ANALES DE LA
Academia de Ciencias de Cuba

CIENCIAS TÉCNICAS

Artículo original de investigación

Aportes teóricos y prácticos de los modelos probabilísticos en la solución de problemas de optimización reales

Julio Madera Quintana [1*] https://orcid.org/0000-0001-5551-690X
Yoan Martínez López [1] https://orcid.org/0000-0002-1950-567X
Gaafar Sadeq Saeed Mahdi [2] https://orcid.org/0000-0003-4834-6900
Pedro Y. Piñero Pérez [2] https://orcid.org/0000-0002-7635-8290
Ansel Y. Rodríguez González [3] https://orcid.org/0000-0001-9971-0237
Ireimis Leguen de Varona [1] https://orcid.org/0000-0002-1886-7644

[1] *Universidad de Camagüey Ignacio Agramonte Loynaz. Camagüey, Cuba*
[2] *Universidad de las Ciencias Informáticas. La Habana, Cuba*
[3] *Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad de Transferencia Tecnológica Tepic. Baja California, México*

**Autor para la correspondencia:* julio.madera@reduc.edu.cu

Editor

Lisset González Navarro
Academia de Ciencias de Cuba.
La Habana, Cuba

Traductor

Darwin A. Arduengo García
Academia de Ciencias de Cuba.
La Habana, Cuba

RESUMEN

Introducción: Los modelos probabilísticos y estadísticos constituyen herramientas de gran importancia para la resolución de problemas reales de optimización o clasificación. Estas herramientas son aplicadas para la toma de decisiones en el transporte, la agricultura, economía o la industria farmacéutica. **Objetivos:** Desarrollar modelos probabilísticos en la solución de problemas de optimización reales, tanto para dominio discreto como continuo, de planificación y de clasificación desbalanceada. **Métodos:** Se evaluaron los algoritmos propuestos en varios problemas de optimización teóricos y prácticos, así como bases de datos internacionales. Se ajustaron los parámetros de los algoritmos y se aplicaron técnicas estadísticas para la validación de los resultados. **Resultados:** Los resultados de esta investigación tributan, principalmente, a todos aquellos sectores y organismos donde se hace necesario una eficiente toma de decisiones en la planificación y uso de sus recursos. Esto incluye problemas de transporte, electricidad, agricultura, economía o de la industria farmacéutica. Esta investigación contribuye a la formación profesional de jóvenes graduados universitarios, enriqueciendo de esta forma el claustro de las universidades. **Conclusiones:** Los algoritmos propuestos que utilizan modelos probabilísticos para la solución de problemas de optimización constituyen una herramienta poderosa para la toma de decisiones en entornos reales. Los resultados demuestran la superioridad de estas técnicas comparadas con otras del estado del arte.

Palabras clave: modelos probabilísticos; algoritmos con estimación de distribuciones; estimación de la matriz de covarianza; problemas de optimización; planificación de tareas

Premio Nacional Anual de la Academia de Ciencias de Cuba, 2023

Theoretical and practical aspects of probabilistic models in solving real optimization problems

ABSTRACT

Introduction: Probabilistic and statistical models constitute tools of great importance for solving real optimization or classification problems. These tools are applied for decision making in transportation, agriculture, economics or pharmaceutical industry. **Objectives:** To develop probabilistic models in the solution of real optimization problems, both for discrete and continuous domain, planning and unbalanced ranking. **Methods:** The proposed algorithms were evaluated in several theoretical and practical optimization problems, as well as international databases. The parameters of the algorithms were adjusted and statistical techniques were applied to validate the results. **Results:** The results of this research contribute, mainly, to all those sectors and organizations that require efficient decision making in the planning and use of their resources. This includes problems of transportation, electricity, agriculture, economy or the pharmaceutical industry. This research contributes to the professional training of young university graduates, thus enriching the universities teaching staff. **Conclusions:** The proposed algorithms using probabilistic models for the solution of optimization problems constitute a powerful tool for decision making in real environments. The results demonstrate the superiority of these techniques compared to others in the state of the art.

Keywords: Probabilistic models; Estimation of Distributions Algorithms; covariance matrix estimation; optimization problems; task scheduling

INTRODUCCIÓN

Los modelos probabilísticos o estadísticos son las formas que pueden tomar un conjunto de datos obtenidos de muestreos con comportamiento que se supone aleatorio. Donde, un modelo probabilístico es un tipo de modelo matemático que usa la probabilidad, y que incluye un conjunto de asunciones sobre la generación de algunos datos muestrales, de tal manera que se asemejen a los datos de una población mayor. Las asunciones o hipótesis de un modelo probabilístico describen un conjunto de distribuciones de probabilidad, que son capaces de aproximar de manera adecuada un conjunto de datos. [1] Las distribuciones de probabilidad inherentes a los modelos probabilísticos son lo que distinguen a estos de otros modelos matemáticos deterministas.

Queda especificado por un conjunto de ecuaciones que relacionan diversas variables aleatorias, y en las que pueden aparecer otras variables no aleatorias. Como tal "un modelo es una representación formal de una teoría". Todas las pruebas de hipótesis estadísticas y los estimadores estadísticos proceden de modelos estadísticos. De hecho, los modelos estadísticos son una parte fundamental de la inferencia estadística. [2,3,4,5]

Asimismo, la estadística es la ciencia que recoge, clasifica y analiza la información que se presenta habitualmente mediante datos agregados; que permiten que las observaciones puedan cuantificarse, medirse, estimarse y compararse utilizando medidas de tendencia central, medidas de distribución, métodos gráficos, entre otros. Además, se ocupa de reunir y organizar datos numéricos, que ayuden a resolver problemas como el diseño de experimentos y la toma de decisiones informadas, a crear políticas de empresa y a comprender distintos aspectos de la vida moderna. [5]

Una de las aplicaciones de las probabilidades y las estadísticas, son los modelos gráficos probabilísticos (MGP), constituyen grafos en los cuales los nodos representan variables aleatorias y los arcos representan relaciones de dependencia condicional. Estos grafos proveen una forma compacta de representar la distribución de probabilidad. [6,7] Como parte de los MGP se encuentran las redes Gaussianas que son modelos gráficos de interacción para la distribución normal multivariada, utilizan la matriz de covarianza para analizar las relaciones entre las variables o pueden ser usados como modelos univariados. [8,9,10, 11, 12,13,14,1516]

Una característica de los problemas de optimización reales es la dependencia que existe entre las variables. [17] Unido a cuestiones relacionadas con la complejidad del espacio de búsqueda, la cantidad de óptimos de la función objetivo, así como

otras cuestiones teóricas definen lo que es un problema de optimización complejo. Otra cuestión importante es que varios de los algoritmos y técnicas que se utilizan para resolver este tipo de problemas, no tomen en cuenta las dependencias entre las variables, cuestión importante que conduce al fallo del método encontrando el óptimo. Se han desarrollado varias investigaciones que incorporan conocimiento de las dependencias entre las variables, para hacer la optimización más eficiente y eficaz.

Es por ello que definimos el problema de investigación y el objetivo general enfocados hacia la solución de esta problemática. Todos los resultados que forman parte de esta propuesta a premio ACC responden al cumplimiento de este objetivo. El problema es que los algoritmos de clasificación y optimización no aprovechan la relación ni la información disponible en las variables, lo cual puede afectar tanto la correcta clasificación de los ejemplos como la optimización de las variables en un problema real.

Se define como el problema científico a resolver las insuficiencias en las estrategias para la resolución de problemas de optimización reales tanto para dominio discretos como continuos, de planificación y clasificación desbalanceada, afines a las demandas de la sociedad cubana actual. Como objetivo general se plantea desarrollar modelos probabilísticos en la solución de problemas de optimización reales, tanto para dominio discreto como continuo, de planificación y de clasificación desbalanceada.

MÉTODOS

Algoritmos con estimación de distribuciones basados en restricciones

A partir del estudio del algoritmo de estimación de distribuciones (EDA) basado en poliárboles se propone e investiga la clase de algoritmos EDA que utilizan pruebas de independencias en el aprendizaje de la estructura probabilística. Estos algoritmos se conocen como EDA basados en restricciones los que definen la clase llamada algoritmos de estimación de distribuciones con restricciones (CBEDA).

Un primer resultado de impacto científico es la creación de un nuevo algoritmo llamado CBEDATPDA que utiliza el método de detección de dependencias de 3 fases para el aprendizaje de redes Bayesianas. [13] Los resultados experimentales demuestran que la nueva propuesta exhibe adecuadas cualidades numéricas para la solución de problemas con codificación entera como son las funciones decepcionantes y el problema de la predicción de estructuras de proteínas (PSP, del inglés, Protein Structure Prediction). Los resultados son comparados con otros algoritmos del estado del arte de la computación evolutiva, incluyendo propuestas del campo de los EDA.

Algoritmos con estimación de distribuciones celulares

Los algoritmos evolutivos (EA) celulares pertenecen a un tipo de algoritmos evolutivos basados en estructuras espaciales, donde cada individuo interactúa con su vecino adyacente. Una vecindad solapada ayuda en la exploración del espacio de búsqueda, mientras que la explotación toma lugar dentro de una vecindad por operadores estocásticos. En este treabajo se investiga una clase de algoritmos evolutivos (EA), los algoritmos de estimación de distribuciones celulares (CEDA) en problemas de optimización discreta y continua.

La intención principal es el estudio de la eficiencia del CEDA desde la perspectiva de disminuir el número de evaluaciones de la función objetivo (eficiencia evaluativa). Como resultado y aporte científico, se crean 3 nuevos algoritmos celulares basados en estimaciones de distribuciones de probabilidad: a) CEDA que utiliza redes Bayesiana (CEBA), b) CEDA basado en redes Gaussianas (CEGA) y c) CEDA basado en las distribuciones normal y Cauchy (CUMDANCauchy). [16,17,18,19,20]

El algoritmo de aprendizaje de CEBA y CEGA primeramente construye un esqueleto no orientado haciendo pruebas de independencia y después lo orienta con un proceso de optimización de métrica. Los algoritmos basados en optimización de métricas utilizan métodos de puntuación (probabilidad máxima con penalización, probabilidad marginal, basados en teoría de la información, etc.) y aproximaciones de búsqueda (tabú, greedy, etc.). Se expone que el aprendizaje a partir de los datos es típicamente proyectado como un problema de optimización en el cual la tarea computacional es encontrar la estructura que maximice el resultado, encontrar la red Bayesiana óptima es posible solo para las redes que contienen unos pocos individuos, ya que las técnicas de aprendizaje dirigidas a la solución de este problema no garantizan conducir al óptimo global. Al agregar estas redes Bayesianas a una célula o rejilla permite obtener un CEBA o un CEGA, algoritmos 1 y 2.

Las redes Gaussianas se describen como modelos gráficos de interacción para la distribución multivariante normal, estos modelos se basan en la independencia condicional. Solo contienen arcos no dirigidos y esto los convierte no solo en uno de los modelos conceptualmente más simples, sino también en uno de los más aplicados (algoritmo 2).

El algoritmo celular de distribución marginal univariado (algoritmo 3) con distribución normal-Cauchy (CUMDANCauchy) es un EDA celular que utiliza una estimación univariada del producto de las distribuciones normal y Cauchy sobre cada variable y produce nuevos individuos mediante el muestreo de la distribución aprendida. El método CUMDANCauchy utiliza la combinación de 2 distribuciones, normal + Cauchy para estimar los nuevos valores de una variable aleatoria, utiliza una combinación de ambas.

```
1   t ← 1
2   Generar  individuos al azar
3   Mientras no se cumpla criterio de terminación hacer
4           Para toda célula hacer
5               Seleccionar localmente  individuos de la vecindad
6               Estimar la distribución Bayesiana  de los  individuos seleccionados
7               Generar TamañoDeLaCelda nuevos individuos según la distribución
8               Insertar los individuos generados en la misma celda de una población auxiliar
9           Fin para
10          Reemplazar la población actual por la auxiliar
11          Calcular y actualizar las estadísticas
12          t ← t + 1
13  Fin mientras
```

Algoritmo 1. Algoritmo celular con estimación Bayesian

```
t ← 1
Generar  individuos al azar
Mientras no se cumpla criterio de terminación hacer
Para toda célula hacer
    Seleccionar localmente  individuos de la vecindad
    Estimar la distribución Gaussiana  de los  individuos seleccionados
    Generar  nuevos individuos según la distribución
    Insertar los individuos generados en la misma celda de una población auxiliar
Fin para
Reemplazar la población actual por la auxiliar
Calcular y actualizar las estadísticas
t ← t + 1
Fin mientras
```

Algoritmo 2: Algoritmo celular con estimación Gaussiana

Dado que los vecindarios clásicos basados en cuadrículas implican un gran número de individuos para definirlos, se introduce una estructura de anillo sobre la que quedan definidos los vecindarios. El aprendizaje de una distribución de probabilidad a partir de un conjunto pequeño de valores es una limitación, por lo tanto, CUMDANCauchy utiliza el tamaño máximo de la vecindad, es decir, el tamaño de la población; CUMDANCauchy puede usar diferentes vecindades y aprender una combinación de distribuciones normal y Cauchy de la población global para generar los nuevos individuos.

Estos algoritmos fueron aplicados a problemas de optimización de energía en redes inteligentes y el enrutamiento del transporte de recursos médicos para enfrentar la COVID-19, donde mostraron un comportamiento numérico superior al resto de los algoritmos del estado del arte comparados. [19,20] El aporte práctico de esta investigación es que los investigadores del campo de la computación evolutiva disponen de una metodología para reducir el número de evaluaciones, en la solución de problemas de optimización discretos y continuos, basada en modelos celulares univariados y en pruebas de independencia en el aprendizaje. [10]

Algoritmos con estimación de distribuciones para la construcción de cronogramas de proyectos

Los procesos de planificación de proyectos se presentan como el problema de organizar un conjunto de tareas respetando sus relaciones de precedencia y asignar recursos humanos y no humanos a las mismas, sin violar la disponibili-

```
1   t ← 1
2   Generar individuos al azar
3   Crear estructura celular anular a partir de individuos
4   Mientras hacer
5       Seleccionar globalmente
6       Estimar la distribución de los individuos seleccionados para cada generación
7       Para toda célula hacer
8           Muestrear los T nuevos individuos según la distribución estimada de cada generación
9           Insertar los individuos generados en la misma celda de una población auxiliar
10      Fin para
11      Reemplazar la población actual por la auxiliar
12      Calcular y actualizar las estadísticas
13      t ← t + 1
14  Fin mientras
```

Algoritmo 3: Algoritmo celular basado en las distribuciones normal y Cauchy

dad de los recursos en cada instante de tiempo. Actualmente aproximadamente el 57 % de los proyectos de tecnologías de la información son renegociados o cancelados, esto trae consigo impacto negativo tanto desde el punto de vista económico como social. Muchos de estos fracasos se deben a deficientes procesos de planificación provocados por la poca alineación a estándares, la insuficiente experiencia y la falta de herramientas que ayuden a construir cronogramas de proyectos óptimos o cuasi óptimos considerando que las tareas pueden ser ejecutadas de múltiples modos.

En este contexto los procesos de planificación han sido tratados en la literatura científica como problemas de planificación de proyectos con recursos limitados; siendo este un problema de optimización combinatoria de la clase NP-completo. Además, estos problemas se caracterizan por la presencia de muchas variables correlacionadas, elemento que motivó el uso de los algoritmos con estimación de distribución (EDA) en la resolución de la problemática planteada.

En este sentido se propusieron 2 nuevos algoritmos con estimación de distribución que incluyen el tratamiento de restricciones dentro del modelo probabilístico: [16, 21]

- algoritmo de con factorización de la distribución de probabilidad y aprendizaje de restricciones (CL_FDA);
- algoritmo marginal con distribución univariada y aprendizaje de restricciones (CL_UMDA).

Se validan los algoritmos diseñados a partir de la experimentación y la comparación de los mismos con los mejores resultados reportados en la bibliografía y con implementaciones de los algoritmos FDA y UMDA.

Propuesta de algoritmo de estimación de distribuciones con aprendizaje de restricciones

El algoritmo CLEDA (constraint learning estimation of distribution algorithms) constituye un algoritmo general que representa a la familia de algoritmos con tratamiento de restricciones en el modelo probabilístico. [16,21] Inicialmente, como entrada del problema, se identifican los objetivos del problema de optimización (algoritmo 4).

En esta investigación se experimenta con los objetivos tiempo y costo, se definen 2 tareas ficticias ubicadas al inicio y al final del cronograma. Otras entradas son max_iterations y error_threshold que representan ejemplos de parámetros que influyen en la definición de la condición de parada. Otra entrada del algoritmo son las restricciones del problema en cuestión.

Se destacan en este sentido 2 tipos de restricciones:

- restricciones de precedencia entre tareas,

```
1   t ← 1
2   Generar individuos al azar
3   Crear estructura celular anular a partir de individuos
4   Mientras hacer
5       Seleccionar globalmente
6       Estimar la distribución de los individuos seleccionados para cada generación
7       Para toda célula hacer
8           Muestrear los T nuevos individuos según la distribución estimada de cada generación
9           Insertar los individuos generados en la misma celda de una población auxiliar
10      Fin para
11      Reemplazar la población actual por la auxiliar
12      Calcular y actualizar las estadísticas
13      t ← t + 1
14  Fin mientras
```

Algoritmo 4. Algoritmo de estimación de distribuciones con aprendizaje de restricciones

– restricciones respecto a la disponibilidad de los recursos renovables y no renovables.

Como parte de las restricciones se define un grupo de modos de ejecución de cada tarea y por cada modo se define la duración, la cantidad de recursos renovables y no renovables que implica la realización de esta tarea en cada modo.

Algoritmo con factorización de la distribución de probabilidad y aprendizaje de restricciones

Este algoritmo se propone como una variante del algoritmo CLEDA, que toma elementos esenciales del algoritmo FDA y se llamará CL_FDA. Este trata de resolver la problemática de la alta complejidad en la solución del problema MMRCPSP asociado a la dependencia entre variables desde el enfoque de aplicar factorizaciones que simplifiquen la búsqueda de soluciones, tomando como base el planteamiento de las propias restricciones del problema en cuestión (algoritmo 5). [15]

Algoritmo marginal con distribución univariada y aprendizaje de restricciones

En el caso del algoritmo CL_UMDA se define como se muestra en el algoritmo 6. La mayoría de los pasos de este algoritmo son similares a los explicados en la subsección anterior para el algoritmo CLEDA, por esta razón solo se explicarán de forma detallada los pasos que son diferentes entre los 2 algoritmos.

Utilización de la matriz de covarianza para balancear conjuntos de datos continuos

En la actualidad, muchas tareas de la vida cotidiana involucran aprendizaje sobre conjuntos de datos desbalanceados, lo cual en la mayoría de los casos es difícil de manejar. Primeramente, en este trabajo se extienden los principales conceptos del desbalance de datos para dominios continuos y clase binaria (datos numéricos), así como los algoritmos que se han desarrollado para resolver este problema, los cuales no emplean la relación de dependencia de los atributos. [21]

Se propone un nuevo algoritmo basado en la técnica de la matriz de covarianza como un método de aprendizaje estadístico óptimo para el balanceo de datos. Asimismo, se presenta un algoritmo basado en la matriz de covarianza para balancear conjuntos de datos con dominios continuos y clase binaria, teniendo en cuenta que los atributos se encuentren fuera y dentro del rango para la generación de sobremuestreo de las clases. [12]

RESULTADOS Y DISCUSIÓN

Para realizar los experimentos se configuraron los parámetros de los algoritmos teniendo en cuenta garantizar el éxito en 95 de 100 corridas de cada parámetro. A partir de ese resultado se ajusta el tamaño de población mínima para alcanzar esa cantidad de éxitos. En todos los casos se utiliza la selección por truncamiento con un valor del 30 % del tamaño de la población global. El criterio de parada es encontrar el óptimo, un valor cercano a este (10^{-5}) o ejecutar un número fijo de generaciones.

El algoritmo de estimación de distribuciones con maximización-minimización y escalador de colinas supera a los algoritmos del estado del arte

Una cuestión interesante es hasta qué punto estas características de los algoritmos podrían suponer una mejora

```
pupul = initial_population(constraints) #Generación población inicial
pupul = evaluation(popul, constraints, objectives) #Evaluación
While not stop_condition(popul, indicators) do
        S = schedule_selection(popul, sel_size) #Selección mejores
        p_e^S(x, t) = distribution_FDA_learn(popul, S, constraints, learn_centrality)
        new_popul = schedule_sow_elitism(popul, S)
        new_popul = schedule_new_solutions(popul, new_popul, p_e^S(x, t))
        new_popul = improve_solutions(new_popul, p_e^S(x, t))
        popul = evaluation(new_popul, constraints, objectives)
End while
Return popul
```

Algoritmo 5. Algoritmo de distribución factorizada con aprendizaje de restricciones

```
pupul = initial_population(constraints) #Generación población inicial
pupul = evaluation(popul, constraints, objectives) #Evaluación
While not stop_condition(popul, indicators) do
        S = schedule_selection(popul, sel_size) #Selección mejores
        p_e^S(x, t) = distribution_UMDA_learn(popul, S, constraints, learn_centrality)
        new_popul = schedule_sow_elitism(popul, S)
        new_popul = schedule_new_solutions(popul, new_popul, p_e^S(x, t))
        new_popul = improve_solutions(new_popul, p_e^S(x, t))
        popul = evaluation(new_popul, constraints, objectives)
End while
Return popul
```

Algoritmo 6. Algoritmo de distribución marginal univariado con aprendizaje de restricciones

significativa respecto a la calidad de los EDA y sus capacidades para resolver problemas del mundo real de altas dimensiones, principalmente problemas en los campos de la minería de datos, el reconocimiento de patrones y la inteligencia artificial. [18] En efecto, la comparación del MMHCEDA con los algoritmos BOA y EBNA (tabla 1) permite concluir que se trata de un algoritmo de optimización evolutiva competitivo. Teniendo en cuenta la naturaleza de las funciones de prueba elegidas, conjeturamos que se obtendrán resultados similares con otras clases de problemas. La investigación aquí presentada es un aporte científico significativo por la obtención de 2 algoritmos EDA competitivos, capaces de resolver problemas de optimización entera.

Comparación del algoritmo celular con estimación Bayesiana con otros enfoques discretos

La eficiencia evaluativa del CEBA es comparada con varios acercamientos, como PADA, SPADA y CUMDA. En Martínez-López se analizan los resultados experimentales con estos métodos (EDA simple, PADAp3, PADAp2, PADAt3, PADAt2, SPADA p = 1, SPADA p = 6), teniendo en cuenta el número de evaluaciones y generaciones. [23] Los resultados sugieren que el EDA celular con vecindarios L9, C9 y C41 con la cuadrícula 20x20x2x2 exhibieron un mejor rendimiento que el EDA simple. Además, los EDA celulares con mejor comportamiento son los CEBA con aprendizaje de la estructura y los parámetros de población local (L5 y C9) con generación (3,86 ± 1,20)

Tabla 1. Comparación del algoritmo de estimación de distribuciones con maximización-minimización y escalador de colinas con otros del estado del arte en la función BF2B30s4-1245

						MMHCEDA
100	3,33	0	0	0	13,3	6,67
200	3,33	30	10	0	16,6	23,3
300	3,33	46,6	23,3	13,3	30	40
400	3,33	53,3	43,3	30	30	66,6
500	3,33	50	56,6	43,3	26,6	53,3
1000	3,33	43,3	50	50	40	73,3
4000	3,33	43,3	50	43,3	33,3	100
8000	3,33	43,3	50	50	43,3	100

y (3,96 ± 1,92) y número de evaluaciones (382 ± 118,88) y (393,04 ± 190,31) respectivamente.

Comparación del algoritmo celular con estimación Gaussiana con otros enfoques continuos

La eficiencia evaluativa del CEGA es comparada con varios acercamientos, como EGNA, EBCOA y MIMICc. En Martínez-López se analizan los resultados experimentales con estos acercamientos para las funciones Sphere y Ackley, teniendo en cuenta el número de evaluaciones y generaciones. [9] Estos resultados sugieren que el EDA celular con vecindarios L9, C13 y C41 con la cuadrícula 20x20x2x2 exhibieron un mejor rendimiento que el EDA simple. Además, los EDA celulares con mejor comportamiento son los CEGA con aprendizaje de la estructura y los parámetros de población local (C13 y C41).

En la figura 1 se muestra una de las principales conclusiones de los CEBA y CEGA para los dominios discreto y continuo, respectivamente.

Comparación del algoritmo celular con distribución marginal univariado con normal-Cauchy con otros enfoques

En este experimento, CUMDANCauchy se compara con los algoritmos desarrollados (y algunos de ellos actualizados) para el marco del GECCO 2019, es decir combinación del algoritmo de Búsqueda de vecinos variable y Optimización de enjambre de partículas evolutiva diferencial (VNS–DEEPSO), Optimización híbrida de búsqueda de vecindad variable con enjambre de partículas Levy (HL –PSVNSO), Optimización de enjambre de partículas de vecindad variable con mapa de Gauss (GM–VNPSO) y Optimización de enjambre de partículas con la mejor perturbación global (PSO–GBP).

CUMDANCauchy utilizó un tamaño de población de 10, un máximo de iteraciones de 499, y el número de escenarios utilizados por evaluación de la función de *fitness* es 10; mientras que los algoritmos de 2019 utilizaron los parámetros definidos por sus desarrolladores. El índice de clasificación se calculó a partir de 20 ejecuciones independientes de cada algoritmo. [18]

Aplicación de los algoritmos de estimación de distribuciones celulares a la distribución de recursos relacionados con la COVID-19 en Camagüey

Para realizar el siguiente experimento se utilizó la biblioteca CVRP [247], implementada en el lenguaje Matlab, versión 19a. 3 metaheurísticas (EDA celular, SA, VNS). Se experimentó con un total de 15 modelos (cantidad de nodos x cantidad de recursos disponibles) para el estudio del comportamiento de estos algoritmos aplicados al problema.

La figura 2 describe como las metaheurísticas tienen comportamientos de la eficiencia evaluativa similares para las diferentes configuraciones de los modelos, los cuales convergen a la solución deseada. Aunque cabe analizar, que tanto los algoritmos VNS, como EDA celular, sus comportamientos fueron ligeramente superior al SA. En este caso el comportamiento de las iteraciones para encontrar el óptimo es similar. Estos experimentos permiten analizar el comportamiento de las diferentes metaheurísticas para luego ser seleccionadas y resolver los diferentes problemas de toma de decisión.

En la aplicación práctica se modeló la flota de la empresa EMSUME en Camagüey. La misma cuenta con una flota de 4 vehículos, todos del mismo tipo (h = 1) y con la misma capacidad (Qh = 100, interpretada como cajas con las mismas dimensiones). Cada viaje de un vehículo tiene un costo fijo de 100 CUP y un costo por km de 10 CUP. Cada vehículo puede recorrer, como máximo, 400 km y los choferes no pueden manejar más de 9 h. La velocidad promedio de los vehículos es de 50 km/h, la misma se relaciona con la calidad de las carreteras y con las condiciones técnicas

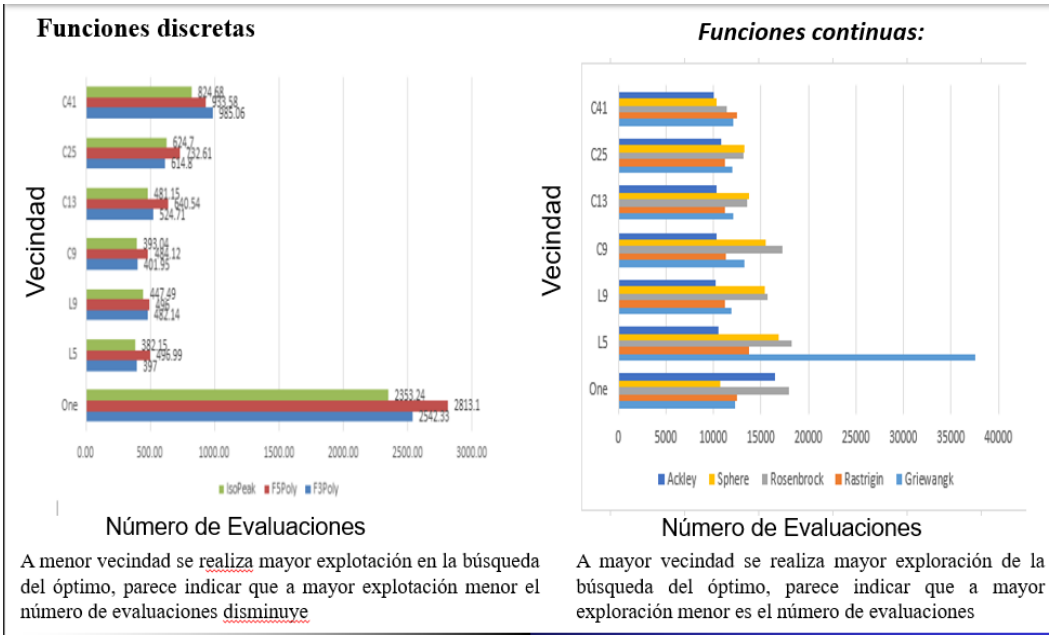

Fig. 1. Conclusiones de los experimentos continuos y discretos para los algoritmos de celular con estimación Bayesiana y algoritmo celular con estimación Gaussiana

de los vehículos. Número de clientes *n* = 12, representa los 12 municipios de la provincia Camagüey.

Se tienen las distancias entre los municipios, las que representan el peso en los arcos del grafo. La ventana de tiempo de carga de los vehículos en el almacén central es [6:00;8:00], indica que los vehículos deben iniciar su carga en ese horario y el tiempo de servicio del vehículo es de 55 minutos.

La ventana de trabajo modelada para los clientes es [8:00;16:00], indica que todas las descargas de vehículos deben comenzar después de las 8:00 am y antes de las 4:00 pm. El tiempo de descarga de la mercancía en los municipios es la misma para todos e igual a 20 min.

Además, se definieron las cantidades de envases a transportar a cada municipio, o lo que es lo mismo, la demanda del municipio. A partir de la información modelada se procedió a la solución del problema FSMVRPTW a través de un algoritmo VNS, obteniéndose la solución óptima.

Aplicación de los algoritmos de estimación de distribuciones celulares en el ordenamiento del transporte urbano en Camagüey

Para realizar los experimentos en esta sección se consultó la literatura de trabajos previos, y se realizó un estudio de los parámetros. Para cada una de las metaheurísticas, los parámetros de configuración fueron: SA(MaxIt = 1200; MaxIt2 = 80; T0 = 100; alpha = 0,98), VNS (MaxIt = 1200; maxK = 80), GRASP (MaxIt = 1200; MaxIt2 = 80) y EDA celular (MaxIt = 1200; MaxPop = 50; numParents = 15). Donde, MaxIt es el número máximo de iteraciones, MaxIt2 es el número de iteraciones del algoritmo, T0 es la temperatura inicial, MaxPop es el máximo de la población, numParents es el número de padres a seleccionar.

Una de las propuestas fue utilizar el algoritmo EDA celular para la ruta 19 con la utilización de 6 ómnibus youtong de 32 asientos y 2 dianas de 25 asientos con una diferencia de 13 min entre ellos, de modo que salgan de la siguiente forma: 3 youtong, 2 dianas y 3 youtong.

Algoritmos con restricciones en el modelo de aprendizaje para la generación de cronogramas

En la experimentación se emplearon un conjunto de 150 instancias recogidas en 15 bases de datos de la librería para los problemas de planificación de proyectos PSPLib. Para garantizar mayor veracidad de los resultados obtenidos se ejecutan 20 corridas por cada algoritmo en cada instancia. Para la comparación de los algoritmos se usan varios indicadores y respecto a cada indicador se aplican *test* estadísticos paramétricos o *test* no paramétricos, dependiendo de la distribución de los datos. Se constata que las propuestas logran mejores resultados que los obtenidos por los EDA tradicionales respecto a lograr planificaciones más cortas de proyectos, la cantidad de veces que se encuentra el óptimo en las bases de datos empleadas en la experimentación y un mejor balance tiempo-costo en la planificación.

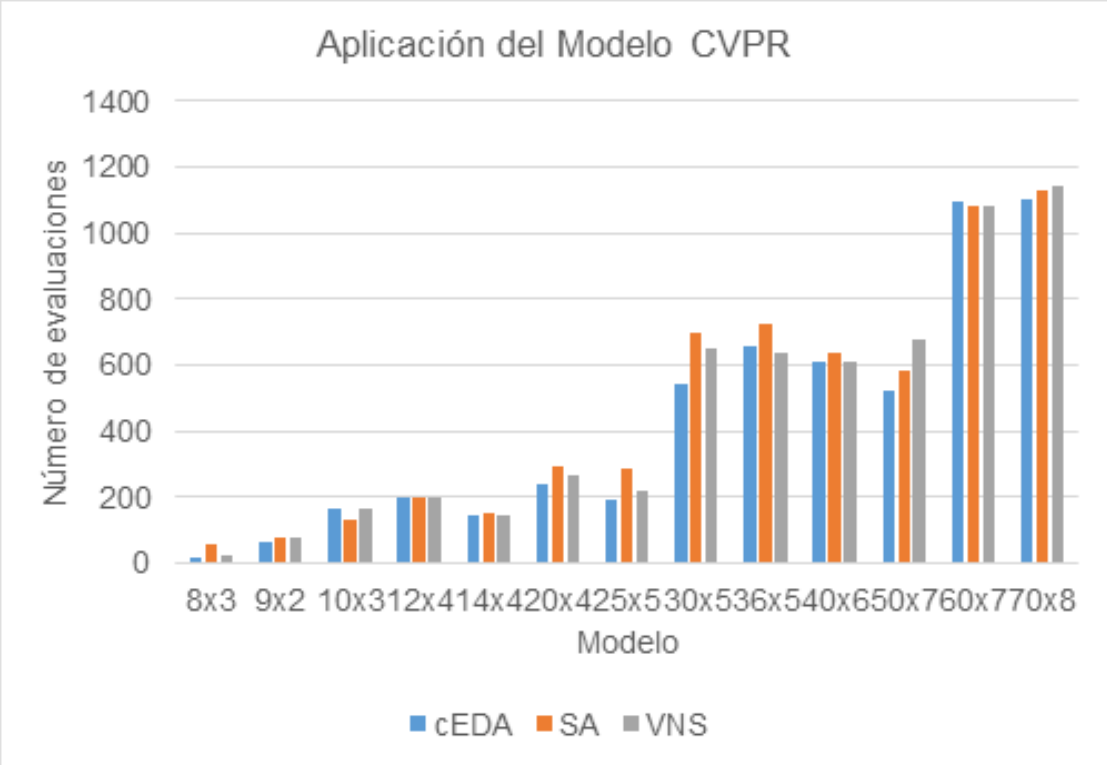

Fig. 2. Resultados del experimento con algoritmo de estimación de distribuciones celular, recocido simulado y búsqueda de vecindad variable

Aplicabilidad de los algoritmos

Se realizó un estudio de la aplicabilidad de los algoritmos en entornos reales a partir de un caso de estudio real de 90 tareas respecto al indicador al indicador mínimo_tiempo_duración del proyecto y tiempo de ejecución de los algoritmos donde se evidencia que los algoritmos propuestos logran minimizar el tiempo de duración del proyecto de 148 días a 80 días. No obstante, el tiempo de ejecución de los algoritmos fue elevado en comparación con los EDA tradicionales justificado por la complejidad del aprendizaje del modelo probabilístico con restricciones.

Respecto a la eficacia en la búsqueda de soluciones, se demuestra que los algoritmos propuestos reportan resultados significativamente mejores que las implementaciones de los algoritmos UMDA y FDA. Además, respecto al indicador mínimo_tiempo_duración, se identifica que no existen diferencias significativas entre los algoritmos propuestos y los mejores algoritmos reportados de la literatura. También, respecto al indicador media_tiempo_duración, los algoritmos propuestos en la BD no superan los resultados de los algoritmos Robust hGMEDA, hGMEDA y SPEA2. Ante la variación en los modos, respecto a los indicadores media_tiempo_duración, balance-tcosto y cantidad_veces_óptimo, se experimentó un cambio donde diferentes variantes de CL_UMDA reportan resultados superiores al resto de los algoritmos. Ante la variación en la cantidad de tareas se identificó que las variantes de CL_UMDA reportan resultados significativamente mejores que el resto de los algoritmos. El enfoque de optimización just_time reporta resultados ligeramente superiores y no se encuentran diferencias entre las estrategias de aprendizaje (figura 3).

Regularización de la covarianza

En Leguen-de-Varona se realiza un estudio experimental para demostrar que el algoritmo propuesto es igual de eficaz que los algoritmos con el mismo fin del estado del arte. Los algoritmos se aplican a problemas de Bioinformática, específicamente a la clasificación de enfermedades utilizando información genética. [12,13] Al demostrarse el uso óptimo de la matriz de covarianza para detectar dependencias entre las variables de un problema de clasificación, permitirá incluir estas técnicas como núcleo de los EDA.

Para el estudio experimental se escogieron 7 conjuntos de datos del repositorio de la UCI con índice de desbalance (IR) (≥ 1,5). [23,24] Estos conjuntos de datos contienen todos sus atributos continuos, la clase binaria y diferentes valores de IR.

Luego se procedió a balancear el conjunto de entrenamiento, generando nuevas instancias sintéticas a partir de la clase minoritaria hasta completar las cantidades de la clase mayoritaria usando la matriz de covarianza, como método de

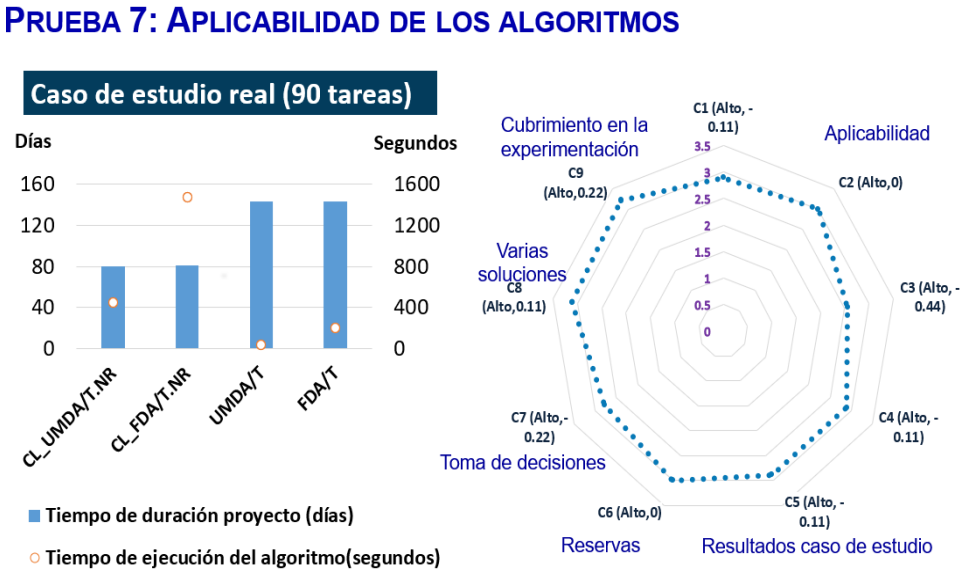

Fig. 3. Aplicabilidad de los algoritmos de distribución factorizada con aprendizaje de restricciones y algoritmo de distribución marginal univariado con aprendizaje de restricciones

aprendizaje, teniendo en cuenta que los valores de cada atributo fueran generados dentro (CovD) o no (CovF) del rango y utilizando una muestra de control *test*, que continúa no balanceada y sin ninguna modificación. [26] Una vez generados los nuevos conjuntos de datos a partir de las instancias obtenidas, usando el algoritmo anteriormente descrito, se estudió el valor AUC con este conjunto de datos ya balanceado. En el primer experimento, se seleccionó el clasificador C4.5 (J48) y en el segundo experimento se seleccionaron además los clasificadores MLP y KNN.

Los resultados del algoritmo de balanceo de datos aplicando la matriz de covarianza con sus variantes CovD y CovF son similares o comparables con respecto a los algoritmos de sobremuestreo del estado del arte. La tabla 2 muestra el *ranking* obtenido por la prueba de Friedman en el experimento 2. En este caso el *p*-value asociado a dicha prueba es de 0,000001, el cual es lo suficientemente bajo como para rechazar la hipótesis de equivalencia, por lo que se puede concluir que existen diferencias significativas entre los algoritmos comparados, resaltando que el de mejor *ranking* fue CovF-MLP, una de las propuestas del artículo.

Tabla 2. *Test* de Friedmann para ordenar los algoritmos evaluados

Algoritmo	Ranking
SMOTEBagging	4.2143
RUSBoost	4.0714
UnderBagging	5.3571
CovF-C4.5	6.2143
CovD-C4.5	6.8571
CovF-MLP	1.7857
CovD-MLP	2.6429
CovF-KNN	5.8571
CovD-KNN	8

Conclusiones

Los resultados de esta investigación confirman que los algoritmos de estimación de distribuciones propuestos —CBEDATPDA, CEBA, CEGA y CUMDANCauchy— incrementan de forma sustancial la eficiencia evaluativa y la calidad de las soluciones al modelar explícitamente las dependencias entre variables en dominios discretos y continuos. Asimismo, la familia CLEDA y sus variantes CL_FDA y CL_UMDA, que integran restricciones de precedencia y recursos dentro del modelo probabilístico, generan cronogramas más cortos y con mejor balance tiempo-costo que los EDA tradicionales y otros métodos de referencia.

El enfoque celular reduce drásticamente el número de evaluaciones sin sacrificar rendimiento, lo que habilita su aplicación práctica en problemas de gran escala, como redes eléctricas inteligentes, logística sanitaria frente a la COVID-19 y transporte urbano.

Por otra parte, el método de balanceo basado en la matriz de covarianza mejora significativamente la AUC en conjuntos de datos continuos desbalanceados, posicionándose como alternativa competitiva frente a técnicas de sobre muestreo convencionales. En conjunto, estos hallazgos validan la tesis central de la investigación: incorporar conocimiento probabilístico sobre la estructura de los datos es clave para resolver eficazmente problemas reales de optimización, planificación y clasificación desbalanceada, proporcionando herramientas robustas para la toma de decisiones en sectores críticos.

REFERENCIAS BIBLIOGRÁFICAS

1. McCullagh P. What is a statistical model? Ann Stat [Internet]. 2002 [citado 19 sep 2023]; 30:1225–310. Disponible en: https://doi.org/10.1214/aos/1035844977
2. Adèr HJ. Modelling. In: Adèr HJ, Mellenbergh GJ, editors. ***Advising on research methods: a consultant's companion***. Huizen, The Netherlands: Johannes van Kessel Publishing; 2008. p. 271-304
3. Burnham KP, Anderson DR. Model selection and multimodel inference. 2nd ed. Springer-Verlag; 2002. ISBN: 0-387-95364-7.
4. Cox DR. Principles of statistical inference. Cambridge University Press; 2006.
5. Konishi S, Kitagawa G. Information criteria and statistical modeling. Springer; 2008.
6. Scutari M. Bayesian network structure learning, parameter learning and inference. 2011.
7. Madera J. Hacia un generación más eficiente de algoritmos evolutivos con estimación de distribuciones: Pruebas de independencia + paralelismo [tesis doctoral]. La Habana: Universidad de La Habana; 2009.
8. Martínez-López Y, Madera J, Rodríguez-González Q, Barigye S. Cellular Estimation Gaussian Algorithm for Continuous Domain. ***J Intell Fuzzy Syst*** [Internet]. 2019 [citado 11 nov 2023];36(5):4957-67. Disponible en: https://doi.org/10.3233/JIFS-179042
9. Richardson Ibañez J. Algoritmos Evolutivos Estimadores de Distribución Celulares para Problemas de Optimización Continuos [tesis]. Camagüey: Universidad de Camagüey; 2017.
10. Rodríguez-González AY, Lezama F, Martínez-López Y, Madera J, Soares J, Vale Z. WCCI/GECCO 2020 Competition on Evolutionary Computation in the Energy Domain: An overview from the winner perspective. Appl Soft Comput [Internet]. 2022 [citado 5 nov 2023]; 109162. Disponible en: https://doi.org/10.1016/j.asoc.2022.109162
11. Zaldívar-Pino O. Biblioteca de clases en R para el trabajo con algoritmos que estiman distribuciones [tesis]. Camagüey: Universidad de Camagüey; 2013.
12. Leguen-de-Varona I, Madera J, Martínez-López Y, Hernández-Nieto J. Over-sampling imbalanced datasets using the Covariance Matrix. EAI Endorsed Trans Energy Web [Internet]. 2020 [citado 18 sep 2023]. Disponible en: https://doi.org/10.4108/eai.13-7-2018.163982
13. Leguen-de-Varona I, Madera J, Martínez-López Y, Hernández-Nieto J. SMOTE-Cov: A new oversampling method based on the Covariance Matrix. In: Litvinchev PI, Marmolejo-Saucedo JA, Rodriguez-Aguilar R, Martinez-Rios F, editors. Data analysis and optimization for engineering and computing problems [Internet]. Cham: Springer; 2020 [citado 18 sep 2023]. p. 207-17. Disponible en: https://doi.org/10.1007/978-3-030-48149-0_15
14. Leguen-de-Varona I. Algoritmo basado en la estimación de la Matriz de Covarianza para balancear conjuntos de datos sobre dominio continuo y clase binaria [tesis]. Universidad de Camagüey; 2018.
15. Leguen-de-Varona I, Madera J, Martínez-López Y, Hernández-Nieto JC. Smote-Cov: A new oversampling method based on the Covariance Matrix. Presentado en: Ciudad de México, México; 2019 nov 28.
16. Mahdi GS. Algoritmos de estimación de distribuciones con tratamiento de restricciones para la construcción de cronogramas de proyectos [tesis doctoral]. La Habana: CUJAE; 2021.
17. Madera J, Ochoa A. Evaluating the Max-Min Hill-Climbing Estimation of Distribution Algorithm on B-Functions. In: Hernández Heredia Y, Milián Núñez V, Ruiz Shulcloper J, editors. Progress in Artificial Intelligence and Pattern Recognition. IWAIPR 2018. Lecture Notes in Computer Science, vol 11047 [Internet]. Cham: Springer; 2018 [citado 21 sep 2023]. p. 1-12. Disponible en: https://doi.org/10.1007/978-3-030-01132-1_3
18. Martínez-López Y, Rodríguez AY, Madera J, Mayedo MB, Moya A, Salgado OM. Applying Some EDAs and Hybrid Variants to the ERM Problem Under Uncertainty. ACM [Internet]; 2020 [citado 5 nov 2023]. Disponible en: https://doi.org/10.1145/3377929.3398393
19. Martínez-López Y, Guevara Yanes L, Madera Quintana J. Aplicación de metaheurísticas en el ordenamiento del transporte urbano en Camagüey. Rev Cubana Transform Digit [Internet]. 2022 [citado 23 oct 2023];3(2):e171. Disponible en: https://rctd.uic.cu/rctd/article/view/171
20. Martínez-López Y, Oquendo H, Mota YC, Guerra-Rodríguez LE, Junco R, Benítez I, Madera J. Aplicación de la investigación de operaciones a la distribución de recursos relacionados con la COVID-19. Retos Dirección [Internet]. 2020 [citado 21 nov 2023]. Disponible en: http://scielo.sld.cu/pdf/rdir/v14n2/2306-9155-rdir-14-02-86.pdf
21. Mahdi GS, Quintana JM, Pérez PP, Al-subhi SH. Estimation of Distribution Algorithm for solving the Multi-mode Resource Constrai-

ned Project Scheduling Problem. EAI Endorsed Trans Energy Web [Internet]. 2020 [citado 12 oct 2023]. Disponible en: https://doi.org/10.4108/eai.13-7-2018.164111.
22. Martínez-López Y, Madera J, Mahdi GS, Rodríguez-González AY. Cellular estimation bayesian algorithm for discrete optimization problems. Investigación Operacional [Internet]. 2020 [citado 21 nov 2023]. Disponible en: https://rev-inv-ope.pantheonsorbonne.fr/sites/default/files/inline-files/41720-10.pdf
23. Guerra M, Madera J. WIFROWAN: Wrapped Imbalanced Fuzzy-Rough Ordered Weighted Average Nearest Neighbor Classification. Computación y Sistemas [Internet]. 2020 [citado 13 sep 2023]. Disponible en: https://doi.org/10.13053/cys-24-3-3054
24. Madera J, Martínez López Y, Fernández-Pardo J. Algoritmo de Estimación de Distribución basado en el Aprendizaje de Redes Bayesianas con Pruebas de Independencias para Problemas de Optimización en Enteros. Rev Cubana Cienc Inform [Internet]. 2020 [citado 2 oct 2023]; 14(4):1–19. ISSN: 2227-1899. Disponible en: https://rcci.uci.cu/?journal=rcci&page=article&op=download&path%5B%5D=2019&path%5B%5D=832

Recibido: 18/10/2024
Aprobado: 18/11/2024

Agradecimientos
Alberto Ochoa Rodríguez, Stephen J. Barigye, Alexis Moya, Bismay Morgado, Miguel Bethencourt Mayedo, José Carlos Hernández-Nieto, Salah Hassan Al-subhi, Roberto García Vacacela, Iliana Pérez Pupo, Hilda Oquendo Ferrer, Yaile Caballero Mota, Luis E. Guerra Rodríguez, Raul Junco Villegas, Isnel Benítez Cortés, Olga Lidia Pérez González, Miguel Á. Álvarez-Carmona, Samantha Barajas, Ramón Aranda, Fernando Lezama, Joao Soares, Zita Vale, Oscar Martínez Santiago, Lenier Guevara Yanes, José Miguel Fernández Pardo, Demetrio Alejandro Rodríguez Fernández.

Conflictos de intereses
Puede estar dado que las figuras ya han sido publicadas en los artículos que avalan el trabajo, por lo que la publicación de este resultado necesita permiso de las revistas.

Contribuciones de los autores
- Conceptualización: Julio Madera Quintana
- Curación de datos: Julio Madera Quintana, Yoan Martínez López, Gaafar Sadeq Saeed Mahdi, Pedro Yobanis Piñero Pérez, Ansel Rodríguez González
- Análisis formal: Yoan Martínez López, Gaafar Sadeq Saeed Mahdi,
- Adquisición de fondos: Julio Madera Quintana, Ansel Rodríguez González
- Investigación: Yoan Martínez López, Gaafar Sadeq Saeed Mahdi, Ireimis Leguen de Varona
- Metodología: Julio Madera Quintana, Yoan Martínez López, Gaafar Sadeq Saeed Mahdi, Pedro Yobanis Piñero Pérez, Ansel Rodríguez González
- Administración del proyecto: Julio Madera Quintana,
- Software: Yoan Martínez López, Gaafar Sadeq Saeed Mahdi, Ireimis Leguen de Varona
- Supervisión: Julio Madera Quintana, Pedro Yobanis Piñero Pérez, Ansel Rodríguez González
- Validación: Yoan Martínez López, Gaafar Sadeq Saeed Mahdi, Ireimis Leguen de Varona
- Visualización: Julio Madera Quintana, Yoan Martínez López, Gaafar Sadeq Saeed Mahdi
- Redacción original: Julio Madera Quintana, Yoan Martínez López, Gaafar Sadeq Saeed Mahdi
- Redacción-revisión y edición: Julio Madera Quintana, Pedro Yobanis Piñero Pérez, Ansel Rodríguez González

Financiamientos
Esta investigación ha sido apoyada por la Cooperación Belga al Desarrollo a través de VLIR-UOS. VLIR-UOS apoya asociaciones entre universidades y colegios universitarios en Flandes (Bélgica) y el Sur que buscan respuestas innovadoras a los desafíos globales y locales (Proyecto Red TIC-VLIR).

Cómo citar este artículo
Madera Quintana J, Martínez López Y, Saeed Mahdi G, Piñero Pérez PY, Rodríguez González AY, Leguen de Varona I. Aportes teóricos y prácticos de los modelos probabilísticos en la solución de problemas de optimización reales. An Acad Cienc Cuba [internet] 2024 [citado día, mes y año];15(2):e1935. Disponible en: http://www.revistaccuba.cu/index.php/revacc/article/view/1935