ISSN 1870-4069

Combinación de rankings como método para la identificación de biomarcadores de vaginosis bacteriana

Jesús Francisco Pérez-Gómez[1], Juana Canul-Reich[1], Erick De-La-Cruz-Hernandez[2]

[1] Universidad Juárez Autónoma de Tabasco,
División Académica de Ciencias y Tecnologías de la Información,
México
[2] Universidad Juárez Autónoma de Tabasco,
División Académica Multidisciplinaria de Comalcalco,
México

juana.canul@ujat.mx

Resumen. La Vaginosis Bacteriana (VB) es una condición patológica que se detecta en mujeres sexualmente activas asintomáticas y sintomáticas que causa secreción vaginal inusual. Esta condición no es potencialmente peligrosa, pero aumenta el riesgo de contraer Enfermedades de Transmisión Sexual (ETS) y nacimientos prematuros. Sus causas no están del todo claras, pero pruebas moleculares [1, 2] sugieren que ciertos microorganismos se expresan durante esta condición. Para hallarlos, se requieren estudios que identifiquen biomarcadores. En este trabajo se utilizaron Algoritmos de Aprendizaje Automático (AAA) para encontrar los mejores atributos predictivos, es decir, biomarcadores de la VB y estudiar su relación con la enfermedad. Para ello, se utilizaron métodos de selección de atributos como *Relief*, *Fisher Score*, *Decision Trees*, *Boruta* y *Random Forests*. Los experimentos de cada método bajo un esquema de Validación Cruzada (VC) generaron cinco "rankings individuales". Con base en ellos se obtuvieron dos rankings combinados. Un primer ranking combinado se obtuvo al escalar y promediar los rankings individuales. Un segundo ranking se basa de un análisis de frecuencias a partir de los rankings individuales. Una comparativa de los rankings obtenidos identifican atributos en común como biomarcadores de la VB. El conjunto de datos analizado es una prueba microbiológica vaginal de 201 mujeres de Tabasco, México. Este es un primer esfuerzo para explorar los atributos relevantes en un conjunto de datos real de vaginosis bacteriana y es la base para planificar diversos experimentos con algoritmos clasificadores. Otros métodos de clasificación y selección de atributos para la detección de la VB se encuentran en investigación.

Palabras clave: Selección de atributos, decision tree, relief, fisher score, Boruta, Random Forest.

Combination of Rankings as a Method for Biomarker Identification of Bacterial Vaginosis

Abstract. Bacterial Vaginosis (BV) is a pathological condition detected in asymptomatic and symptomatic sexually active women that causes an unusual

vaginal discharge. This condition is not potentially dangerous, but it increases the risk of contracting Sexually Transmitted Diseases (STDs) and premature births. Its causes are not entirely clear, but molecular evidence [1, 2] suggests that certain microorganisms are expressed during this condition. To find them, studies of biomarkers identifying are required. In this work, Machine Learning Algorithms (AAA) were used to find the best predictive features and to study its relationship with the disease. For this, feature selection (FS) methods such as Relief, Fisher Score, Decision Trees, Boruta, and Random Forests were used. The experiments of each method under a Cross-Validation (VC) scheme generated five "individual rankings." Based on them, two combined rankings were obtained. A first combined ranking was obtained by scaling and averaging the individual rankings. A second combined ranking was obtained based on an analysis of frequencies from the individual rankings. A comparison of the combined rankings identifies features in common as biomarkers of BV. The dataset analyzed is a vaginal microbiological test of 201 women from Tabasco, Mexico. This is a first effort to explore the most relevant features in a real bacterial vaginosis dataset and it is the basis for the implementation of experiments with classifying algorithms. Divers classifiers algorithms for the detection of BV and other feature selection methods are under further investigation.

Keywords: Feature selection, decision trees, relief, fisher score, Boruta, Random Forest.

1. Introducción

La Vaginosis Bacteriana es una alteración del microbiota vaginal. Es una disbiosis que comúnmente se detecta en mujeres sexualmente activas asintomáticas y sintomáticas dónde se observan hallazgos clínicos como flujo vaginal anormal en color (gris o verde) y olor a pescado [1]. Las mujeres con esta infección tienen un 60% más riesgos de contraer Virus de Inmunodeficiencia Humana (VIH) y aumenta en un 30% las probabilidades de transmitir el VIH a una pareja no infectada [3]. La epidemiología de la vaginosis bacteriana como un estado comunitario de microbiota vaginal sigue siendo poco conocida, y muchas veces controvertida [2].

Hasta ahora, algunos estudios moleculares mencionan organismos anaerobios como *Lactobacillus*, *Atopobium vaginae*, *Gardnerella vaginalis* y *Megasphaera* tipo 1 y 2 implicados en la expresión de esta afección [1, 2]. Un biomarcador o marcador biológico se refiere a cualquier sustancia, estructura o proceso que pueda medirse en el cuerpo o sus productos, que permita influir o predecir la enfermedad [4]. Esta respuesta medible puede ser funcional, fisiológica, bioquímica a nivel celular o una interacción molecular [5]. Para identificarlos, se requiere determinar el nivel de información clínicamente relevante de todos los datos obtenidos en torno a la enfermedad [6].

A partir de este enfoque los mejores predictores o atributos relevantes se pueden determinar al utilizar métodos de aprendizaje automático. Por tanto, esta investigación se aborda como un problema de selección de atributos.

El objetivo principal es explorar los atributos y determinar los más relevantes en el conjunto de datos de VB utilizando cinco Algoritmos de Selección de Atributos (ASA). Específicamente se utilizaron los métodos *Relief*, *Fisher Score*, *Decision Tree*, *Boruta* y *Random Forests*. Con cada algoritmo se obtuvo el nivel de relevancia de cada

atributo, lo que conforman los "rankings individuales". Con base en éstos se obtuvieron dos rankings combinados de atributos, que muestran los predictores más relevantes de la BV. Un primer ranking combinado se basa en el promedio de los 5 valores de importancia obtenidos de los rankings individuales después de ser escalados al rango entre 0 y 1. El segundo ranking combinado se basa en un análisis de frecuencia y la obtención de la moda estadística a partir de las posiciones de los atributos en los rankings individuales.

Los dos rankings combinados de atributos con los mejores predictores de BV que se derivan de los experimentos son provistos. La base de datos utilizada en los experimentos [1] son datos de diagnóstico molecular de Vaginosis Bacteriana conformado de 201 instancias y 57 atributos. Las muestras corresponden a mujeres de Comalcalco, Tabasco y fueron obtenidas y analizadas en el Laboratorio de Investigación en Enfermedades Infecciosas y Metabólicas de la DAMC-UJAT.
Finalmente, los objetivos que sigue esta investigación se resumen en los siguientes:

1. La obtención de un primer ranking combinado de atributos con los predictores más relevantes de la VB a partir de la combinación de cinco algoritmos de selección de atributos.
2. La obtención de un segundo ranking combinado de atributos con los predictores más relevantes de la VB basado en un análisis de frecuencias a partir de los rankings individuales.

La motivación se fundamenta en analizar los datos proporcionados desde la perspectiva del Aprendizaje Automático. Experimentos adicionales, métodos y técnicas del área de la Inteligencia Artificial están siendo propuestos para la ampliación de la investigación a partir de los resultados obtenidos.

Este documento se organiza de la siguiente manera. La Sección 2 describe algunas investigaciones relacionadas a los métodos y desarrollos experimentales enfocados a la selección de atributos. La Sección 3 detalla el conjunto de datos utilizado y los métodos de aprendizaje automático implementados en esta investigación. La sección 4 explica a detalle las fases experimentales de la investigación. La Sección 5 muestra los resultados obtenidos en todos los experimentos desarrollados, y finalmente, en la Sección 6 se proporcionan las conclusiones generales del proyecto.

2. Trabajos relacionados

En esta sección se describen algunos proyectos de investigación relacionados a la selección de atributos o reducción de dimensiones propios del área de aprendizaje automático, y que han motivado la utilización de los algoritmos propuestos en el desarrollo experimental.

El propósito del trabajo de Yolanda Baker y demás [7] fue descubrir los atributos relevantes de la VB y aplicar algunos algoritmos de clasificación para su diagnóstico. Los autores aplicaron veinte algoritmos de selección de atributos en combinación con nueve algoritmos de clasificación utilizando la herramienta WEKA. La precisión -proporción de instancias correctamente clasificadas-, recall -proporción de verdaderos positivos-, y el número de atributos reducidos fueron algunas de las métricas de rendimiento que se consideraron en este estudio. Los autores encontraron que los

Tabla 1. Lista de atributos que conforman el conjunto de datos de Vaginosis Bacteriana [1].

Atributos	Valores
VBPCR	Etiqueta clase: 1=positivo, 2=Negativo, 3=Indeterminado
EDADENA, EDAD30	Edad del paciente
Citolog, CitologiaOrd, CitologiaBICAT	Citologia normal, ordinaria o anormal
Crispatus, L. Gasseri, L. Iners, L. Jensenii, CripatusCq, GasseriCq, JenseniiCq, InersCq, Megasphaera Phylotipo1, Atopobium, Gardnerella V., CT, NG, MH, UP, UU	Microorganismos obtenidos mediante qPCR.
BVNumero	Numero de patógenos
BVCombination	Combinación de patógenos
HSV12	Herpes tipo 1
RMY0911ELSY	Relacionado con MYDE0911-ELSY
ELSY, HPV, HPVgenotypes, SingleHPVComplete, MultipleHPVComplete, LRIHPVComplete, PHRHPVComplete, HRHPVComplete	Relacionados con VPH
@6, @11, @42, @44, @84, @E626	Relacionados con VPH
E653, E666, E616, E618, E631, E633, E635, E639, E645, E651, E652, E656, E658, E659, E673	Relacionados con VPH

algoritmos *Functional Trees (FT)* y *WrapperSubSetEval* resultaron ser la mejor combinación de acuerdo con las métricas utilizadas.

Beck y Foster [8] utilizaron los algoritmos *Random Forests (RF)* y *Logistic Regression (LR*) para diagnosticar la VB. Para rankear los atributos consideraron *purity increase in the node* como medida de rendimiento en *RF*. Para *LR*, los atri-butos fueron rankeados al considerar el cociente medio y la desviación estándar en todos los experimentos de la validación cruzada. *Relief* se implementó como una alternativa para calcular un tercer ranking. Como resultado se obtuvo una tabla con los atributos más relevantes. Los autores categorizaron atributos como *Aeroccocus*, *Atopobium*, *Dialister, Eggerthella* y *Gardnerella* como los más relevantes.

Muhammad y otros [9] presentaron un sistema de diagnóstico clínico para enfermedades del corazón en el cual propusieron un método de selección de atributo basado en tres algoritmos selectores: *Fisher Score-based feature selection, Forward Feature Selection (FFS)* y *Reverse Feature Selection (RFS*). En la fase de clasificación, los autores utilizaron *SVM* con el kernel de base radial. Para evaluar el rendimiento de las técnicas propuestas se consideraron medidas como *Mathews's correlation*, precisión, especificidad y sensibilidad. Las bases de datos para realizar las pruebas fueron obtenidas del repositorio *UCI (University of California Irvine*). Finalmente, los autores obtuvieron mejores resultados con la metodología propuesta en comparación con los métodos de selección de atributos de manera individual, tanto en rendimiento como en costos computacionales.

La investigación de Kumar y Shaikh [10] demuestra que el algoritmo *random forests* utilizado en los experimentos mejoró su rendimiento al utilizar los atributos seleccionados con *Boruta*. Para llegar a esta conclusión, implementaron algoritmos de selección de atributos como *Relief, Recursive Feature Elimination (RFE) y Boruta*. El conjunto de datos para los experimentos fue obtenido del repositorio de aprendizaje automático UCI que incluye 303 observaciones y 13 atributos. Para los experimentos, primero particionaron los datos en conjuntos de entrenamiento (70%) y conjuntos de prueba (30%).

El método de validación cruzada de 10 pliegues se implementó para evitar el sobreajuste del modelo. Los autores concluyeron que todos los métodos selectores obtuvieron resultados similares, la diferencia se centró en el orden de importancia de los atributos. Sin embargo, la combinación de *random forests* como clasificador y la utilización del subconjunto de atributos confirmados por *Boruta* obtuvieron medidas de rendimiento superiores a las demás pruebas.

3. Materiales y métodos

En esta sección se detalla el conjunto de datos utilizado para la implementación de los experimentos propuestos, así como los métodos de selección de atributos utilizados en el desarrollo experimental. Los métodos descritos a continuación se implementaron en el lenguaje R, y para ello se utilizó el ambiente de trabajo R-Studio en la versión 1.2.5001.

3.1. Conjunto de datos

La base de datos que se utilizó para los experimentos de esta investigación se basa en un estudio de diagnóstico molecular de Vaginosis Bacteriana [1]. Está integrado por muestras cervicales de 201 exámenes ginecológicos. Los microorganismos implicados en esta condición fueron determinados por la técnica *Quantitative Polymerase Chain Reaction* (*qPCR*). El conjunto de datos se conforma de 201 instancias y 57 atributos. Las muestras y el análisis microbiológico se realizaron en los Laboratorios de Investigación en Enfermedades Infecciosas y Metabólicas de la División Académica Multidisciplinaria de Comalcalco, Tabasco. Un resumen de los atributos que conforman la base de datos se muestra en la Tabla 1.

3.2. Preprocesamiento de datos

El conjunto de datos proporcionado contiene algunos datos faltantes, por lo que antes de ser utilizado se sometió a un preprocesamiento. Para esto, las instancias y atributos con valores nulos se eliminaron. De acuerdo con los proveedores de los datos [1], este proceso no representa la reducción de información relevante para la detección de la VB. Los atributos eliminados corresponden a datos relacionados con el VPH, por lo que su eliminación no representa una pérdida de información para el estudio de la VB.

Originalmente los datos proporcionados contienen tres etiquetas: positivo, negativo e indeterminado. Para los propósitos particulares de esta investigación, la clase

indeterminada se eliminó, ya que en la fase de entrenamiento de los métodos utilizados el interés es identificar entre pacientes enfermos y no enfermos.

Finalmente, la base de datos preprocesada contiene 1 etiqueta clase, 34 atributos y 173 instancias.

3.3. Algoritmos de selección de atributos

Los ASA, por sus siglas, son técnicas para el descubrimiento del conocimiento que proporcionan el entendimiento del problema a través del análisis de los atributos más relevantes [11]. En tareas de clasificación, estos métodos permiten mejorar el rendimiento de un algoritmo clasificador al reducir los costos y sobrecargas operacionales. Muchos ASA incluyen un ranking de atributos como un mecanismo de selección principal, lo que denota los atributos con mayor y menor relevancia [12]. Un ranking de atributos evalúa la relevancia individual de los atributos, sin considerar las posibles interacciones entre ellos [13]. Los procedimientos de cada ASA para calcular un ranking de atributos se describen a continuación.

Árboles de Decisión (*Decision Tree*)

Un árbol de decisión (*DT*, por sus siglas en inglés) es un algoritmo de clasificación. Sin embargo, en sus procedimientos evalúa la importancia de los atributos. En el proceso de creación de un árbol, la relevancia de los atributos se obtiene de acuerdo a la entropía. Max Bramer [14] describe la entropía como una medida teórica de "incertidumbre" de la información que contiene un conjunto de datos de entrenamiento. El paquete en R *caret* [15] proporciona una implementación de árboles de decisión en el algoritmo *J48*.

Bosques Aleatorios (*Random Forests*)

Este método de ensamble es una combinación de árboles de decisión relacionados entre ellos [16]. Su enfoque genera una gran cantidad de árboles, donde cada árbol depende de los valores de un vector de instancias aleatorias, mide su desempeño y selecciona los mejores del conjunto [14]. El paquete en R *FSelector* [17] proporciona una implementación del algoritmo *Random Forests*.

Boruta

Este es un método de selección de atributos tipo envoltura que se construye en torno al algoritmo de clasificación *Random Forest*. La idea general es que en cada iteración se genere una serie de atributos sombra a partir de los predictores, copia cada uno de ellos y permuta entre sí los elementos de cada nueva columna. Un modelo con *Random Forest* se ajusta, y las importancias relativas de cada atributo se calculan. Si una variable queda por debajo de las sintéticas (ruido), será indicativo de que su aportación al modelo es dudosa, y por tanto se elimina. El proceso continua hasta que todas las variables son aceptadas, rechazadas o finalizan las iteraciones [18]. El paquete en R *Boruta* [19] proporciona una implementación del algoritmo *Boruta.*

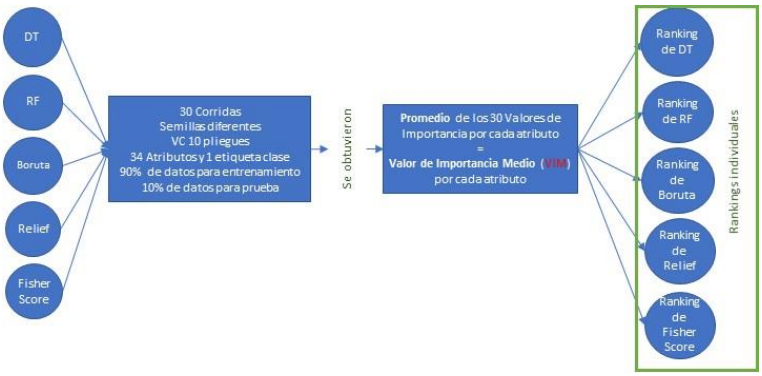

Fig. 1. Proceso de creación de los rankings individuales. En letra roja se especifica el criterio que se utilizó para el ordenamiento de mayor a menor relevancia. DT: *decision tree*; RF: *random forests*; VC: validación cruzada.

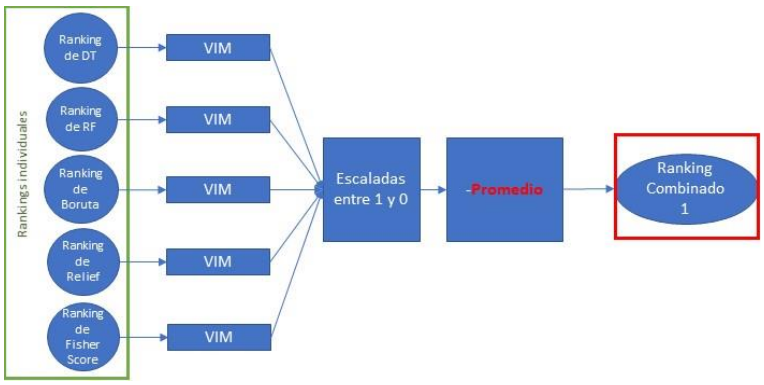

Fig. 2. Proceso de creación del primer ranking combinado de atributos. En color rojo se realza la medida considerada para ordenar los atributos del ranking combinado. DT: *decision tree*, RF: *random forests*; VIM: valor de importancia medio.

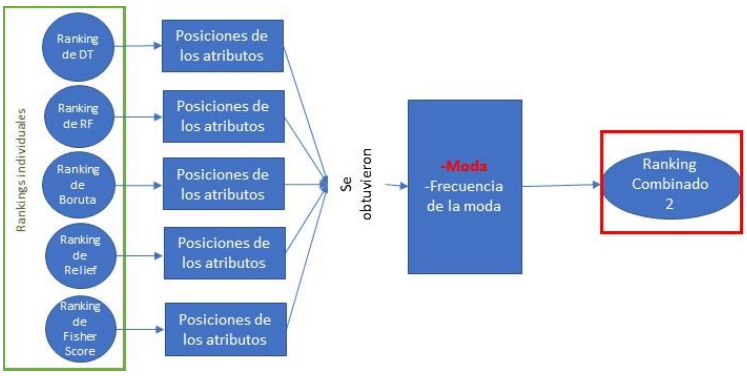

Fig. 3. Proceso de creación del segundo ranking combinado de atributos. En color rojo se realza

Relief

Relief es un método multivariado tipo filtro para la selección de atributos. Selecciona atributos relevantes basándose en la diferencia de valores de atributos entre pares de instancias utilizando el algoritmo *nearest-neighbor* y proporciona una puntuación para cada atributo [20]. El algoritmo *nearest-neighbor* se basa en la distancia Manhattan. El paquete en R *FSelector* [17] proporciona una implementación del algoritmo *Relief.*

Fisher Score

Fisher score es un método supervisado de selección de atributos que se utiliza para la reducción de dimensionalidad. Éste crea una puntuación para cada atributo de manera independiente bajo el criterio de Fisher, que conduce a un subconjunto de características subóptimo [21]. La idea general de este método es encontrar los atributos donde la distancia entre los puntos de clases diferentes sea tan grande como sea posible, mientras que la distancia entre los puntos de datos de la misma clase sea tan pequeña como sea posible [22]. Matemáticamente, el *F-score* -otro nombre con el que se conoce- proporciona una medida de qué tan bien un atributo a la vez puede discriminar entre diferentes clases [23]. Cuanto más alto sea el *F-score*, mayor el poder de discriminación. El paquete en R *PredPsych* [23] proporciona una implementación del algoritmo *Fisher Score*.

4. Diseño Experimental

La fase experimental para explorar los predictores más relevantes, es decir, los biomarcadores de la Vaginosis Bacteriana consisten de 30 corridas de cada ASA bajo un esquema de validación cruzada de 10 pliegues. A través de las 30 corridas se utilizaron semillas diferentes para asegurar la aleatoriedad de los datos. Esto significa que en cada corrida el conjunto de datos completo es fraccionado en 10 partes iguales, y cada fracción considera el 90% de las instancias para datos de entrenamiento y el 10% restante para datos de prueba. Al final, una corrida de validación cruzada promedia las 10 medidas de rendimiento obtenidas. Los ASA se aplicaron al subconjunto de entrenamiento de cada validación cruzada, el cual varía en cada iteración. Esto permitió calcular 30 Valores de Importancia (VI) para cada atributo de acuerdo con las métricas de cada ASA. Posteriormente, todos los VI se promediaron y se determinó así un Valor de Importancia Medio (VIM) para cada atributo. De esta manera se obtuvieron los rankings de atributos individuales. Este proceso se muestra gráficamente en la Fig. 1.

Para obtener un primer ranking combinado de atributos, los VIM de cada atributo -uno por cada ASA- se escalaron al rango entre 0 y 1. El escalado consistió en la obtención del cociente al dividir el VIM entre el número 100. Una vez escalados, se obtuvo el promedio de los VIM. El promedio obtenido es la base para la creación del primer ranking combinado. Este proceso se muestra gráficamente en la Fig 2.

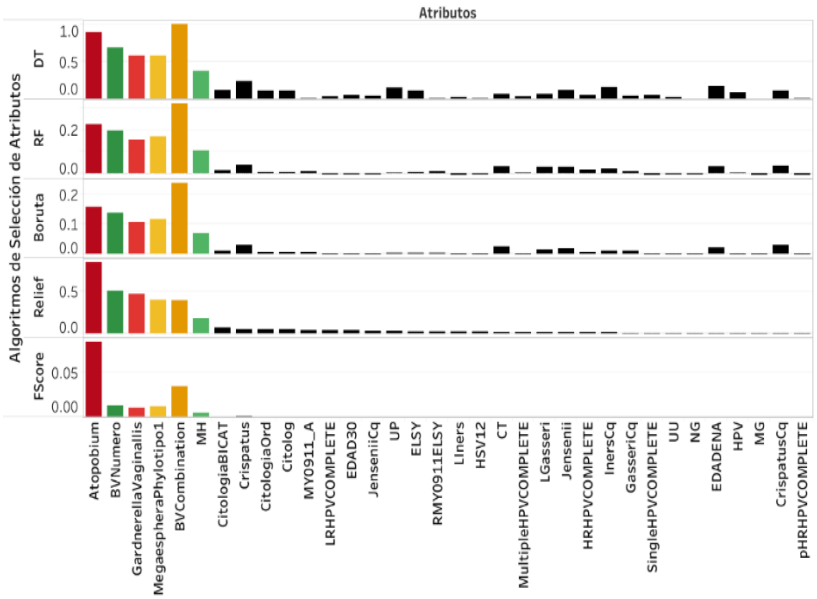

Fig. 4. Representación gráfica de los rankings individuales de atributos obtenidos mediante los algoritmos de selección de atributos (ASA). El cálculo de los valores de importancia medio (VIM) se representa por las barras verticales. Cada algoritmo emplea su propia escala de valores. DT: *decision tree*; RF: *random forests*.

Tabla 2. Primer ranking combinado de atributos. Se basa en el escalado y promedio de los cinco valores de importancia medios (VIM) obtenido por cada atributo en los rankings individuales. El VIM es el criterio de ordenamiento.

Atributos	**Promedio de VIM**
Atopobium	0.641467676
BVCombination	0.547947691
BVNumero	0.443284461
MegaespheraPhylotipo1	0.343827918
GardnerellaVaginallis	0.335289495
MH	0.21830713
Crispatus	0.064334192
EDADENA	0.040210846
CitologiaBICAT	0.037516153
InersCq	0.035282781
UP	0.033222477
CrispatusCq	0.030599123
Jensenii	0.030567158
Citolog	0.028379047
CitologiaOrd	0.028356376

Al tomar como base los rankings individuales, se realizó un análisis de frecuencias. Éste se basa en las posiciones conseguidas por los atributos dentro de cada ranking individual. Para esto, se estimaron la moda estadística y la frecuencia de la moda. El segundo ranking combinado de atributos se calculó a partir de este análisis. La moda estadística se considera como criterio de ordenamiento, ya que calcula el valor que ocurre con más frecuencia en un conjunto de observaciones. Este proceso se muestra gráficamente en la Fig. 3.

5. Resultados

Los resultados de los experimentos con Algoritmos de Selección de atributos se muestran a continuación. En la Fig. 4 se muestran los rankings individuales obtenidos. El eje X representa los atributos del conjunto de datos. El eje Y representa los ASA implementados. Los VIM obtenidos por cada atributo, es decir, el promedio de todos los valores de importancia a través de las 30 corridas de validación cruzada, se representa por las barras verticales en escala de cada método. Los seis atributos con mayor VIM se identifican con barras de color rojo, verde y amarillo en diferentes tonalidades.

Con el fin de crear el primer ranking combinado de atributos se consideraron los VIM obtenidos a partir de los rankings individuales. Los algoritmos emplean diferentes escalas de valores para mostrar sus resultados, por lo que se adaptaron a una misma escala. Para esto, los cinco VIM se escalaron y posteriormente se promediaron. El resultado se muestra en la Tabla 2. Los atributos se ordenan con base en el valor del VIM obtenido.

Por razones de espacio, se muestran solo los 15 atributos con los valores más altos. Posteriormente se realizó un análisis de frecuencia. Para crearlo, se calcularon las posiciones obtenidas por los atributos en cada uno de los rankings individuales. A partir de estas posiciones, se calculó la moda estadística y la frecuencia de la moda. Los resultados se muestran en la Tabla 3. Por razones de espacio, solo se muestran los primeros 15 atributos. Finalmente, los dos rankings combinados de atributos resultantes de los experimentos se comparan en la Tabla 4.

6. Conclusión

En este trabajo se identificaron los atributos más relevantes en un conjunto de datos microbiológico acerca de la vaginosis bacteriana. Para determinarlos, se implementaron cinco métodos de selección de atributos propios del área de aprendizaje automático en diversos experimentos. Con base en el análisis de los resultados obtenidos se crearon dos rankings combinados de atributos.

Es de importancia el considerar que entre ellos identifican al menos 10 atributos en común como los de mayor relevancia, sobre todo en las primeras seis posiciones.

Tabla 3. Segundo ranking combinado de atributos. Se basa en un análisis de frecuencias a partir de los rankings individuales. La moda se considera como criterio de ordenamiento. R: Relief, FS: fisher score, J48: decision tree, B: *boruta*, RF: *random forests*.

Atributos	Algoritmos de Selección de Atributos					Moda	Frecuencia de la moda
	R	FS	J48	B	RF		
BVNumero	2	2	1	1	1	1	3
Atopobium	1	1	2	2	2	2	3
GardnerellaVaginallis	3	3	3	3	3	3	5
MegaespheraPhylotipo1	4	4	4	4	4	4	5
BVCombination	5	5	5	5	5	5	5
MH	6	6	6	6	6	6	5
CitologiaBICAT	7	7	7	7	7	7	5
Crispatus	8	9	8	10	10	8	2
LRHPVCOMPLETE	12	20	13	8	8	8	2
ELSY	16	10	18	9	9	9	2
EDAD30	13	8	12	11	11	11	2
LIners	18	11	19	12	12	12	2
CitologiaOrd	10	13	9	13	13	13	3
RMY0911ELSY	17	14	14	20	19	14	2
Citolog	9	15	11	14	15	15	2

Tabla 4. Comparativa de los dos rankings combinados de atributos obtenidos. Por límites de espacio, sólo los primeros 15 atributos son mostrados. La etiqueta "a" denota que el atributo aparece en ambos rankings.

Ranking combinado 1	Ranking combinado 2
Atopobium[a]	BVNumero[a]
BVCombination[a]	Atopobium[a]
BVNumero[a]	GardnerellaVaginallis[a]
MegaespheraPhylotipo1[a]	MegaespheraPhylotipo1[a]
GardnerellaVaginallis[a]	BVCombination[a]
MH[a]	MH[a]
Crispatus[a]	CitologiaBICAT[a]
EDADENA	Crispatus[a]
CitologiaBICAT[a]	LRHPVCOMPLETE[b]
InersCq	ELSY[b]
UP	EDAD30[b]
CrispatusCq	Liners
Jensenii	CitologiaOrd[a]
Citolog[a]	RMY0911ELSY
CitologiaOrd[a]	Citolog[a]

Dichos atributos -considerados como biomarcadores de la vaginosis bacteriana- se analizarán para determinar su significancia biológica, ya que están potencialmente relacionados con la presencia o expresión de la infección.

Los rankings obtenidos son la base para planificar diversos experimentos con algoritmos clasificadores con el fin de definir los atributos o subconjunto de atributos óptimos en la detección de la VB. Esta es una investigación en curso que forma parte de un análisis exploratorio extenso. Más experimentos con otros métodos de selección de atributos y técnicas de clasificación están siendo investigadas.

Referencias

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)
2. Verstraelen, H., Swidsinski, A.: The biofilm in bacterial vaginosis: Implications for epidemiology, diagnosis and treatment: 2018 update. Curr. Opin. Infect. Dis. 32, pp. 38–42 (2019)
3. 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)
4. Anonymous: Biomarkers in risk assessment: Validity and Validation (2001)
5. World Health Orgnization: Biomarkers and risk assessment: concepts and principles. Environmental Health Criteria, 155, pp. 82 (1993)
6. Strimbu, K., Tavel, J.A.: What are biomarkers? Curr. Opin. HIV AIDS. 5, pp. 463–466 (2010)
7. Baker, Y.S., Agrawal, R., Foster, J.A., Beck, D., Dozier, G.: Detecting bacterial vaginosis using machine learning. In: Proceedings of the ACM Southeast Regional Conference on - ACM SE '14. pp. 1–4 (2014)
8. Beck, D., Foster, J.A.: Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis. BioData Min. 8, pp. 1–9 (2015)
9. Saqlain, S.M., Sher, M., Shah, F.A., Khan, I., Ashraf, M.U., Awais, M., Ghani, A.: Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. Knowl. Inf. Syst. 58, pp. 139–167 (2019)
10. Kumar, S.S., Shaikh, T.: Empirical evaluation of the performance of feature selection approaches on random forest. In: Int. Conf. Comput. Appl. ICCA. Pp. 227–231 (2017)
11. Khaire, U.M., Dhanalakshmi, R.: Stability of feature selection algorithm: A review. J. King Saud Univ. Comput. Inf. Sci. (2019)
12. Iguyon, I., Elisseeff, A.: An introduction to variable and feature selection (2003)
13. Duch, W., Grabczewski, K., Winiarski, T., Biesiada, J., Kachel, A.: Feature selection based on information theory, consistency and separability indices. In: ICONIP, Proc. 9th Int. Conf. Neural Inf. Process. Comput. Intell. E-Age. 4, pp. 1951–1955 (2002)
14. Bramer, M.: Principles of data mining. Springer, United Kingdom (2016)
15. Kuhn, M.: Building predictive models in R using the caret package. J. Stat. Softw. 28, pp. 1–26 (2008)
16. Medina-Merino, F., Ñique Chacón, I.: Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. pp. 165–190 (2017)
17. Romanski, P.: Package FSelector. http://cran.r-project.org/web/packages/FSelector/FSelector.pdf (2013)
18. Guerrero, J.A.: El problema de la dimensionalidad. Rev. Estadística y Soc. 1, pp. 22–24 (2016)

19. Kursa, M.B., Rudnicki, W.R.: Feature selection with the Boruta package. J. Stat. Softw. 36, pp. 1–13 (2010)
20. Robnik-Sikonja, M., Kononenko, F.: Theoretical and empirical analysis of relieff and rreliefF. Mach. Learn, 53, pp. 23–69 (2003)
21. Gu, Q., Li, Z., Han, J.: Generalized fisher score for feature selection. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI, pp. 266–273 (2011)
22. Valizade-Hasanloei, M.A., Sheikhpour, R., Sarram, M.A., Sheikhpour, E., Sharifi, H.: A combined fisher and laplacian score for feature selection in QSAR based drug design using compounds with known and unknown activities. J. Comput. Aided. Mol. Des., 32, pp. 375–384 (2018)
23. Koul, A., Becchio, C., Cavallo, A.: PredPsych: A toolbox for predictive machine learning-based approach in experimental psychology research. Behav. Res. Methods, 50, pp. 1657–1672 (2018)