginallis
MegaespheraPhylotype1	MegaespheraPhylotype1
Gardnerellavaginalis	BVCombination
MH	MH
Crispatus	CitologiaBICAT
EDADENA	Crispatus
CitologiaBICAT	LRHPVCOMPLETE
InersCq	ELSY
UP	EDAD30
CrispatusCq	LIners
Jensenii	CitologiaOrd
Citolog	RMY0911ELSY
CitologiaOrd	Citolog

3.3 Support Vector Machine

SVM is a classification algorithm that creates a model representing the sample points in the feature space by separating the classes as much as possible in that space [17]. The decision limit represented by a hyperplane is placed to leave the largest possible margin on each side [18] (See Figure 1). When a new instance is evaluated using an SVM model, this new instance will be classified into either of the classes. By maximizing the margin between the two classes the classification performance improves [19].

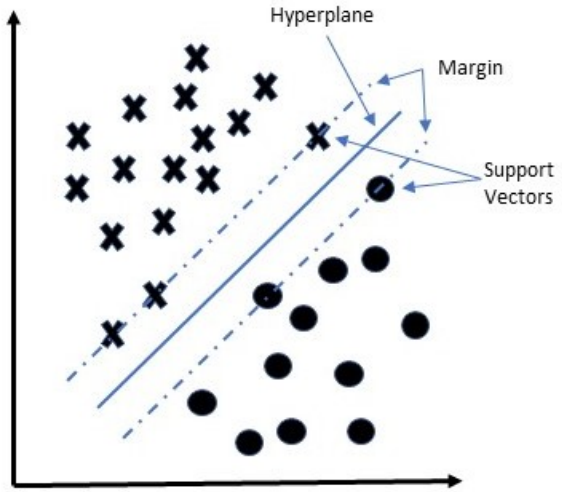


Figure 1. Graphic representation of how a Support Vectors works.

SVM is based on a Kernel function. This function transforms the data from given space -also named Input Space- to a new high dimensional space -known as Feature Space- where data can be separated with a linear surface – called hyperplane- [20] (See Figure 2).

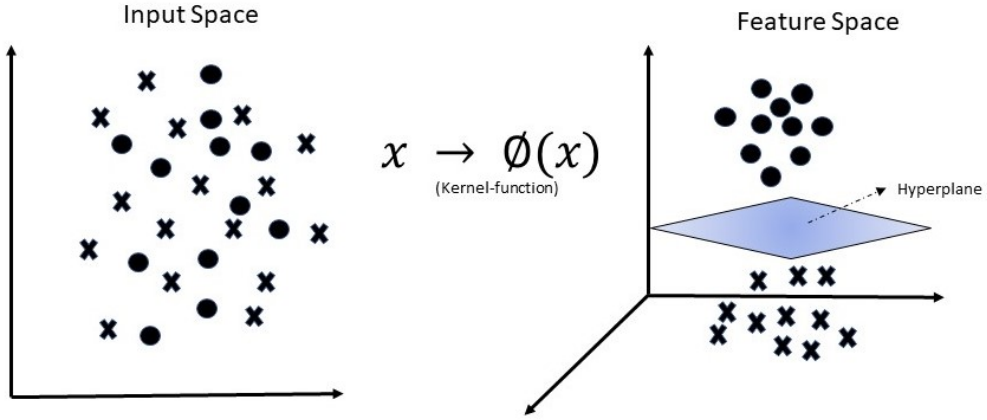


Figure 2. The data is transformed from Input Space (Left) to Feature Space (Right) with a Kernel function. Then, where the data in two-dimensions were inseparable, now in three-dimension space is separable by a hyperplane.

Suppose x_1 and x_2 are two data point, $\emptyset$ is a mapping and K denotes Kernel which is given by Equation 1.

$$K\ (x_1, x_2) = \emptyset(x_1)^T\ \emptyset(x_2) \tag{1}$$

A kernel takes two arguments, apply a mapping on the arguments and then return the value of their dot product. Kernels where mapping is identity mapping -Input Space and Feature Space are equal- is called linear Kernel and SVM using linear Kernel is called linear SVM [20]. The linear Kernel is calculated with Equation 2.

$$K\ (x_1, x_2) = x_1^T x_2 \ \Rightarrow\ \ \emptyset(x) = x \tag{2}$$

In this work, the SVM algorithm was implemented with the parameters by default. This is; Type: C-classification, kernel: linear, cost: 1. An implementation of the SVM algorithm is provided in the *e1071* R software package [21].

3.4 Performance measures

Accuracy. This performance measure is the number of instances that a classification method predicts correctly, expressed as a proportion of all instances to which it applies [22]. The accuracy is obtained in Equation (3).

$$Accuracy = \frac{(tp+tn)}{tp+\ fp+tn+fn} \tag{3}$$

Where *tp* is true positive, *tn* is true negative, *fp* is false positive, and *fn* is false negative for the prediction values in the confusion matrix obtained from a classifier model.

Balanced accuracy. The dataset [1] used in this paper is imbalanced, that is, the cardinalities of the classes are far apart. In other words, the number of instances between classes is remarkedly different. The balanced accuracy - or also named weighted accuracy - is the average of the accuracies obtained across all classes [22]. The balanced accuracy is calculated in Equation (4).

$$Balanced\ accuracy = \frac{\left(\frac{tp}{tp+fn}+\frac{tn}{fp+tn}\right)}{2} \tag{4}$$

Sensitivity. According to Bramer [23], this performance measure is the proportion of positive instances that are correctly classified as positive. It is interpreted as the level of confidence that a test will obtain a positive result correctly. The sensitivity is obtained in Equation (5).

$$Sensitivity = \frac{tp}{tp+fn} \tag{5}$$

Specificity. It is the proportion of negative instances that are correctly classified as negative [23]. The specificity is interpreted as the level of confidence that a test will obtain a negative result correctly. The specificity is calculated in Equation (6).

$$Specificity = \frac{tn}{tn+fp} \tag{6}$$

***K*-Folds Cross-Validation.** It is a method for obtaining reliable estimates for small datasets and prevent over fitting. Torgo [24] describes this method as follows: Obtain *k* equally sized and random subsets of the dataset. For each of these *k* subsets, a model is built using the remaining *k*-1 sets to evaluate this model. The performance of the model is stored and the process is repeated for all remaining subsets. In the end, there are *k* performance measures, all obtained by testing a model. These *k* performances are averaged, with which mean performances are obtained. In this work, the value used for *k* is 10, as similarly was used in [8][9][10], presented in Section 2.

4 Experimental design

The classification experiments with SVM to investigate its capability to accurately learn to identify positive and negative BV instances were performed in three scenarios as described below:

4.1 Scenario 1: Entire features set.

Thirty runs of SVM under a 10-fold cross-validation scheme were performed. This scheme was described in Section 3.4. Across all 30 runs a different seed was used to ensure different data splitting on each run. In this scenario, the entire feature set was used, that is 34 features plus the class label feature. The performance of all 30 cross-validation processes were averaged, which is given as final performance of this scenario. All metrics such as accuracy, balanced accuracy, specificity, and sensitivity are also reported.

4.2 Scenario 2: Sub-datasets from two feature rankings.

Were performed 30 runs of SVM with 10FCV, but this time, the feature rankings resulting in Pérez-Gómez [8] and described in Section 3.2 were used as sub-datasets. These sub-datasets, with only 15 features, corresponding to the BV features determined as the most important predictors by the feature selection methods used in that research. The performance measures calculated and the validation scheme used were the same as those described in Scenario 1.

4.3 Scenario 3: One feature taken at a time.

Were performed 30 runs of SVM by each feature in the dataset using the class label and one feature at a time. This, to evaluate the performance of the SVM to distinguish between both classes of BV using only one feature individually. As in Scenario 1, the validation scheme and the performance measures were calculated in the same way from each experiment performed.

The process of all scenarios is shown in Figure 3.

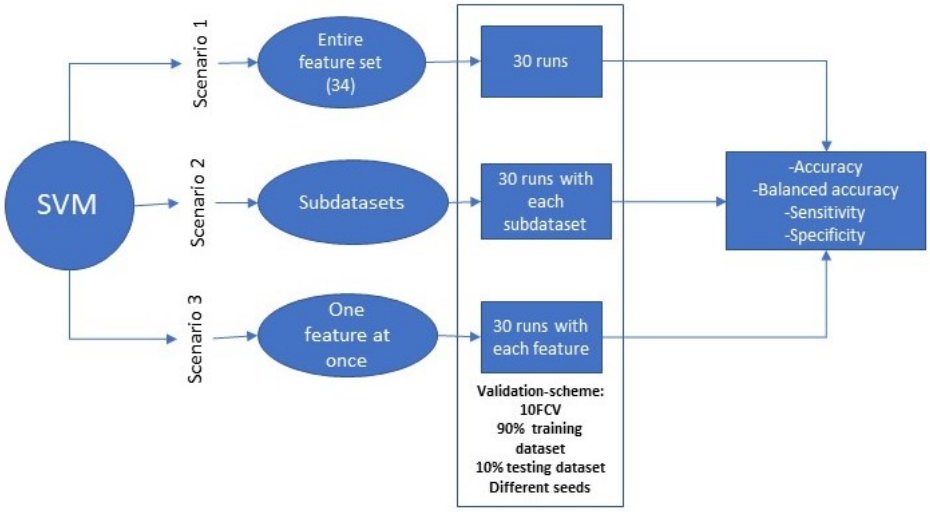


Figure 3. Experimental design performed with Support Vector Machine (SVM) to determine the capability to classify between positive or negative Bacterial Vaginosis. The experiments were implemented in three different scenarios. 10FCV: 10- folds cross validation.

5 Results

Experiments were completed, and detailed results are given in this section.

A sample of how an SVM with a linear kernel separates the training instances between classes positive and negative is shown in Figure 4. This separation is the hyperplane created by SVM. Here, only two features -*MegaespheraPhylotype1* and *InersCq*- were used to training a SVM model.

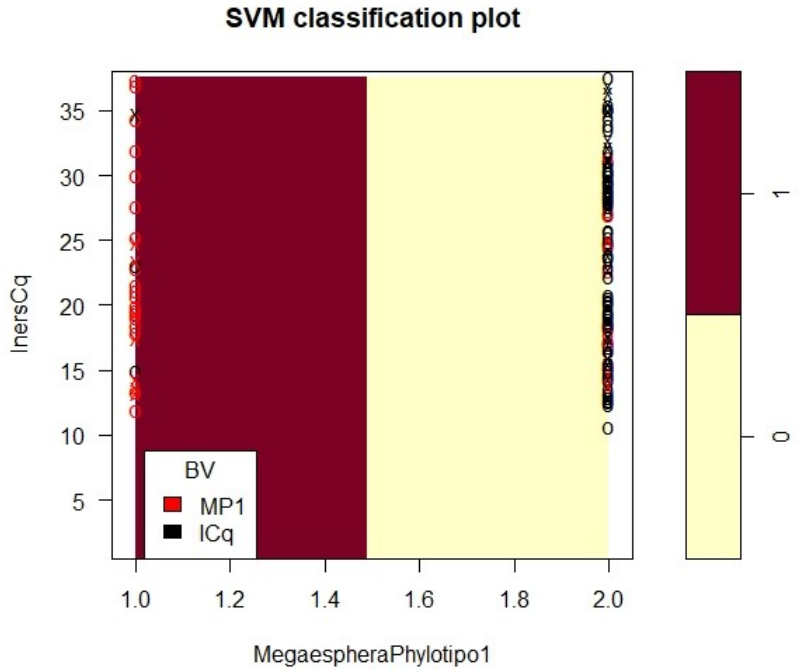


Figure 4. Hyperplane traced by a linear-Kernel support vector machine (SVM) model using two features of the bacterial vaginosis training set. The red area in the plane represents the training instances classified as BV positive. The yellow area represents the training instances classified as BV negative.
MP1: MegaespheraPhylotipo1, ICq: InersCq.

5.1 Scenario 1.

Experiments with 30 runs of SVM under a 10-fold cross-validation scheme were performed. Here, the entire set of features was used to evaluate the performance of the classifier. The results of this scenario are described in Table 3.

Table 3. Performance of support vector machine (SVM) obtained in Scenario 1. For the evaluation of the classifier, the entire feature set in the bacterial vaginosis (BV) dataset [1] was used.

Dataset	Features number	Accuracy	Balanced accuracy	Sensitivity	Specificity
Entire BV dataset	34	1	1	1	1

According to results, the implementation of SVM with the use of the entire feature set obtained the highest level for a classifier across all performance measures computed.

5.2 Scenario 2

The 30 runs of SVM with 10FCV by each sub-dataset were performed. The sub-datasets, created with only the 15 most relevant features from the feature rankings in the previous work [5] were used in these experiments. The results of this scenario are shown in Table 4.

Table 4. Performance of Support Vector Machine (SVM) using two sub-datasets crated with only the 15 most relevant features of bacterial vaginosis (BV) from a previous paper [8].

Dataset	Features number	Accuracy	Balanced accuracy	Sensitivity	Specificity
Subdataset 1	15	1	1	1	1
Subdataset 2	15	1	1	1	1

In the SVM experiments where the sub-datasets were used, similar results to those first were obtained. With the two both sub-datasets created with only the 15 most important features of bacterial vaginosis, 100% accuracy was obtained. Balanced accuracy, sensitivity, and specificity also obtained 100%.

5.3 Scenario 3

The results of 30 iterations of SVM under a 10FCV scheme are provided. In this scenario, one feature at a time was used to create a classification model with SVM. By each feature in the dataset this step was repeated. The results of this scenario are shown in Table 5. The features are sorted according to the accuracy obtained in the experiment.

Table 5. Classification performance obtained by support vector machine (SVM) with the use of one feature at a time in the bacterial vaginosis (BV) dataset.

Features in BV dataset	Accuracy	Balanced accuracy	Sensitivity	Specificity
BVCombination	1	1	1	1
Atopobium	0.953622	0.948720	0.960106	0.937333
MegaespheraPhylotype1	0.872826	0.790585	0.976004	0.605166
GardnerellaVaginalis	0.863221	0.874914	0.848162	0.901666
BVNumero	0.844196	0.789775	0.912051	0.6675
MH	0.820758	0.689790	0.984081	0.3955
ELSY	0.740012	0.537670	0.992008	0.083333
CitologiaBICAT	0.722712	0.5	1	0
Crispatus	0.722712	0.5	1	0
Citolog	0.722712	0.5	1	0
CitologiaOrd	0.722712	0.5	1	0
MY0911_A	0.722712	0.5	1	0
LRHPVCOMPLETE	0.722712	0.5	1	0
EDAD30	0.722712	0.5	1	0
JenseniiCq	0.722712	0.5	1	0
UP	0.722712	0.5	1	0
RMY0911ELSY	0.722712	0.5	1	0
LIners	0.722712	0.5	1	0
HSV12	0.722712	0.5	1	0
CT	0.722712	0.5	1	0
MultipleHPVCOMPLETE	0.722712	0.5	1	0
LGasseri	0.722712	0.5	1	0
Jensenii	0.722712	0.5	1	0
HRHPVCOMPLETE	0.722712	0.5	1	0
InersCq	0.722712	0.5	1	0
GasseriCq	0.722712	0.5	1	0
SingleHPVCOMPLETE	0.722712	0.5	1	0
UU	0.722712	0.5	1	0
NG	0.722712	0.5	1	0
HPV	0.722712	0.5	1	0
MG	0.722712	0.5	1	0
pHRHPVCOMPLETE	0.721971	0.499487	0.998974	0
CrispatusCq	0.717743	0.496538	0.993076	0
EDADENA	0.713480	0.493653	0.987307	0

Based on the accuracies obtained by SVM and the features individually evaluated, features such as “BVCombination”, “Atopobium”, “Megaesphera”, “Gardnerella vaginalis”, “BVNumero” and “MH” are highlighted between the most related to the BV diagnosis.

Finally, a comparative graphic of all experimented scenarios with SVM is provided in Figure 5.

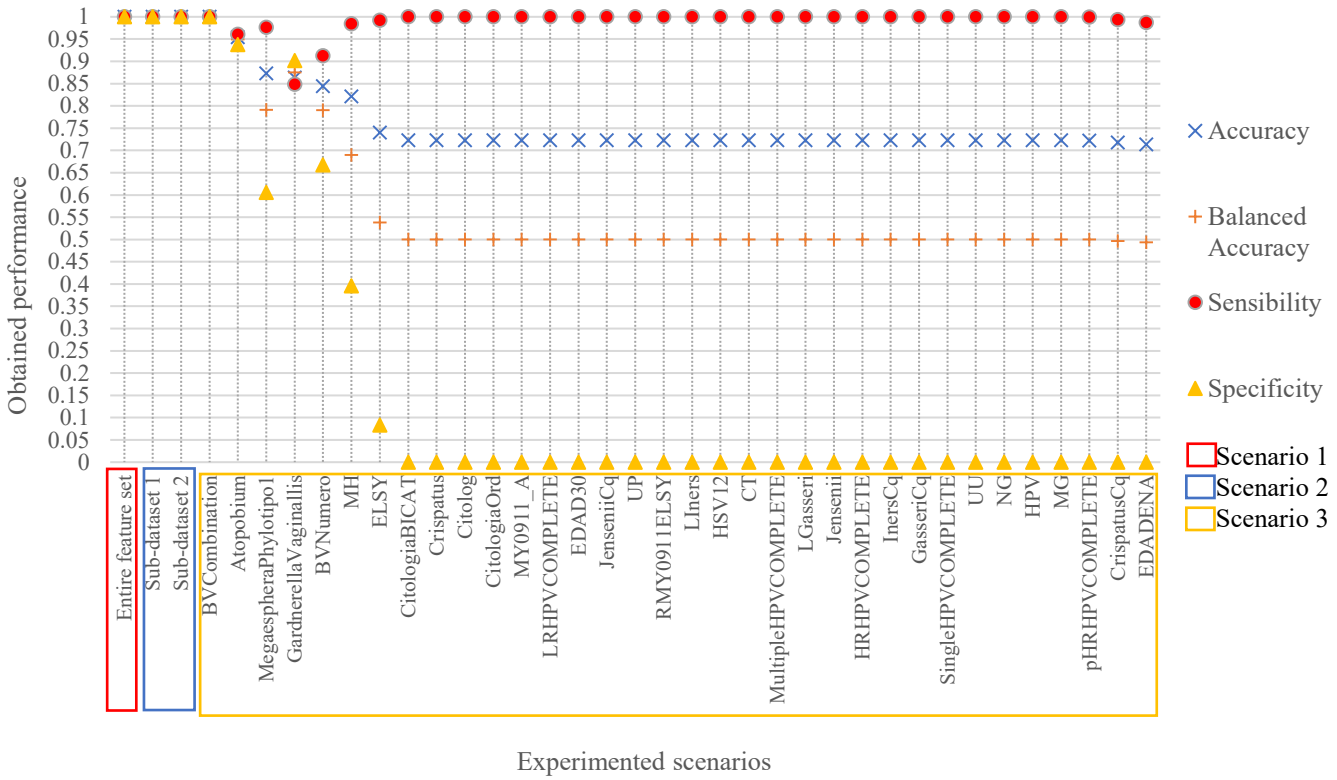


Figure 5. Performance obtained by support vector machine (SVM) into Scenario 1 described in section 5.1, those of Scenario 2 described in section 5.2, and those of Scenario 3 described in section 5.3.

According to the previous graphic, the highest results with SVM were obtained in Scenario 1, Scenario 2, and in the experiments performed with the feature named "BVCombination".

6 Conclusions

In this work, the capability of SVM to identify microorganisms associated with bacterial vaginosis in an effort to create a computer-based diagnosis was investigated. The experiments were divided into three scenarios. In the first one, the entire feature set of the dataset was used to evaluate the classifier. In the second scenario, the classifier used two sub-datasets with the most relevant features of BV. In the third scenario, the features were individually used to create models with SVM by each. The performance measures obtained by SVM in all scenarios were compared.

Results confirm SVM is a classifier highly accurate in the use of the entire dataset of BV. Even if sub-datasets with only the fifteen most relevant features of bacterial vaginosis are used, accuracy of up to 100% can be reached. This means that the use of a reduced set of features can generate models with high BV classification accuracy. To knowing the optimal, maximum, and minimal number of features, more experiments are necessary. However, these results highlight the potential usefulness of SVM to identify those microorganisms related to BV etiology, thus reducing the number of laboratory assays necessary to determine the presence of BV with diagnostic accuracy.

The implementation of SVM with only one feature at a time to determine its classification capability is analyzed below. "*BVCombination*" as a BV feature used to create a model with SVM obtained the possible

highest results in the performance of a classifier. The importance of this feature is that it denotes the combination of previously identified microorganisms related to bacterial vaginosis (anaerobic bacteria as *Gardnerella vaginalis*, *Prevotella spp.*, *Mycoplasma hominis*, among others), and it represents the presence or absence of those microorganisms in the vaginal sample. If in contrast, the importance of the features representing the count of microorganisms in the dataset is highlighted, *Atopobium* can be considered the most related feature to the diagnosis of BV, followed by species such as *Megaesphera* and *Gardnenerella vaginalis*. This approach results essential for BV diagnosis because in most cases the transition between normal and disease status lies on the microbial density of organisms usually present in cervicovaginal microenvironment. From a biological point of view, although the presence Gardnerella vaginalis is commonly related to the development of BV, most of the studies suggest that the main feature that distinguishes the role of Gardnerella vaginalis is the high density observed in vaginal samples, which is frequently associated with its pathological behavior. On the other hand, although the presence of Atopobium vaginae and/or Megasphaera phylotype 1 are not frequently associated to clinical signs of BV, molecular studies have demonstrated their high rate of prevalence in women with a confirmed diagnosis of BV. Additional studies are required to determine the biological significance of the findings obtained with the present work.

Based on the variability observed between the accuracy and balanced accuracy performance measures, experiments with techniques for data balancing as possible future work are proposed. This proposal is founded with the aim to improve the results obtained in this work.

This is an ongoing research that is part of an extensive exploratory analysis on machine learning methods applied in the bacterial vaginosis study. More experiments with other feature selection methods and classification techniques are being investigated.

References

1. Sanchez-garcia, E.K., Contreras-paredes, A., Martinez-abundis, E., Garcia-chan, D., De La Cruz-Hernandez, E.: Molecular Epidemiology of Bacterial Vaginosis and Its Association with Sexually Transmitted Pathogens in Healthy Women. J. Med. Microbiol. Mol. 68, (2019). https://doi.org/https://doi.org/10.1099/jmm.0.001044.
2. Hoang, T., Toler, E., DeLong, K., Mafunda, N.A., Bloom, S.M., Zierden, H.C., Moench, T.R., Coleman, J.S., Hanes, J., Kwon, D.S., Lai, S.K., Cone, R.A., Ensign, L.M.: The cervicovaginal mucus barrier to HIV-1 is diminished in bacterial vaginosis. PLoS Pathog. (2020). https://doi.org/10.1371/journal.ppat.1008236.
3. Srinivasan, S., Fredricks, D.N.: The Human Vaginal Bacterial Biota and Bacterial Vaginosis. Interdiscip. Perspect. Infect. Dis. 2008, 1–22 (2008). https://doi.org/10.1155/2008/750479.
4. Onderdonk, A.B., Delaney, M.L., Fichorova, R.N.: The human microbiome during bacterial vaginosis. Clin. Microbiol. Rev. 29, 223–238 (2016). https://doi.org/10.1128/CMR.00075-15.
5. Kusters, J.G., Reuland, E.A., Bouter, S., Koenig, P., Dorigo-Zetsma, J.W.: A multiplex real-time PCR assay for routine diagnosis of bacterial vaginosis. Eur. J. Clin. Microbiol. Infect. Dis. 34, 1779–1785 (2015). https://doi.org/10.1007/s10096-015-2412-z.
6. Parra, G.I.M.: Aspectos clínicos y diagnóstico de laboratorio de la vaginosis bacteriana. Rev. Habanera Ciencias Medicas. 14, 611–623 (2015).
7. Money, D.: The laboratory diagnosis of bacterial vaginosis. Can. J. Infect. Dis. Med. Microbiol. 16, 77–79 (2005). https://doi.org/10.1155/2005/230319.
8. Perez-Gomez, J.F., Canul-Reich, J., Hernandez-De la Cruz, E.: Combinación de Rankings como Método para la Identificación de Biomarcadores de Vaginosis Bacteriana. Res. Comput. Sci. (2020).
9. Beck, D., Foster, J.A.: Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PLoS One. 9, (2014). https://doi.org/10.1371/journal.pone.0087830.
10. Beck, D., Foster, J.A.: Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis. BioData Min. 8, 1–9 (2015). https://doi.org/10.1186/s13040-015-0055-3.
11. 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. Appl. Sci. 10, 3291 (2020). https://doi.org/10.3390/app10093291.

12. Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S.K., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., Brotman, R.M., Davis, C.C., Ault, K., Peralta, L., Forney, L.J.: Vaginal microbiome of reproductive-age women. Proc. Natl. Acad. Sci. U. S. A. 108, 4680–4687 (2011). https://doi.org/10.1073/pnas.1002611107.
13. Baker, Y.S., Beck, D., Agrawal, R., Dozier, G., Foster, J.A.: Detecting Bacterial Vaginosis using machine learning. In: Proceedings of the 2014 ACM Southeast Regional Conference, ACM SE 2014. pp. 1–4 (2014). https://doi.org/10.1145/2638404.2638521.
14. Srinivasan, S., Hoffman, N.G., Morgan, M.T., Matsen, F.A., Fiedler, T.L., Hall, R.W., Ross, F.J., McCoy, C.O., Bumgarner, R., Marrazzo, J.M., Fredricks, D.N.: Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PLoS One. 7, (2012). https://doi.org/10.1371/journal.pone.0037818.
15. Team, R.C.: R: A language and environment for statistical computing, http://www.r-project.org/, (2013).
16. Team, Rs.: RStudio: Integrated Development for R, http://www.rstudio.com/, (2020).
17. Wang, H., Zheng, B., Yoon, S.W., Ko, H.S.: A support vector machine-based ensemble algorithm for breast cancer diagnosis. Eur. J. Oper. Res. 267, 687–699 (2018). https://doi.org/10.1016/j.ejor.2017.12.001.
18. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422 (2002). https://doi.org/10.1023/A:1012487302797.
19. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media (2009).
20. Chauhan, V.K., Dahiya, K., Sharma, A.: Problem formulations and solvers in linear SVM: a review. Artif. Intell. Rev. 52, 803–855 (2019). https://doi.org/10.1007/s10462-018-9614-6.
21. Chang, C.C., Lin, C.J.: LIBSVM: A Library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, (2011). https://doi.org/10.1145/1961189.1961199.
22. Witten, I.H., Frank, E., Geller, J.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Elsevier (2002). https://doi.org/10.1145/507338.507355.
23. Bramer, M.: Introduction to Data Mining. Presented at the (2013). https://doi.org/10.1007/978-1-4471-4884-5_1.
24. Torgo, L.: Data Mining with R, learning with case studies. Chapman and Hall/CRC. (2010).