42

Fecha de presentación: abril, 2020
Fecha de aceptación: junio, 2020
Fecha de publicación: julio, 2020

PROPUESTA DE ALGORITMOS

PARA LA REDUCCIÓN DE ESPACIO MUESTRAL EN LA CEFALOSPORINA

PROPOSED ALGORITHMS FOR THE REDUCTION OF SAMPLE SPACE ON CEFALOSPORIN

Tonysé de la Rosa Martín[1]
E-mail: tdelarosa@umet.edu.ec
ORCID: https://orcid.org/0000-0002-0881-6034
[1] Universidad Metropolitana. Ecuador.

Cita sugerida (APA, sexta edición)

De la Rosa Martín, T. (2020). Propuesta de algoritmos para la reducción de espacio muestral en la cefalosporina. *Revista Universidad y Sociedad*, 12(4), 316-324.

RESUMEN

En el presente trabajo se presentan la implementación y evaluación de varios métodos de reduc-ción de la dimensionalidad basados en técnicas de inteligencia artificial, y aborda uno de sus complejos problemas, como es el identificar y reducir un conjunto representativo de atributos para así contribuir al mejoramiento de los modelos de clasificación y predicción. La búsqueda de subconjuntos óptimos de atributos para la clasificación de conjuntos de datos presenta el incon-veniente de su complejidad temporal. Se implementaron procedimientos de búsqueda por algo-ritmos genéticos, enfriamiento simulado, búsqueda secuencial y una hibridación entre este último y algoritmos genéticos, con tal de alcanzar mayor robustez y eficiencia. Se implementan además varias medidas de asociación entre subconjuntos variables, a partir de conceptos de la estadística clásica o tomadas de la Teoría de la Información de Shannon. En todos los casos experimentados se reduce el espacio muestral en más del 65%. Los mejores resultados se alcanzan con el algo-ritmo Enfriamiento Simulado, empleando Máquinas de Soporte Vectorial como clasificador. Todos estos procedimientos de búsqueda presentan una complejidad temporal de orden polinomial, esto demuestra la viabilidad práctica en costo y recursos computacionales de cada procedimiento im-plementado.

Palabras clave: Atributos, clasificación, dimensionalidad, inteligencia artificial, predicción, reducción.

ABSTRACT

This paper presents the implementation and evaluation of various methods of dimensionality re-duction based on artificial intelligence techniques, and addresses one of their complex problems, such as identifying and reducing a representative set of attributes to assist in the improvement of the classification and prediction models. The quest for optimal subsets of attributes for the classi-fication of data sets have the disadvantage that its time complexity. Search procedures were im-plemented by genetic algorithms, simulated cooling, sequential search and a hybrid between this and genetic algorithms, so to achieve greater robustness and efficiency. It also implemented sev-eral measures of association between variable subsets, based on concepts borrowed from classical statistical theory of Shannon Information. In all cases tested the sample space is reduced by more than 65%. The best results are achieved through the Simulated Annealing algorithm using support vector machines classifier. All these search procedures present a polynomial time complexity of order, this demonstrates the practical feasibility and cost of each procedure computing resources deployed.

Keywords: Artificial intelligence, classification, dimensionality, reduction, prediction, sub-attributes.

INTRODUCCIÓN

Los avances en el sector de la bioinformática, junto con los extraordinarios progresos de la fisiología, la bioquímica, la medicina y las técnicas de computación han promovido una revolución en el ámbito del diseño y producción de fármacos. Entre las muchas funciones de la farmacología la más importante es la creación de medicamentos de alta calidad para la preservación de la salud de los seres humanos, de ahí que los medicamentos son la base para casi cualquier programa de salud pública intencionado a reducir la morbilidad o mortalidad.

La predicción de la actividad biológica de compuestos químicos es hoy día un objetivo principal dentro de la Industria Médico Farmacéutica Mundial. El alto costo del proceso de investigación - desarrollo de nuevos fármacos, ha obligado a este sector económico a adoptar la estrategia del uso de técnicas de la computación y la informática para acelerar el proceso y disminuir los costos. En los últimos años, la industria farmacéutica ha reorientado sus investigaciones y prestado más atención a aquellos métodos que permitan una selección racional o el diseño de nuevos compuestos con propiedades deseadas.

Esta situación ha obligado al país a pensar y crear estrategias para solucionar el problema de la creación de medicamentos de alta calidad, que ayuden al pueblo cubano a tener una vida más duradera y para aquellos pacientes que no tengan cura efectiva para su enfermedad puedan convivir con ella más tiempo. También la creación de medicamentos representaría una ayuda importante en concepto de bienes monetarios para el país al exportarlos y con ellos se ayudaría a otras naciones amigas necesitadas de estos medicamentos que a tan alto precio se obtienen en el mercado mundial.

Por lo anteriormente expuesto, el presente trabajo se encuentra enmarcado dentro del proyecto de investigación científica conjunta entre docentes de la Universidad de las Ciencias Informáticas (UCI) y la Universidad Metropolitana del Ecuador (UMET).

Actualmente existen bases de datos de regular tamaño formada por moléculas y sus descriptores asociados. Estos son utilizados por los métodos de inteligencia artificial implementados en disímiles plataformas para la predicción de actividad biológica asociando esta a la estructura química. Dichos métodos realizan la predicción utilizando una cantidad elevada de descriptores topológicos, topográficos e híbridos, aunque solo algunos de ellos aportan información útil para el establecimiento de los modelos. La generalidad de esos descriptores parte de formulismos que se basan en la matriz de conectividad de los vértices o aristas del grafo químico por lo que se encuentra elevada redundancia en la información que ellos contienen. Otro problema es el elevado consumo de los recursos de cómputo cuando se necesita procesar una cifra tan elevada de datos. Por lo tanto, se hace necesario contribuir a la reducción del espacio muestral de descriptores, con el fin de eliminar gran parte de la redundancia de información en la base de datos y para mejorar la eficiencia y costo computacional del establecimiento de los modelos y la realización de las predicciones.

Son diferentes los procedimientos que se emplean en la actualidad para la reducción de la dimensión en una muestra dada. Entre los más modernos se destacan las técnicas de inteligencia artificial, que se emplean solas o combinadas con técnicas clásicas de la estadística avanzada. Mediante las cuales se determina la presencia de variables irrelevantes o redundantes.

El objetivo del artículo es proponer algoritmos de búsqueda y evaluación para la reducción del espacio muestral en la Cefalosporina.

DESARROLLO

Los algoritmos genéticos son procesos de búsqueda basados en los principios de la selección y la evolución natural. Las posibles soluciones a un problema son codificadas en forma de cadenas binarias, y la búsqueda se inicia con una población de posibles soluciones generadas aleatoriamente (Holland, 1975).

Los algoritmos genéticos son algoritmos matemáticos altamente paralelos que transforman un conjunto de objetos matemáticos individuales con respecto al tiempo. Estos usan operaciones modeladas de acuerdo con el principio Darwiniano de reproducción y supervivencia del más apto y tras haberse presentado de forma natural una serie de operaciones genéticas de entre las que destaca la recombinación sexual. Cada uno de estos objetos matemáticos suele ser una cadena de caracteres (letras o números) de longitud fija que se ajusta al modelo de las cadenas de cromosomas, y se les asocia con una cierta función matemática que refleja su aptitud.

Los pasos para construir un algoritmo genético, siguiendo la propuesta de pseudocódigo, son:

- Diseñar una representación.
- Decidir cómo inicializar una población.
- Diseñar una forma de evaluar un individuo.
- Diseñar un operador de mutación adecuado.
- Diseñar un operador de cruce adecuado.

- Decidir cómo seleccionar los individuos para ser padres.
- Decidir cómo reemplazar a los individuos.
- Decidir la condición de parada.

Definiéndose como parámetros fundamentales a introducir: número de población y generaciones, probabilidad de mutación y cruce.

Se debe garantizar que los mejores individuos tengan una mayor posibilidad de ser padres (reproducirse) frente a los individuos menos buenos. Se debe ser cuidadoso para dar una oportunidad de reproducirse a los individuos menos buenos. Estos pueden incluir material genético útil en el proceso de reproducción. Esta idea define la presión selectiva que determina en qué grado la reproducción está dirigida por los mejores individuos. Existen varios esquemas de selección, dentro de los más empleados se encuentran:

» Selección por Torneo (TS): escoge al individuo de mejor aptitud de entre N individuos seleccionados aleatoriamente (N = 2, 3, …).

» Orden Lineal (LR): la población se ordena en función de su aptitud y se asocia una probabilidad de selección a cada individuo que depende de su orden.

» Selección Aleatoria (RS): un padre lo escoge aleatoriamente, para el otro selecciona N padres y escoge el más lejano al primer (N = 3, 5, …). Está orientado a generar diversidad.

» Selección por Ruleta: se asigna una probabilidad de selección proporcional al valor de aptitud del cromosoma. Siendo este último el esquema de selección empleado, a continuación, se muestra el seudocódigo de este.

function GeneticSearch(eval)

t := 0;

Inicializar *P(t)*;

Evaluar *P(t)*;

Escalar *P(t)*;

Obtener_Mejor_Individuo *P(t)*;

Para t := 1 **hasta** cantidad Generaciones(max) **hacer** :

Seleccionar *P(t)* desde *P(t – 1)*;

Cruzar *P(t)*;

Mutar *P(t)*;

Evaluar *P(t)*;

Escalar *P(t)*;

converge: = Obtener_Mejor_Individuo *P(t)*; Estadisticas *P(t)* ;

Si (i = max) **or** (converge = true) **entonces**

break;

fin Si

fin Para

atributos: = Listar(Mejor Individuo);

return atributos;

fin function

Donde, *P(t)* es la población en la iteración ***t***.

El enfriamiento simulado (Simulated Annealing) (Kirpatrick, 1983.) es una metaheurística para problemas de optimización global que se basa en conceptos de la mecánica estadística y es una generalización del Método de Monte Carlo. Fue propuesto por primera vez por Metrópolis (Langley, 1994) y usado en optimización combinatoria por Kirkpatrick (1983) Este método heurístico se basa en los conceptos descritos originalmente por el proceso físico sufrido por un sólido al ser sometido a un baño térmico.

Se sabe en ingeniería, que una manera de encontrar los estados de energía de sistemas complejos, tales como sólidos, consiste en utilizar la técnica de enfriamiento, en la que el sistema se calienta primero a una temperatura en la que sus granos deformados recristalizan para producir nuevos granos; luego se enfría suavemente y de esta manera, cada vez que se baja la temperatura, las partículas se reacomodan en estados de más baja energía; hasta que se obtiene un sólido con sus partículas acomodadas conforme a una estructura de cristal (estado fundamental). En la fase de enfriamiento, para cada valor de la temperatura, debe permitirse que el sistema alcance su equilibrio térmico (Kirpatrick, 1983).

De forma análoga, en el algoritmo de enfriamiento simulado los estados del sistema corresponden a las soluciones del problema, la energía de los estados a los criterios de evaluación de la calidad de la solución (generalmente se utiliza la función objetivo), el estado fundamental a la solución óptima del problema, los estados metaestables a los óptimos locales, y la temperatura a una variable de control. *"El éxito del Enfriamiento Simulado se basa en la escogencia de una buena temperatura inicial y una adecuada velocidad de enfriamiento".* (Kirpatrick, 1983)

"La característica principal de este algoritmo es que al buscar una nueva solución Sn+1 dada una solución Sn, acepta en ocasiones una de inferior calidad a la de Sn por medio de una función probabilística la cual depende del parámetro variable de temperatura y de la calidad

ofrecida por las dos soluciones Sn y Sn+1. Mientras más bajo sea el parámetro de temperatura, menor será la probabilidad de aceptar una solución peor, y viceversa". (Kirpatrick, 1983)

El Enfriamiento Simulado es una poderosa herramienta de búsqueda estocástica que se ha hecho muy popular dado el amplio espectro de problemas que puede resolver. En particular en el área de la optimización combinatoria y la selección de variables o de características. A continuación, se muestra la estructura de este.

function SimmulatedAnnealing (T0, Tf, k, Vecinos) T = T0

Sactual = Genera solución aleatoria;

Mientras T >= Tf **hacer:**

Para i en nVecinos (T) **hacer:**

Scandidata = Genera un vecino (Sactual)

λ = coste (Scandidata) – coste (Sactual)

Si U (0, 1) < e–λ/T or λ < 0 **entonces:**

Sactual = Scandidata;

fin Si

fin Para

Estadisticas();

T = k*(T);

fin Mientras

atributos: = Listar(SActu al);

return atributos;

fin funcion

Donde:

Sactual: solucion actual Scandidata: solucion candidata T0 es la temperatura inicial

Tf: la temperatura final

k: es el coeficiente de enfriamiento elegido

nVecinos(T): el número de vecinos generados en cada ciclo según T

U (0,1): es un generador de números aleatorios uniformemente distribuidos.

Se decidió utilizar como herramienta para apoyar en la investigación a Weka, que es un software que posee una colección extensa de algoritmos de máquinas de conocimiento, conteniendo las herramientas necesarias para la realización de minería de datos, transformaciones necesarias en los datos, tareas de clasificación, regresión, agrupamiento, asociación y visualización.

Librerías de las Máquinas de Soporte Vectorial (LibSVM)

Software integrado para la clasificación, regresión, estimación de la distribución de los datos y soporta la clasificación multiclase empleando las MSV. Dentro de sus prestaciones se encuentran:

- Diferentes formularios de MSV.
- Validación para la selección de los modelos.
- Estimaciones Probabilísticas.

En primer lugar, el entrenamiento de datos es separado en varios segmentos. Secuencialmente un segmento está considerado como el conjunto de validación y el resto son para el entrenamiento (Langley, 1994).

Los tipos de SVM son:

- C-SVC.
- nu-SVC.
- one-class SVM.
- épsilon-SVR.
- nu-SVRT.

Los Tipos de Kernel son:

- Lineal: K(u, v) = u' * v.
- Polinomial: K (u, v) = (gamma * u' * v + coef0) ^degree.
- Función de Base Radial (RBF): K (u, v) = exp (-gamma * |u – v|^2).
- Sigmoidal: K (u, v) = tanh (gamma * u * v + coef0).

Esta librería brinda la posibilidad de integrarse al software Weka, permitiendo una mejor interpretación de los resultados y usabilidad (Langley, 1994).

Los algoritmos genéticos son por construcción métodos de búsqueda ciega, el proceso de optimizar es una caja negra que asigna a cada individuo una aptitud. Esta opacidad en la medida que proporciona un algoritmo de propósito general y permite realizar la búsqueda con información mínima, tienen la contrapartida de que son intrínsecamente débiles. Como la debilidad es intrínseca, cualquier intento de mejora cualitativa implica incorporarle al algoritmo un mecanismo de explotación de la solución, después de explorar el espacio de búsqueda.

La idea general de esta técnica de hibridación consiste en utilizar el algoritmo genético para realizar la búsqueda global y encargar la búsqueda local greedy (secuencial) para explotar la solución. Para esto fue necesario llevar

a cabo la hibridación de forma modular, incorporando el procedimiento de búsqueda secuencial como un operador más del algoritmo genético.

El procedimiento de búsqueda local toma como punto de partida las soluciones brindadas por el algoritmo genético en cada generación después de aplicarles los operadores probabilísticos, así el método de búsqueda secuencial explota los estados vecinos que generan estas soluciones globales considerando solo aquellas que sean mejores.

A todos los algoritmos de búsqueda planteados anteriormente se le incorporó un mecanismo de almacenamiento de las mejores soluciones durante su ejecución. O sea, se implementó un proceso de almacenamiento de aquellas soluciones cuya aptitud fuera superior a la aptitud promedio del conjunto de soluciones almacenadas. Para lograr mayor eficiencia en consultas de selección e inserción al conjunto de soluciones, estas se almacenan en Tablas Hash. Este mecanismo de estadística simple permite obtener un subconjunto de soluciones (ordenadas por aptitud) finales al problema, permitiéndole al usuario escoger cualquiera de estas (Liu, 1997).

Los métodos de evaluación antes referidos necesitan una medida o criterio de evaluación por lo que fueron implementadas las siguientes medidas de evaluación:

- Para las variables individuales se implementaron las de CHI2, Correlación de Pearson, Incertidumbre Simétrica y la de Gain Ratio.
- Para subconjuntos fueron implementadas las de Correlación de subconjuntos y la de consistencia.

A continuación, se muestra un diagrama de flujo (Figura 1) donde se explica de modo general la funcionalidad de reducción de espacio muestral y la de ordenamiento de las características independientes de acuerdo con la relevancia que estas presentan con respecto a la clase o característica dependiente (actividad biológica) empleando la Metodología de Filtro (Kohavi, 1997).

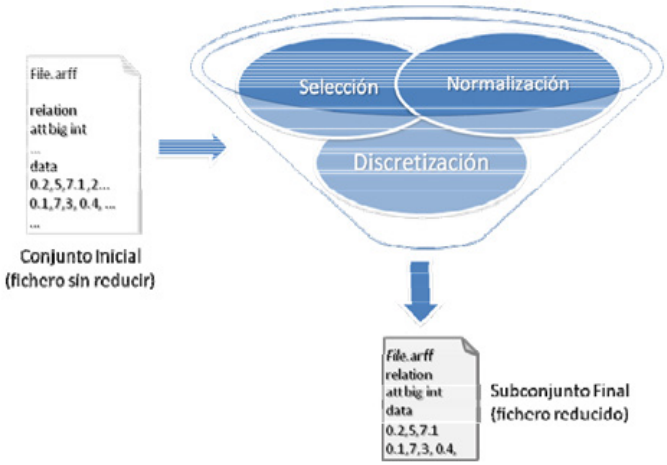

Figura 1. Diagrama general del flujo de eventos.

De manera general este diagrama muestra como son cargados al sistema los ficheros de tipo. arff y por los procesos que pasan para ser reducidos u ordenados en dependencia de la orden que el administrador le pase al sistema. Al final del proceso se entrega un fichero en el que se encuentran los datos reducidos, este fichero es de tipo. arff.

De manera más específica el siguiente diagrama de flujo (Figura 2) muestra el proceso de la reducción de las características para tener una visión ampliada del proceso de reducción y lograr un mejor entendimiento de la solución del problema.

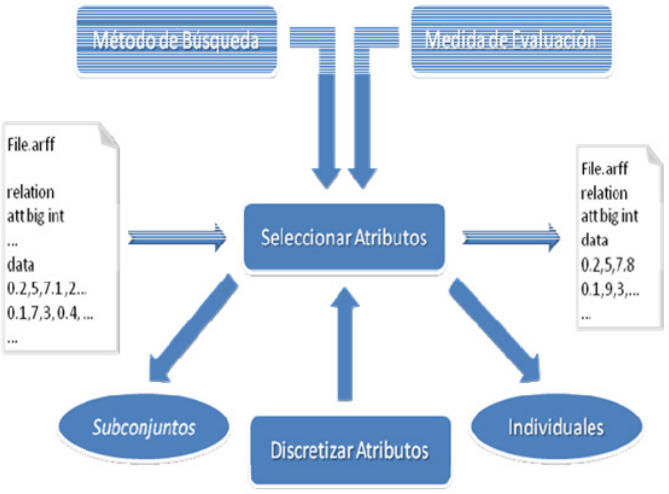

Figura 2. Diagrama de flujo para el proceso de selección de atributos.

En el diagrama anterior se describe como el fichero de entrada de tipo. arff es incorporado desde una base de datos al sistema, con el objetivo de realizarle una

reducción de sus atributos. El administrador selecciona el método de búsqueda y la medida de evaluación, en caso de que este fichero presente datos con valores continuos y la variable dependiente o clase posea valores discretos; entonces se procede a la discretización de los datos. Este proceso de discretización no es más que convertir los valores continuos a valores de tipo discreto; buscando así uniformidad entre los datos de las variables independientes y la variable dependiente (clase).

Las medidas de evaluación son utilizadas para la selección de características de tipo individual o de subconjuntos, de acuerdo con la que se haya seleccionado se realiza la reducción correspondiente.

Al terminarse la selección los nuevos datos son guardados en un fichero nuevo de tipo. arff y este es guardado en una base de datos.

Implícito dentro del proceso de selección se encuentra el proceso de ordenamiento de los atributos, el cual se explica de la siguiente manera y basándose en el diagrama anterior:

- El administrador selecciona el método de búsqueda y la medida de evaluación; en caso de que este fichero presente datos con valores continuos y la variable dependiente o clase posea valores discretos, entonces se procede a la discretización de los datos.
- Al finalizar este proceso de ordenamiento en vez de crearse un fichero. arff se crea un fichero .txt con los datos ordenados de acuerdo con la relevancia que estos presentan con la variable dependiente o clase. Este proceso es muy importante pues con sus resultados se pueden realizar estudios estadísticos por parte de los especialistas en la parte de predicción y clasificación para de alguna manera tener una visión de la relación que tienen los atributos o variables independientes con respecto a la variable dependiente o clase.

Para comprobar la eficiencia y rapidez de los métodos implementados se tomaron muestras de datos reales de una familia de cefalosporinas (34). A continuación, se muestran las características de la familia:

Instancias: 104

Número inicial de variables: 180

Esta muestra mantiene dentro de sus características principales que todos sus compuestos son reportados como activos en los ensayos realizados, por lo que este estudio va encaminado a determinar, cuáles son las características estructurales distintivas dentro de los activos. Para emplear un criterio de clasificación sobre esta muestra se tomó la variable dependiente perteneciente a la actividad biológica y se consideraron como activos aquellos compuestos cuyo valor fuese mayor o igual que el promedio e inactivo en caso contrario.

Para la clasificación de las muestras se emplean las máquinas de soporte vectorial C-SVC y nu-SVC para la clasificación, pertenecientes ambas a la librería libSVM en su versión 6.8. Dicha librería posee varias funciones Kernels que le permiten redimensionar los valores de entrada:

- De base radial (RBF).
- Polinomial.
- Lineal.
- Sinusoidal.

De estas funciones y partiendo de las características fundamentales de la muestra se emplearon la RBF y la Sinusoidal debido a que las mismas poseen más de 50 espacios de nueva dimensión lo que les permite encontrar los mejores valores de clasificación, además que soportan la no linealidad entre los datos.

Existen diferentes parámetros que evalúan la eficiencia del clasificador validando así la calidad del modelo. Las medidas más conocidas para evaluar la clasificación están basadas en la matriz de confusión que se obtiene cuando se prueba el clasificador en el conjunto de datos del entrenamiento como se muestra en la Tabla 1.

Tabla 1. Matriz de Confusión.

	Positivos	**Negativos**
Positivos	Verdaderos Positivos (VP)	Falsos Positivos (FP)
Negativos	Falsos Negativos (FN)	Verdaderos Negativos (VN)

Se consideran verdaderos positivos aquellos compuestos cuyos valores de actividad son positivos bien clasificados mientras que los verdaderos negativos son aquellos compuestos cuyos valores negativos de actividad son igualmente bien clasificados. Mediante un análisis contrario, se definen los falsos positivos y los falsos negativos.

Otra forma de evaluar el rendimiento de un clasificador es por las curvas ROC. En esta curva se representa el valor de la razón de VP contra la razón de FP, mediante la variación del umbral de decisión. Se denomina umbral de decisión a aquel que decide si una instancia x, a partir del vector de salida del clasificador, pertenece o no a cada una de las clases. Esta última y la precisión serán los criterios empleados para evaluar el clasificador.

La MSV es una técnica de aprendizaje supervisado en la que los parámetros C (Costo), nu (Valor empleado por nu-SVC) y alfa (Valor empleado por la función kernel) son fundamentales para garantizar el entrenamiento de esta, evitando así la memorización o sobre entrenamiento de las máquinas. El método para determinar las muestras de entrenamiento y prueba que se empleó es la validación cruzada (cross validation). La cantidad de subconjuntos destinados para la misma, según la cantidad de instancias presentadas por la muestra, alcanzaron valores entre 2 y 100, mientras que nu y gamma tomaron valores entre 0.5 y 0.9, el costo se fijó en 100. Esos resultados se muestran en la Tabla 2.

Tabla 2. Resultados de la clasificación para costo 100, nu 0.9 y ganma 0.5.

Valor de Validación	MSV	%clasificación
2	C-SVC nu-SVC	57 57
5	C-SVC nu-SVC	57 57
10	C-SVC nu-SVC	62 61
20	C-SVC nu-SVC	57 57
100	C-SVC nu-SVC	56 56

Los mejores resultados se alcanzaron con la validación cruzada en 10, los resultados se muestran en la Tabla 3, encontrándose además los valores de la exactitud, precisión y Área ROC.

Tabla 3. Calidad de la clasificación para la muestra completa.

MSV	Exactitud	Precisión	Área ROC	%clasificación
nu-SVC C-SVC	0.50 0.50	0.702 0.69	0.615 0.60	61.53 60.71

Según los resultados de la clasificación, se infiere que en la muestra se encuentran descriptores cuya información no es significativa para la descripción. Dentro de las ventajas de la reducción de variables se encuentra la mejora de la eficiencia del clasificador, y para ello se emplean los algoritmos propuestos en el capítulo anterior (AG, ES, AH):

Para evaluar la calidad de las variables seleccionadas se emplearon las medidas de evaluación basadas en Consistencia y la correlación de Pearson o CFS para evaluar subconjuntos de atributos. Cada algoritmo posee parámetros que son influyentes dentro de los resultados de este, en el caso de AG y AH sus resultados dependen de la Probabilidad de Cruzamiento (Pc), la cual permite el cruzamiento de dos individuos para lograr uno con mejores potencialidades que los dos anteriores, por lo que sus valores deben permanecer por encima de 0.50. Para estudiar el comportamiento de estos con respecto a la generación de los modelos se utilizaron los valores de 0.6, 0.7, 0.8 y 0.9. Otro de los parámetros que influyen en estos algoritmos es la Probabilidad de Mutación (Pm), la cual permite la creación de nuevos individuos a partir de las características de los anteriores. No obstante, la Pm puede dar lugar a combinaciones de variables que generen malas soluciones por correlaciones casuales. Estas van a provocar que los algoritmos pierdan el sentido de la búsqueda al brindar respuestas ajenas a la fenomenología estudiada. Algunos autores (referencia) han propuesto minimizar el valor de Pm por debajo de 0.1. Otros han propuesto la simplificación del AG por eliminación de la Pm (referencia). En la presente investigación se definió el valor de Pm como 0.01.

Como posibilidad dentro de los algoritmos presentados se encuentra la generación de un número de posibles combinaciones de solución, fijándose en este trabajo un máximo de las diez mejores soluciones posibles. A partir de estas condiciones se realizó la selección de variables especificando el método de búsqueda, la Pc, la medida de evaluación a valor de Pm constante. Para cada caso se obtuvieron los valores correspondientes de cantidad final de variables, porciento de reducción, valor de ajuste (valor de los parámetros evaluados en la medida de evaluación). Esos resultados se muestran en la tabla 4.

Tabla 4. Resultados de la selección de variables.

Método de búsqueda	Medida de Evaluación	Cant. de Variables	% Reducción	Pc	Valor de Ajuste
AG	CFS con PL	54 46 9 15	70 75 95 91	0.6 0.7 0.8 0.9	0.84 0.70 0.74 0.78
AG	Consistencia	57 71 50 74	57 71 50 74	0.6 0.7 0.8 0.9	0.91 0.92 0.91 0.91
AH	CFS con PL	13 12 9 13	92 91 95 92	0.6 0.7 0.8 0.9	0.34 0.34 0.34 0.34
AH	Consistencia	9 10 10 10	95 94 94 94	0.6 0.7 0.8 0.9	0.92 0.92 0.92 0.92

Para validar la calidad de los modelos reducidos se le aplicará el clasificador Máquinas de Soporte Vectorial con los parámetros definidos en la Tabla 5. Los resultados para el Kernel y los dos tipos de máquinas de soporte vectorial se muestran a continuación, donde la identificación del modelo viene dada por: método de búsqueda, más la medida de evaluación y la probabilidad.

Tabla 5 Clasificación de la selección por AG.

Modelo	No. Variables	MSV	Presc.	Area ROC	% Clasificación
A G - CSF-0.6	54	C-SCV nu-SCV	0.98 0.934	0.98 0.933	98 93
A G - CSF-0.7	46	C-SCV nu-SCV	0.915 0.898	0.92 0.894	91 89
A G - CSF-0.8	9	C-SCV nu-SCV	0.78 0.838	0.779 0.837	78 84
A G - CSF-0.9	15	C-SCV nu-SCV	0.929 0.885	0.923 0.884	92 89
AG-Cons-0.6	54	C-SCV nu-SCV	0.952 0.904	0.952 0.904	95 90
AG-Cons-0.6	46	C-SCV nu-SCV	0.942 0.914	0.942 0.913	94 91
AG-Cons-0.6	9	C-SCV nu-SCV	0.894 0.825	0.885 0.837	89 84
AG-Cons-0.6	15	C-SCV nu-SCV	0.924 0.875	0.923 0.875	92 88

Los resultados de la clasificación se comportan en este algoritmo entre un 78 y 98% de clasificación correcta, de la misma manera que la precisión y el área debajo de la curva se mantienen entre rangos de valores que permiten validar la eficiencia del clasificador empleado para cada uno de los modelos. De los dos tipos de máquinas de soporte vectorial empleadas, los mejores resultados se obtienen con C-SVC. Según los valores, el mejor modelo es el 1, al alcanzar un 98 % de buena clasificación; sin embargo, este modelo cuenta con 54 variables; mientras que el modelo 4, con solo 15 variables, alcanza un 92 % para ambas medidas de evaluación y, teniendo como precedente el principio de parsimonia, este es el mejor de los modelos creados por los Algoritmos Genéticos demostrándose que mientas mayor es la Pc y menor la Pm se tienen mejores resultados. Para este algoritmo, el promedio de reducción de variables por ambas técnicas fue de 31 variables. Basado en los mismos criterios, el AH se comporta de la siguiente manera.

Tabla 6. Clasificación de la selección por AH.

Modelo	No. Variables	MSV	Presc.	Area ROC	% Clasificación
A H - CSF-0.6	13	C-SCV nu-SCV	0.706 0.799	0.702 0.798	70 80
A H - CSF-0.7	12	C-SCV nu-SCV	0.799 0.799	0.798 0.798	70 80
A H - CSF-0.8	9	C-SCV nu-SCV	0.714 0.799	0.712 0.798	71 80
A H - CSF-0.9	13	C-SCV nu-SCV	0.799 0.799	0.798 0.798	70 80
AH - Cons-0.6	9	C-SCV nu-SCV	0.695 0.827	0.692 0.808	70 83
AH - Cons-0.6	10	C-SCV nu-SCV	0.695 0.808	0.692 0.808	70 81
AH - Cons-0.6	10	C-SCV nu-SCV	0.695 0.808	0.692 0.808	70 81
AH - Cons-0.6	10	C-SCV nu-SCV	0.695 0.808	0.692 0.808	70 81

Los valores de los resultados se encuentran entre un 70 y 81 %, siendo la mejor máquina de soporte vectorial nu-SVC quien mantiene todos sus valores entre un 80 y 83% de los mismos. Se propone como mejor modelo el 5, quien con 9 variables alcanza los mejores valores de clasificación. Los resultados arrojados por este algoritmo demuestran que su funcionamiento no es óptimo para muestras donde no exista linealidad entre sus datos evidenciándose en los resultados obtenidos con la medida basada en consistencia. Sin embargo, este algoritmo logra reducir aún más la muestra que el anterior siendo el promedio de variables 10.

El método de búsqueda ES, es una técnica que para explorar todo el espacio de búsqueda se basa en una probabilidad, donde las dos condiciones fundamentales para la realización de una buena exploración son:

- Definir una temperatura inicial alta para garantizar que se cubra todo el espacio de búsqueda.
- Mantener una temperatura final baja. Dentro de la probabilidad de moverse o no hacia una mejor o peor solución juega un papel fundamental el valor de alfa. Se reporta que los valores más acertados son 0.7, 0.8 y 0.9.

Los que fueron los empleados en el trabajo. Los resultados de la reducción de variables se presentan en la tabla 7.

Tabla 7. Resultados de la selección de variables utilizando enfriamiento simulado.

Método de búsqueda	Medida de Evaluación	Cant. de variables	% de reduccion	alfa	Valor de ajuste
ES	CFS	20 9 1	88.8 95 99.4	0.7	0.14 0.18 0.18
ES	Cons	69 74 64	61.6 58.8 64.4	0.8	0.92 0.92 0.92

A los modelos generados se le aplicó el clasificador con los mismos parámetros obteniendo los resultados significativos, en la Tabla 8 se describen los resultados teniendo en el nombre del modelo el algoritmo seguido de la medida de evaluación y la variación de alfa.

Tabla 8. Clasificación de la selección de variables con ES.

Modelo	No. Variables	MSV	Presc.	Area ROC	% Clasificación
ES-CSF-0.7	20	C-SCV nu-SCV	1 0.991	1 0.99	100 99.3
ES-CSF-0.8	46	C-SCV nu-SCV	0.933 0.904	0.933 0.904	91 89
ES-CSF-0.9	9	C-SCV nu-SCV	0.751 0.751	0.751 0.751	78 84
ES-Cons-0.7	54	C-SCV nu-SCV	1 1	1 1	100 100
ES-Cons-0.8	46	C-SCV nu-SCV	1 1	1 1	100 100
ES-Cons-0.9	9	C-SCV nu-SCV	1 1	1 1	100 100

Los resultados de la clasificación se comportan con ES entre un 75 y 100% de clasificación correcta, de la misma manera que la precisión y el área debajo de la curva se mantienen entre rangos de valores que permiten validar la eficiencia del clasificador empleado para cada uno de los modelos.

Aunque los modelos 4, 5 y 6 poseen 100% de clasificación en ambas máquinas de soporte vectorial la cantidad de variables que poseen los hace imprácticos; mientras que los modelos 1 y 2 poseen 20 y 9 variables respectivamente y tienen 100 y 93 % de clasificación. El promedio de reducción de variables es de 16.

Conclusiones

En el estudio se propone como método de selección de variables el Enfriamiento Simulado, teniendo en cuenta que, para las Máquinas de Soporte Vectorial como clasificador, brinda los mejores modelos en las muestras estudiadas.

Se propone e implementa un novedoso algoritmo híbrido que combina algoritmo genético con algoritmos de búsqueda secuencial. Con este método se logra la máxima reducción de dimensionalidad, lo cual no implica obtener los mejores modelos. Se demostró así mismo que los algoritmos de complejidad polinomial evaluados (entre ellos el híbrido) son menos consumidores de tiempo de cómputo que los algoritmos de complejidad exponencial reduciéndose el tiempo de ejecución de horas a minutos.

Desde el punto de vista informático, se implementaron y evaluaron procedimientos de búsqueda que utilizan algoritmo genético y enfriamiento simulado para la reducción de la dimensionalidad de los datos; así como las medidas de evaluación basadas en la correlación, consistencia, y Teoría de la Información de Shannon. Con Algoritmo Genético se obtuvo una reducción entre 77 y 90% de la muestra original, mientras que con Enfriamiento Simulado y Algoritmo Híbrido la reducción estuvo entre 95 y 99 %.

Se proponen modelos de clasificación de antibióticos del tipo de las cefalosporinas y de compuestos activos frente a cáncer de próstata empleando Máquinas de Soporte Vectorial como clasificador utilizando Enfriamiento Simulado como método de reducción de dimensionalidad para las medidas de evaluación propuestas.

REFERENCIAS BIBLIOGRÁFICAS

Holland, J. H. (1975). ***Adaptation in Natural and Artificial Systems.*** University of Michigan Press.

Kirpatrick, S. G. (1983). Optimization by Simulated Annealing. Science, 220(4598), 671-680.

Kohavi, R. J. (1997). ***Wrappers for Feature Subset Selection. Artificial Intelligence, 97***, 273-324.

Langley, P. (1994). ***Selection of Relevant Features in Machine Learning.*** Procedings of the AAAI Fall Symposium on Relevance. AAAI Press.

Liu, H. D. (1997). Feature selection for Classification. ***Intelligent Data Analysis***, 1(1-4), 131-156.