RIELAC, Vol. 40 3/2019 p. 79-91 Septiembre – Diciembre ISSN: 1815-5928

RIELAC
Revista de Ingeniería Electrónica, Automática y Comunicaciones

Obtención de predicados difusos con un enfoque multiobjetivo: comparación de dos variantes

Orenia Lapeira, Taymi Ceruto, Alejandro Rosete

RESUMEN / *ABSTRACT*

FuzzyPred es un algoritmo de aprendizaje no supervisado de minería de datos, que permite extraer predicados difusos en forma normal a partir de los datos. Este método se utiliza para resolver una tarea descriptiva donde no se conoce a ciencia cierta qué tipo de relaciones se van a encontrar. Se trata de encontrar patrones que describan los datos y sus relaciones. Debido al gran conjunto de soluciones o espacio de búsqueda que puede tener, fue modelado como un problema de optimización, donde se aplican las metaheurísticas como vía de solución para encontrar buenas soluciones. FuzzyPred brinda como resultado un conjunto de predicados, evaluados en cada una de las medidas de calidad, aunque solo optimiza una de estas medidas. Este trabajo analiza vías para enfocar FuzzyPred como un problema de optimización multiobjetivo. Por esto, se introducen en el problema dos de las técnicas principales de optimización multiobjetivo: la técnica basada en Pareto (o multiobjetivo puro) y la de los factores ponderados. Se realiza un estudio experimental comparativo entre ambas técnicas en este problema para conocer la eficacia de estas técnicas. Los resultados en varias bases de datos internacionales demuestran que se obtienen mejores resultados con la técnica multiobjetivo puro.

Palabras claves: Minería de Datos, Predicados Difusos, Técnicas de Optimización Multiobjetivo.

FuzzyPred is an unsupervised-learning data-mining algorithm, which allows extracting fuzzy predicates in normal forms from the data. This method is used to solve a descriptive task where it is not known the kind of relationships that are to be found. The goal is to find patterns that describe the data and their relationships. Due to the large set of solutions or search space that may have, it was modeled as an optimization problem, where metaheuristics are applied to find good solutions. FuzzyPred provides as a result a set of predicates, evaluated in each of the quality measures, in spite of the fact that only one of the measures is optimized (truth value). This paper introduces a multiobjective approach for FuzzyPred by using two of the most known multi-objective optimization techniques: Pareto technique (or pure multi-objective) and weighted factors. An experimental study is presented in order to compare the efficacy of both techniques in this problem. The results in several international databases show that better results are obtained by the pure multi-objective technique.

Keywords *Data Mining, Fuzzy Predicates, Multiobjective Optimization Techniques.*

1. -INTRODUCCIÓN

La teoría de conjuntos difusos está cobrando una importancia creciente en el ámbito de la minería de datos [1]. Los conjuntos difusos permiten establecer límites flexibles entre los distintos niveles de significado, sin ignorar ni enfatizar en exceso los elementos cercanos a la frontera, tal y como ocurre con la percepción humana. En todo proceso de extracción de conocimiento hay un componente de interacción humana y los conjuntos difusos representan este conocimiento de forma lingüística e incorporan conocimiento previo de forma fácil y proporcionan soluciones interpretables.

FuzzyPred es un algoritmo de aprendizaje no supervisado de Minería de Datos, que permite extraer predicados difusos en forma normal a partir de los datos [2]. El mismo ha sido modelado como un problema de optimización monobjetivo [2] que

Recibido: 18/2/2019 Aceptado: 27/5/2019

se resuelve a través de metaheurísticas. La calidad de los resultados se mide empleando una familia de medidas basadas en el uso de los operadores difusos [2].

Este algoritmo, está diseñado bajo los principios del proceso de KDD propuesto en [1], que utiliza los predicados difusos en forma normal conjuntiva y disyuntiva, como modo de representación del conocimiento. Este método se utiliza para resolver una tarea descriptiva donde no se conoce a ciencia cierta qué tipo de relaciones se van a encontrar. Así, se trata de encontrar patrones que describan los datos y sus relaciones. Los datos que se utilizan para construir la vista minable deben ser transformados a valores difusos. Es decir, como precondición, cada atributo a utilizar debe estar representado por etiquetas lingüísticas y su grado de pertenencia.

En los problemas de minería de datos existen diferentes características que hacen factible el uso de metaheurísticas [3,4] . Debido al gran conjunto de soluciones o espacio de búsqueda que puede tener, FuzzyPred fue modelado como un problema de optimización combinatorio, donde se aplican las metaheurísticas como vía de solución para encontrar buenas soluciones, aunque sin garantía de optimalidad.

FuzzyPred brinda como resultado un conjunto de predicados, evaluados en cada una de las medidas de calidad, pero solo utiliza una de esas medidas (el valor de verdad [2], principalmente) para optimizar la búsqueda de los mejores predicados. La necesidad de incluir otras medidas de calidad para optimizar la búsqueda de los mejores predicados, hace que el problema de la obtención de predicados difusos necesite ser modelado como un problema de optimización multiobjetivo. De esta manera, no solo se obtendrán predicados que sean buenos según su valor de verdad, sino que pueden ser buenas en otros criterios (funciones asociadas a objetivos).

Existen varias técnicas que permiten resolver un problema con un enfoque multiobjetivo. Dos de las más conocidas son: la de Pareto (o multiobjetivo puro) y la de factores ponderados [5, 6, 7, 8, 9]. En la primera, se considera que las diferentes funciones son incomparables, mientras que en la segunda se definen valores de importancia para cada objetivo.

Como la obtención de predicados difusos no se ha modelado como un problema multiobjetivo, tampoco se han realizado pruebas experimentales que permitan conocer la eficacia de las técnicas de optimización multiobjetivo en este problema. Es por ello que en este trabajo, se propone una comparación entre dichas técnicas (basada en Pareto y factores ponderados, específicamente), con el objetivo de poder conocer la calidad de los resultados obtenidos en cada una de ellas. El resto del trabajo está estructurado de la siguiente manera. En la sección II, se introduce un marco teórico acerca de la búsqueda de predicados difusos. En la sección III se abordan aspectos generales acerca de la optimización multiobjetivo y IV, se explica el algoritmo multiobjetivo de FuzzyPred, dígase esquema de codificación, función de evaluaciones y proceso de optimización y búsqueda. En la sección V se muestran los resultados experimentales obtenidos en cada una de las técnicas. Finalmente, en la sección VI se presentan las conclusiones y reflexiones del trabajo.

2.- Búsqueda de predicados difusos en una base de datos

A pesar de la enorme cantidad de datos y algoritmos que existen, es numerosa la información que aún no ha sido compilada y analizada. Si hubiera un método específico que fuera más conveniente y sin desventajas, entonces no habría necesidad de seguir desarrollando otros métodos diferentes. Disponer de un método que encuentre patrones que puedan interpretarse como varios modelos, se considera como una variante interesante [2].

En la minería de datos existen diferentes modelos para representar el conocimiento (árboles de decisión, reglas, grupos) y para obtener cada uno de ellos se pueden usar diferentes algoritmos (ID3, Apriori, K-Medias, entre otros) de acuerdo al tipo de tarea (clasificación, agrupamiento, asociación, entre otros) que se esté realizando [10,11] . Ninguno de estos modelos brinda la posibilidad de obtener conocimiento donde (por ejemplo) la relación lógica principal sea la doble equivalencia o la conjunción o la disyunción o cualquier combinación libre de operadores entre variables de una base de datos. Es por ello que FuzzyPred [2] se propone encontrar un nuevo modelo que represente en Forma Normal (conjunciones de disyunciones simples, o disyunciones de conjunciones simples) el conocimiento extraído de una base de datos. Esto es posible debido a que todos los modelos lógicos son representables en Lógica Clásica usando formas normales, esto da un posible alcance general al conocimiento que se puede obtener usando FuzzyPred. Este enfoque es flexible y brinda la posibilidad de descubrir un conocimiento diferente de los métodos anteriores, debido a que los mismos se encuentran enfocados en un modelo determinado y tienen restricciones asociados al mismo. Este método tiene como característica que trabaja con bases de datos fuzzificadas, por tanto, cada una de las variables de la base de datos debe tener un valor real entre 0 y 1.

FuzzyPred se utiliza para resolver una tarea de minería de datos descriptiva, en la cual no se sabe a ciencia cierta qué resultados se quieren obtener. La extracción automática de predicados se puede realizar con diferentes técnicas de

aprendizaje, no obstante, en FuzzyPred se propone el uso de metaheurísticas. Las mismas presentan una serie de ventajas, entre las que se destacan su potente capacidad de búsqueda en espacios de soluciones grandes [12,13,14].

La optimización multiobjetivo ha sido probada como un mecanismo interesante en varias investigaciones de diferentes aplicaciones [2,9]. Sin embargo aunque los predicados pueden evaluarse usando diferentes métricas, en FuzzyPred la búsqueda se orienta empleando un solo objetivo (maximizar el valor de verdad [2].

Si se contara con un variante multiobjetivo de FuzzyPred, se podría obtener una familia de predicados con diferentes compromisos entre las diferentes funciones a optimizar, por ejemplo, podrían haber predicados más legibles y otros con mayor valor de verdad, o que sean buenos cualquier otra de las medidas definidas para evaluar los predicados difusos [2]. Así, cada predicado (solución) tiende a satisfacer un criterio de la búsqueda en mayor medida que cualquier otra solución.

3.- Optimización multiobjetivo

La optimización multiobjetivo trata de solucionar problemas que involucran múltiples objetivos. Un problema de optimización multiobjetivo, es aquel que incluye un conjunto de funciones objetivos a optimizar (dos o más funciones). Uno de los retos del momento es que la mayoría de los problemas de optimización que se presentan, en diversos sectores tienen en cuenta más de un objetivo, por tanto requieren de una solución que satisfaga todos los objetivos, en determinada medida [7].

Los problemas de optimización multiobjetivo (de ahora en adelante MOPs) pueden ser definidos según [7] como:
"…un vector de variables de decisiones las cuales satisfagan restricciones y optimizan un vector cuyos elementos representan dos o más funciones objetivos. El objetivo de este tipo de problema es encontrar parámetros necesarios que optimicen el vector de funciones objetivos…"

Precisamente, los MOPs son problemas donde se desea optimizar k funciones objetivos simultáneamente. Puede incluir la maximización de todas las k funciones, o la minimización de todas las k funciones, o la combinación de la de las maximización y minimización k funciones. En el caso de los MOPS, la existencia de múltiples soluciones conlleva la imposibilidad de decidir automáticamente cuál de ellas es la mejor, siempre y cuando se considere a todos los objetivos con igual importancia.

Los métodos que en esta sección se describen le permiten al usuario especificar sus preferencias como forma de representación de la optimización multiobjetivo. Estas técnicas están divididas en dos grupos [2,5]: técnicas a priori y técnicas a posteriori. Las técnicas a priori son aquellas en las que las preferencias del usuario tienen que ser conocidas antes de que comience la búsqueda, entre estas están el ordenamiento lexicográfico y factores ponderados. En las técnicas a posteriori las preferencias se van dando conforme la búsqueda avanza y el tomador de decisiones indica si una solución le parece adecuada o no y el proceso actualiza las preferencias conforme el tomador de decisiones lo va indicando, guiando así el proceso de búsqueda.

La técnica de factores ponderados combina los resultados de las distintas funciones objetivo en un único valor de aptitud. Este valor de aptitud se obtiene con una combinación lineal de las funciones objetivo. El problema es que se debe de introducir un factor de pesos para las funciones objetivo y en la práctica resulta difícil realizar una buena asignación para las funciones objetivos.

La idea básica de la técnica basada en Pareto [15,16, 17, 18, 19, 20] (u multiobjetivo puro) es encontrar las soluciones no dominadas. Las soluciones no dominadas son aquellas para las cuales no se conoce otra solución que sea mejor que ella en algún criterio y que a la vez no sea peor en los demás criterios, es decir, que la domine. De esta manera, en lugar de obtenerse una solución, el objetivo de la búsqueda es encontrar un conjunto de soluciones no dominadas (o Frente de Pareto). Las soluciones en el Frente de Pareto son aquellas para las cuales no exista ninguna otra solución que la domine, o sea, son las soluciones no dominadas. Esta técnica no necesita que el decisor asigne pesos a priori a diferencia de la técnica antes comentada. Luego el decisor, teniendo el Frente de Pareto, decide la solución que mejor considere para su caso de interés.

4.- Enfoque multiobjetivo de FuzzyPred

4.1.- Esquema de codificación

El esquema de codificación es la estructura de datos que representa los parámetros de una posible solución, y el criterio de evaluación es la función que permite evaluar cada individuo. En este trabajo una solución (individuo) codifica un único predicado. Luego, la solución global que se le ofrece al usuario, es el conjunto de predicados obtenidos en diferentes ejecuciones del algoritmo. Cada solución del problema, es un vector que utiliza codificación entera para representar los posibles valores de cada variable. En este vector, todos los atributos posibles son codificados; pero se establece una marca especial (valor cero) que indica la ausencia de la variable en el predicado.

La cantidad de cláusulas para cada predicado está definida para valores entre 1 y 3 cláusulas. La codificación de un predicado consta de tres partes principales:

- La primera parte corresponde a la sucesión de cláusulas (Sc). En esta parte de la codificación se modelan tres cláusulas, aunque, se puede representar que alguna no esté presente mediante el valor cero (ver figura 2 cláusula 3). Los valores de modificadores posibles de cada cláusula son: 0 (ausencia), 1 (presencia), 2 (negado), 3 (muy), 4 (híper), 5 (algo), 6 (medio), 7 (variable negada con muy), 8 (variable negada con híper), 9 (variable negada con algo), 10 (muy medio), 11 (híper medio), 12 (algo medio).
- La segunda parte representa la cantidad de cláusulas del predicado (Cc).
- La tercera parte representa la forma normal (Fn) que relaciona a las variables de Sc. Cuando Fn vale 0 el predicado esta en FNC (Forma Normal Conjuntiva) y cuando vale 1 en FND (Forma Normal Disyuntiva).

Los valores tomados por cada una de las variables representan modificadores los cuales intensifican o disminuyen el valor de verdad de la variable.

La Fig. 1 muestra el esquema de una solución con dos variables: a y b. La figura 2 representa un predicado en forma normal conjuntiva debido a que Fn tiene valor 0, con dos cláusulas (Cc=2). En la tercera cláusula no interviene ni la variable A ni la B ya que toman valor 0 en la codificación, por lo que el predicado quedaría de la siguiente manera (A o no B) y (muy B). Lo mismo sucede con la variable A en la segunda cláusula.

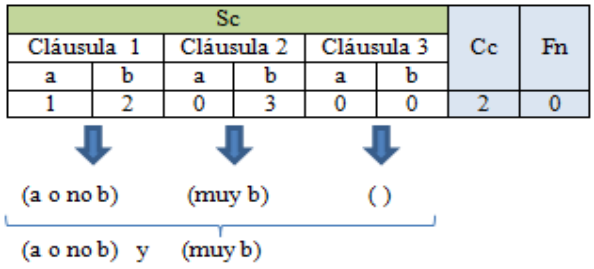

Figura 1

Ejemplo de codificación de un predicado.

Es importante aclarar que todos los valores de la codificación pueden cambiar en cada ejecución, creando así disímiles predicados. Esto le permite incorporar flexibilidad al algoritmo, aunque también se pueden fijar una cantidad de cláusulas y un tipo de forma normal, si se quiere esto para contextos específicos.

4.2.- Evaluación

El criterio de evaluación de cualquier modelo se define a través de un conjunto de medidas de calidad propuestas en [2] . Cada una de estas funciones son las que permiten establecer hasta qué punto se ajusta bien un patrón a los objetivos. En el caso especial de los predicados difusos, los valores de cada variable oscilan entre 0 y 1, donde el cero indica absolutamente falso y el uno indica que es absolutamente verdadero. Los valores intermedios indican una pertenencia gradual. El valor de verdad es la principal medida utilizada para evaluar los predicados difusos. Para evaluar el valor de verdad de un predicado, se evalúa el valor del predicado para cada ejemplo presente en la base de datos, y luego se busca el valor de verdad global, aplicando un cuantificador universal, que se implementa como una conjunción de los valores de verdad correspondientes a cada ejemplo [2].

La medida valor de verdad posee determinadas limitantes (como el pesimismo de la aplicación del cuantificador universal y el criterio del veto) como fue estudiado en [2]. El cuantificador universal (la conjunción de todos los resultados) tiende a restringir el resultado final (trata de que todos los elementos cumplan), que a su vez está determinado por el operador de lógica difusa que sea seleccionado (criterio del veto), teniendo en cuenta que cada operador posee sus propias características.

En [2] se proponen un conjunto de medidas de calidad que no sufren de estas limitantes y que se pueden usar para evaluar la calidad de los patrones. Estas medidas, le ofrecen ayuda al usuario a la hora de tomar decisiones y fueron adaptadas a partir de otras definidas para las reglas de asociación y otras medidas estadísticas. Estas medidas son: soporte, soporte binario, poda central de predicados difusos, poda superior de predicados difusos, y poda inferior de predicados difusos. En FuzzyPred se contemplan todas estas métricas para evaluar la calidad de los predicados obtenidos.

4.3.- Búsqueda y optimización

Para el proceso de búsqueda de los predicados se utilizan metaheurísticas. Esto se debe a que no se han definido algoritmos exactos que lo hagan y porque el espacio de soluciones puede llegar a ser grande para hacer una exploración exhaustiva, teniendo en cuenta los aspectos que determinan la complejidad del problema:

- la cantidad de variables involucradas, que a su vez depende de la cantidad de cláusulas del predicado.
- los posibles valores que puede tomar cada variable.

Independientemente de que la cantidad de combinaciones no fuera tan grande, de todos modos no es factible desde el punto de vista computacional hacer el cálculo de cada una de las funciones objetivo para todas las combinaciones posibles, por la dimensión que pueda tener la vista minable.

Las metaheurísticas se clasifican en dos grupos principales: basados en un punto o de trayectoria, y basados en poblaciones de puntos [14]. Los métodos de búsqueda basados en un punto son aquellos que partiendo de un punto, buscan la vecindad y actualizan la solución actual en función de esta, formando una trayectoria, de punto a punto. Los algoritmos basados en poblaciones, trabajan con un conjunto de soluciones en cada iteración, y su resultado debe ser el mejor individuo de la población o sea el individuo mejor adaptado.

Hay muchos métodos metaheurísticos que se pueden usar para resolver problemas combinatorios [21, 22]. Algunos son variantes de otros, introduciendo nuevas ideas o heurísticas. Hay algunos más simples, otros tienen más parámetros. Algunos tienen buena fama y otros están subvalorados. De forma general hay mucha propaganda y sectarismo en este tema; pero la realidad dice que no hay superioridad posible de ninguno sobre otro en general [23].

Según el Teorema NFL (de “No Free Lunch Theorem”) [23, 24] ningún algoritmo es mejor que otro en la totalidad de los problemas en los que son aplicables. Es decir, si un algoritmo A es mejor que uno B sobre un grupo determinado de problemas, entonces debe esperarse que existan un conjunto igual de problemas donde ocurra lo contrario.

Esto hace que un algoritmo se comporte mejor o no con respecto a un problema determinado, según la función objetivo, así según los operadores que permiten variar de un estado a otro del problema. En general, esto sugiere que lo adecuado es evaluar varias metaheurísticas en cada problema concreto.

Para buscar el conjunto de soluciones que necesita FuzzyPred es posible utilizar varias metaheurísticas. En este trabajo se evaluaran un conjunto de metaheurísticas: algoritmos metaheurísticos monobjetivo y multiobjetivos (tanto basados en un punto como basados en poblaciones de puntos) para la técnica de factores ponderados y basada en Pareto, respectivamente. La explicación de cada una de las metaheurísticas aplicadas se encuentra a continuación:

Algoritmos metaheurísticos monobjetivo

- Escalador de Colinas (EC) [13]: El escalador de colinas es un método de propósito general y su principal limitación es la convergencia al máximo local más cercano. Para ello, en caso de que la evaluación de la solución candidata no sea mejor que la solución actual, indica que la solución actual es un máximo local y el EC nunca saldrá de ese punto.
- Recocido Simulado (RS) [13]: Es una de las metaheurísticas más antiguas y fue una de los primeros algoritmos que contenían una estrategia explicita para escapar de óptimos locales, basándose en la idea fundamental de permitir movimientos que no producían mejoras, o sea aceptar soluciones peores que la solución actual. La probabilidad de aceptación de soluciones peores disminuye a lo largo de la búsqueda y depende de la evaluación en la función objetivo de la solución actual y un parámetro de control que regula la proporción de malas soluciones.
- Estrategia Evolutiva (EE) [13]: Inicialmente la EE parte de generar una población inicial aleatoria de individuos. Luego se va evolucionando la población actual como resultado de aplicar los operadores: selección, mutación, y reemplazo, hasta que no se cumpla la condición de parada que puede ser un número máximo de generaciones, y finalmente se devuelve la mejor solución o individuo mejor adaptado de la población actual.
- Algoritmo Genético (AG) [13]: La estructura del AG es muy similar al de la EE, con la única diferencia que incorporan el proceso de recombinación de individuos el operador de cruzamiento, entre selección y mutación. Los individuos seleccionados se recombinan para intercambiar parte de su información genética. A partir de este intercambio se obtienen nuevos individuos que tienen una codificación que posee algo de ambos padres. La ocurrencia de esto se controla con un parámetro llamado probabilidad de cruzamiento.

Algoritmos metaheurísticos multiobjetivos

- Escalador de Colinas Estocástico Multiobjetivo (ECMO) propuesto en [13]: Se trata de una metaheurística que compara la dominancia entre la solución actual y una solución candidata, y si esta última no es dominada, entonces se compara con las soluciones no dominadas encontradas hasta ese momento. Se elimina del listado aquellas soluciones que son dominadas, y si la solución candidata no es dominada se agrega al listado de soluciones no dominadas.
- Recocido Simulado Multicaso (RSMO) [13]: El algoritmo propone una regla de aceptación que se basa en la comparación de la solución actual con la nueva solución, tomando en cuenta el caso en que las dos soluciones sean indiferentes (hay mejoras en algunos objetivos y deterioros en otros). En esta situación se derivan tres sub-casos. En el primer sub-caso, el algoritmo chequea el número de soluciones de Pareto que son dominados por la nueva solución (llamado contador de dominancia de la nueva solución (R (xc)). En el segundo sub-caso, si la nueva solución no domina ninguna solución de Pareto, entonces se chequea si puede ser una solución del conjunto de Pareto. En el sub-caso 3, si la nueva solución no puede constituir una solución del conjunto de Pareto, entonces la probabilidad de aceptación está basada en la diferencia entre el rango de dominancia de la nueva solución y la solución actual.
- Multi-Objective Genetic Algorithm (MOGA)[13]: Específicamente, a cada individuo se le asigna un rango el cual será ordenado para la selección. El rango se asigna según un criterio de no dominancia. Los individuos dominados son penalizados de acuerdo a la densidad de las regiones que lo dominan. La población es ordenada según el rango de los individuos. Posteriormente, se asigna el valor de fitness para cada individuo por interpolación desde el mejor rango hasta el peor. Se promedia la adaptación de los individuos con el mismo rango, para que tengan igual valor de fitness.
- Non-dominated Sorting Genetic Algorithm (NSGAII) [13, 25, 26]: es una versión mejorada del algoritmo NSGA que implanta la idea de un método de selección basado en clases de dominancia para todas las soluciones. Este algoritmo varía del algoritmo genético en el operador de selección ya que antes que la selección sea efectuada, se clasifica la población usando la no dominancia de un individuo. Además, la nueva versión NSGAII difiere en que construye una población de individuos, asigna un rango y clasifica cada individuo según los niveles de no dominancia. Aplica operaciones evolutivas para crear una nueva agrupación de descendencias, y entonces combina a los padres y a sus hijos antes de dividir los grupos en partes frontales. Los mejores individuos son aquellos que tienen rangos menores, y por lo tanto más posibilidades de reproducirse en la siguiente generación.

5.- Resultados experimentales

Para el estudio experimental se utilizaron 8 bases de datos obtenidas de la UC Irvine Machine Learning. En la tabla 1 se muestran las características de cada una de las bases de datos. En la columna Ejemplos se muestran la cantidad de instancias de cada base de datos y la columna Atrib, el número de atributos desglosados en reales (R), entero (E) o nominales (N). Es importante aclarar que las bases de datos fueron fuzzificadas con el método de normalización min-max, para los valores enteros y reales. En este tipo de normalización el mayor valor de cada atributo toma valor 1, el menor valor 0, y los valores intermedios son proporcionales.

Tabla 1: Características de las bases de datos.

Nombre BD	Ejemplos	Atrib (R/E/N)
Balance Scale	625	5 (5/0/0)
Bolts	40	8 (2/6/0)
Quacke	2178	4 (3/1/0)
Basketball	90	5 (3/2/0)
Breast Cancer	278	5 (0/5/0)
Cloud	108	4(4/0/0)
Iris	150	4(4/0/0)
Autos	206	5(5/0/0)

Los siguientes valores fueron considerados en cada experimento:

- Se realizaron 30 repeticiones, cada una con 500 iteraciones.
- Se experimentó con dos objetivos. Las medidas de calidad escogidas fueron FPTV (el valor de verdad) y el SPD (Soporte para predicados difusos) [6]. Para la técnica de Factores Ponderados se asumieron 3 combinaciones diferentes para los pesos de cada una de las funciones objetivos: 0.3 a FPTV y 0.7 para SPD (Esquema 1), 0.5 para FPTV y 0.5 para SPD (Esquema 2), y 0.7 para FPTV y 0.3 para SPD (Esquema 3). Estas combinaciones se realizaron con el fin de conocer cómo se comportan los valores en cada objetivo independientemente de los pesos que se le otorguen.
- Se obtuvieron predicados de 1 hasta 3 cláusulas, tanto en Forma Normal Conjuntiva (FNC) como en Forma Normal Disyuntiva (FND).
- Metaheurísticas usadas para la obtención de predicados difusos con los tres variantes del enfoque de factores ponderados: EC, RS, AG, EE. A partir de los 30 frentes de Pareto obtenidos de cada ejecución de cada una de estas metaheurísticas se obtuvo un frente de Pareto de cada una de las variantes. Es decir, para cada variante se obtuvo el frente de Pareto uniendo los 30*4=120 frentes resultantes de las 30 ejecuciones de cada metaheurística monobjetivo.
- Metaheurísticas usadas para la obtención de predicados difusos con el enfoque multi-objetivo puro: RSMO, ECMO, NSGAII, MOGA. A partir de los 30 frentes de Pareto obtenidos de cada ejecución de cada una de estas metaheurísticas se obtuvo un frente de Pareto. Es decir, se obtuvo el frente de Pareto uniendo los 30*4=120 frentes resultantes de las 30 ejecuciones de cada metaheurística multi-objetivo.

Los parámetros de configuración de cada uno de las metaheurísticas fueron:

- RS y RSMO: temperatura inicial 15, temperatura final 0, α: 0 .93 y 20 iteraciones manteniendo la misma temperatura.
- EC y ECMO: 2 estados vecinos (esto define el tamaño de la vecindad).
- GA, EE, NSGAII y MOGA: se utilizaron 25 individuos, con probabilidad de cruzamiento de 1.0 y probabilidad de mutación 0.5, un factor de truncamiento del 30% de la población inicial empleando mutación uniforme, cruzamiento uniforme y reemplazo de estado estable.

En la Tabla 2 se muestran algunos de los predicados encontrados (resultados de la evaluación en cada una de las funciones objetivos) en cada una de las bases de datos para ambos enfoques (basado en Pareto y los tres esquemas de Factores Ponderados). La columna ID identifica los predicados mostrados.

Para la técnica de Pareto se muestra una de las soluciones no dominadas encontradas. Para la técnica Factores ponderados se muestran algunas de las mejores soluciones encontradas en los diferentes esquemas). En general, los algoritmos tienden a converger, obteniendo soluciones similares en diferentes ejecuciones. Se resaltan en negrita los mejores valores encontrados para cada uno de los objetivos.

Tabla 2: Ejemplos de algunos predicados encontrados en algunas de las bases de datos.

Nombre BD	Técnica	Predicado	FPTV	SPD	ID
Basketball	Basada en Pareto	(algo (AsistMin) o algo (Altura) o no (Points))	**1.0**	**1.0**	**Ba1**
	Factores Ponderados Esquema 1	(AsistMin o algo (Heigth) o muy (TiempoJugado) o no (Edad))	0.43	0.90	Ba2
	Factores ponderados Esquema 2	(AsistMin y no (Edad)) o (algo (Puntos))	0.20	0.95	Ba3
	Factores Ponderados Esquema 3	(algo (AsistMin) o muy (Puntos))	0.20	0.95	Ba4
Bolts	Basada en Pareto	(something (Run) or not (Speed) or something (Total) or very (Speed) or something (Number) or hyper (Sens) or not Time) or Bolts)	**1.0**	**1.0**	**Bo1**
	Factores ponderados Esquema 1	(something (Run) or not (Speed) or not (Total) or very (Speed) or Number or not (Sens) or something (Time) or not (Bolts))	0.92	0.98	Bo2
	Factores Ponderados Esquema 2	(Run or not (Speed) or something (Bolts)) and (very (Speed) or Sens)	0.30	0.76	Bo3
	Factores Ponderados Esquema 3	(Run or very (Speed) or not (Total) or very (Speed) or not (Number))	0.05	0.89	Bo4
Quacke	Basada en Pareto	(not (A) or something (B) or very (C) or something (D))	**0.83**	0.87	**Q1**
	Factores ponderados Esquema 1	(not (A) or hyper (B) or something (C) or not (D))	0.46	0.76	**Q2**
	Factores Ponderados Esquema 2	(not (A) and C and D) or (something (B))	0.18	0.70	Q3
	Factores Ponderados Esquema 3	(very (A) or something (B) or hyper (C) or D)	0.54	**0.94**	Q4

De los resultados obtenidos anteriormente puede comentarse lo siguiente:

- De las soluciones mostradas para Basketball, la única solución no dominada es Ba1, para Bolts es Bo1, para Quacke las no dominadas son Q1 y Q4, mientras que para Balance Scale las soluciones no dominadas son BS1 y BS2. Esto implica que en general, entre las 6 soluciones no dominadas encontradas (en las 4 bases de datos analizadas) hay 4 que las obtiene la técnica de Pareto.
- De manera general, las evaluaciones fueron mejores en la técnica basada en Pareto que, en la técnica de factores ponderados, en los tres esquemas diseñados.

- La métrica de FPTV obtiene peores valores que la métrica SPD, esto se debe a la información que brinda cada una de ellas y las limitantes que posee la métrica FPTV, aunque existen casos (Ej. Bolts) que llegan a ser muy cercanos sus evaluaciones.

Conjuntamente, para el análisis experimental, se tuvieron en cuenta el conjunto de soluciones no dominadas encontradas por todos los algoritmos metaheurísticos para cada técnica. Cada conjunto estará conformado por las mejores soluciones encontradas en cada técnica de optimización multiobjetivo para cada una de las bases de datos (frente de Pareto actual). El conjunto conocido como frente de Pareto verdadero, estará conformado por las soluciones no dominadas encontradas en ambas técnicas de optimización multi-objetivo.

En la Figura 2 se muestran el conjunto de soluciones no dominadas encontradas en ambas técnicas de optimización en cada una de las bases de datos. En dicha figura el conjunto de soluciones no dominadas por la técnica de factores ponderados (en cualquiera de sus tres esquemas) se encuentran representadas por círculos y las encontradas por la técnica basada en Pareto por asteriscos, los colores indican cada una de las bases de datos. En la gran mayoría de las bases de datos las soluciones no dominadas encontradas por la técnica basada en Pareto dominan a las obtenidas por las encontradas en los diferentes esquemas de la técnica factores ponderados. Este comportamiento es diferente solamente en las bases de datos Quacke y Cloud.

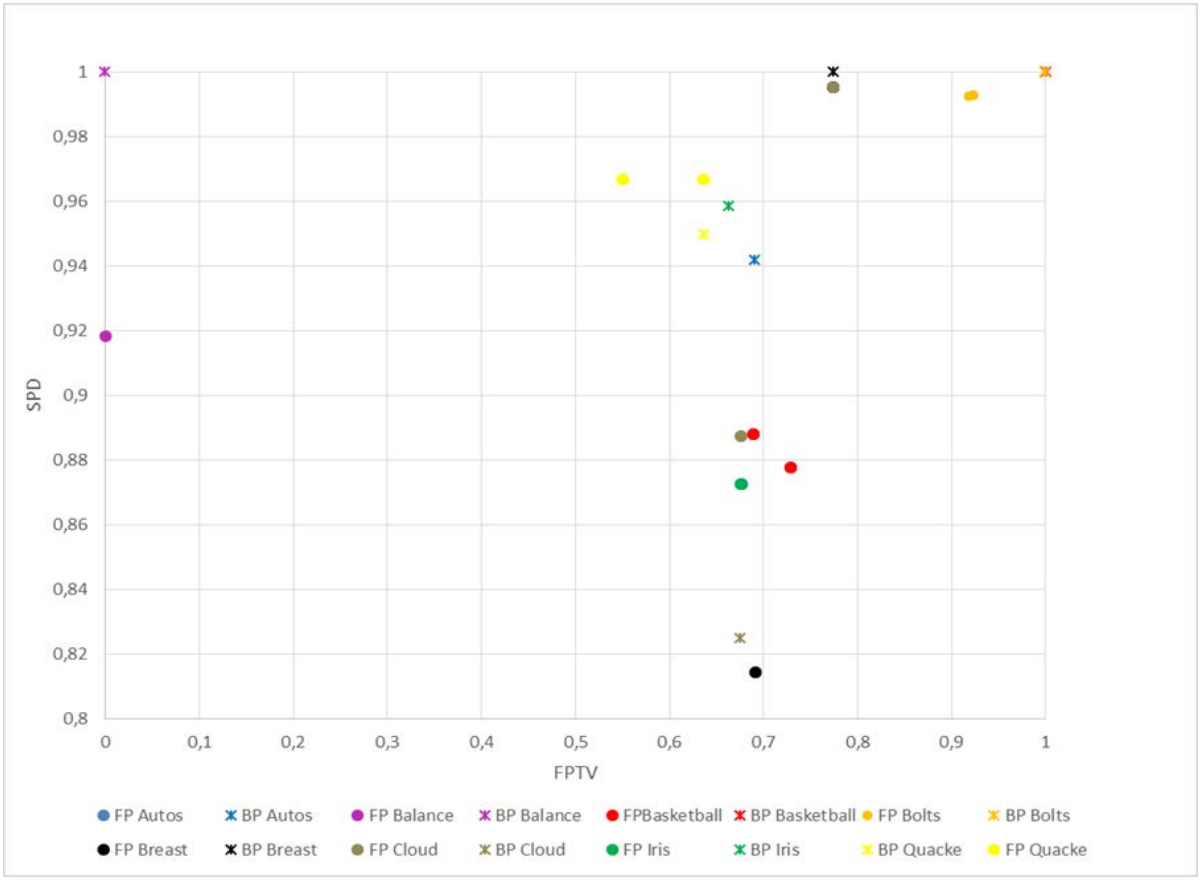

Figura 2: Conjunto de soluciones no dominadas encontradas por ambas técnicas en cada una de las bases de datos.

Para cada conjunto se calcularon algunas métricas de rendimiento que permiten medir la calidad (Tasa de error [20], Distancia generacional [21], y Spread [21]) de los frentes de Pareto obtenidos con respecto al frente de Pareto verdadero.

- Tasa de error (TE) [20]: indica el porcentaje de soluciones del frente de Pareto actual que no son miembros del frente de Pareto verdadero. Toma como referencia al frente de Pareto verdadero. Un valor 0, significa que todos los puntos obtenidos están en el frente de Pareto verdadero y 1 significa que ninguno lo está.
- Distancia Generacional (DG) [21]: Mide "que tan lejos" está el frente de Pareto actual del frente de Pareto verdadero. Toma como referencia al frente de Pareto actual. Si la distancia generacional es 0 indica que todos los elementos generados pertenecen al frente de Pareto verdadero, cualquier otro valor indica que tan lejos está del frente de Pareto verdadero.
- Spread o Delta indicador [21]: Combina conceptos como la dispersión y la cardinalidad para medir la distribución del frente de Pareto actual. Utiliza como información adicional la distancia a los "extremos" del frente de Pareto verdadero para tener una medida más precisa de la cobertura del frente de Pareto actual. La interpretación de esta métrica es que una dispersión igual a 0 indicaría una aproximación al frente de Pareto perfectamente distribuida.

En la Tabla 3 se muestran los resultados obtenidos en cada una de las métricas de rendimiento seleccionadas, para cada uno de los esquemas FP1 (Frente de Pareto Esquema 1: peso 0.3 para FPTV y 0.7 para SPD), FP2 (Frente de Pareto Esquema 2: peso 0.5 para FPTV y 0.5 para SPD), FP3 (Frente de Pareto Esquema 3: peso 0.7 para FPTV y 0.3 para SPD) y BP (Basada en Pareto). Sombreadas en negrita se encuentran los mejores valores para cada una de las instancias.

Tabla 3: Resultados de los frentes obtenidos en cada una de las métricas de rendimiento.

Autos				**Balance**			**Basketball**			**Iris**		
	DG	Spread	TE	DG	Spread	TE	DG	Spread	TE	DG	Spread	TE
FP1	0,09	1	1	0,04	1	1	0,005	1,14	1	0,008	**1**	1
FP2	0,01	1	1	0,05	1	1	0,016	1,14	1	0,010	**1**	1
FP3	0,05	1	1	0,05	1	1	0,019	**1**	0,66	0,013	**1**	1
BP	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	1,98	**0**
Bolts				**Breast**			**Cloud**			**Quacke**		
	DG	Spread	TE	DG	Spread	TE	DG	Spread	TE	DG	Spread	TE
FP1	0,00056	1,17	1	0,00064	1	1	0,017	1,01	1	4,174E-08	1,59	0,8333
FP2	0,00739	1	1	0,2036	1	1	**0**	**0**	**0**	**0**	**1**	**0**
FP3	0,00082	1	1	0,00064	1	1	0,019	1,01	1	**0**	**1**	**0**
BP	**0**	**0**	**0**	**0**	**0**	**0**	0,019	1	1	0,0075	**1**	1

Analizando los resultados obtenidos en cada métrica de calidad, se llegan a las siguientes conclusiones:

- En la métrica **Tasa de Error**, de manera general la técnica basada en Pareto obtiene los mejores resultados en todas las bases de datos, donde se puede observar que la misma aporta un gran porcentaje de soluciones al frente de Pareto verdadero. Esto ocurre debido a que muchas de las soluciones obtenidas en la técnica factores ponderados fueron dominadas por soluciones obtenidas mediante la técnica de factores ponderados.

- Los resultados obtenidos en la métrica **Spread**, indican que el frente de Pareto obtenido por la técnica basada en Pareto se encuentran más cercanos al frente de Pareto verdadero, además de encontrarse mejor distribuidos, y su nivel de cobertura fue mejor que la obtenida por la técnica de factores ponderados. Sin embargo, existen casos como en la base de datos Quacke y Cloud que sus valores llegan a ser muy cercanos.

- Para la métrica **Distancia Generacional**, los mejores valores fueron obtenidos por la técnica basada en Pareto para todas las bases de datos. Estos valores indican que los frentes de Pareto actuales de la técnica basada en Pareto se encuentran más cercanos al frente de Pareto verdadero. Estos resultados tienen mucho sentido, ya que dichos frentes de Pareto actuales fueron los que aportaron mayor cantidad de soluciones al frente de Pareto verdadero.

Posteriormente se realizaron pruebas no paramétricas (Test de Friedman, específicamente) y los posthoc Holm y Finner, con el fin conocer si los valores obtenidos en cada una de las instancias (las tres instancias diseñadas para factores ponderados)

y la técnica de Pareto, poseen diferencias significativas. Para estas pruebas el p-valor definido es 0.05. La herramienta KEEL fue usada para la realización de las pruebas no paramétricas.

Primeramente, se aplica la prueba de Friedman para comprobar si existen diferencias significativas entre un conjunto de algoritmos (en este caso, cada uno de los esquemas multiobjetivos). Dicha prueba obtiene un ranking promedio, el cual permite ordenar los esquemas en cuanto a su rendimiento en las métricas Spread, Distancia generacional y Tasa de error, para cada una de las bases de datos. Luego, con el objetivo de evaluar si el mejor esquema tiene un rendimiento significativamente diferente a las restantes, se aplicaron los test post-hoc de Holm y Finner (siempre que p-valor obtenido en la prueba de Friedman, indique que existen diferencias significativas entre ellos). En el caso de las métricas por separado, los p-valores fueron superiores a 0.05 por lo que no se detecta diferencia significativa entre los métodos, también esto influido por contar con solo 8 valores en comparación. Sin embargo, cuando se consideran todas métricas en su conjuntos (Columna Global de la Tabla 4) se observa que sí hay diferencias en general, y también en cada uno de los posthoc donde los p-valores obtenidos fueron menores que 0,05. Esto indica que indica que la hipótesis nula se rechaza u queda demostrada la superioridad general del esquema basado en Pareto.

Tabla 4: Ranking resultante de las pruebas de Friedman para cada una de las métricas de rendimiento.

Esquema	Distancia Generacional	Spread	Tasa de Error	Global
Factores ponderados 1	2.4375 (2)	3.3125 (4)	3.0625 (4)	2.9583 (4)
Factores ponderados 2	2.6875 (3)	2.4375 (2)	2.625 (2)	2.5833 (2)
Factores ponderados 3	3.125 (4)	2.5625 (3)	2.6875 (3)	2.7708 (3)
Basada en Pareto	1.75 (1)	1.6875 (1)	1.625 (1)	1.6875 (1)
p-valor	0.190037	0.094725	0.142425	**0.003363**

En la **Tabla 4** se muestra el ranking obtenido en cada una de las pruebas de Friedman para cada una de las métricas de rendimiento y en la columna global uniendo todos los resultados. Los resultados obtenidos muestran que el esquema que mejor ranking obtuvo tanto de manera independiente por cada métrica como de manera global fue la técnica basada en Pareto, posteriormente se encuentra el esquema de factores ponderados 2 (con igual peso para ambas funciones objetivos), factores ponderados 3 (0.7 para FPTV y 0.3 para SPD), y por último factores ponderados 1 (0.3 para FPTV y 0.7 para SPD). La última fila de tabla indica los valores de p-valor obtenidos en cada una de las pruebas de Friedman aplicadas. Como puede verse, la diferencia no fue significativa en el análisis de las métricas individualmente, sin embargo a nivel global la diferencia fue significativa. Con los post-hoc de Finner y Li se comprobó (p-valor inferior a 0.05 en todos los casos) que el esquema basado en Pareto es superior significativamente a los otros esquemas en el sentido global. Esto permite afirmar que la técnica basada en Pareto obtuvo en general, resultados mejores que los enfoques basado en factores ponderados, con diferencia estadísticamente significativa.

5.-CONCLUSIONES

En este trabajo se ha realizado una comparación entre dos técnicas para modelar problemas multiobjetivos. Específicamente fueron aplicados al problema de la obtención de predicados difusos como vía para encontrar patrones dentro de una base de datos. Los resultados experimentales en varias bases de datos reales para dos medidas de calidad de estos patrones, demuestran que la técnica basada en Pareto obtiene mejores resultados de manera global, aunque existen casos en los que los resultados resultaron similares a la técnica de factores ponderados.

REFERENCIAS

1. Fernández A, Carmona C, del Jesus M, Herrera F. A view on fuzzy systems for big data: progress and opportunities. International Journal of Computational Intelligence Systems. 2016; 9(sup1): 69-80.
2. Ceruto T, Lapeira O, Rosete A. Quality measures for fuzzy predicates in conjunctive and disjunctive normal forms. Ingeniería e Investigación. 2014; 34(3):63-69.

3. Guerine M, Rosseti I, Plastino A. Extending the Hybridization of Metaheuristics with Data Mining to a Broader Domain. ICEIS. 2014; (1):395-406.
4. Tsai C, Lai C, Chao H, Vasilakos A. Big data analytics: a survey. Journal of Big data. 2015; 2(1): 21.
5. Mukhopadhyay A, Maulik U, Bandyopadhyay S, Coello C. A survey of multiobjective evolutionary algorithms for data mining: Part I. IEEE Transactions on Evolutionary Computation. 2014; 18(1): 4-19.
6. Mukhopadhyay A, Maulik U, Bandyopadhyay S, Coello C. A survey of multiobjective evolutionary algorithms for data mining: Part II. IEEE Transactions on Evolutionary Computation. 2014; 18(1): 20-35.
7. Deb K. Multiobjective Optimization using Evolutionary Algorithms. Search methodologies, Springer; 2014.
8. Kuturel KS, Smith AE. Multiobjective tabu search using a multinominal probability mass function. European Journal of Operational Research. 2016; 169 (14).
9. Bandaru S, Ng A, Deb K. Data mining methods for knowledge discovery in multi-objective optimization: Part A-Survey. Experts Systems with Applications. 2017; 70: 139-159.
10. PhridviRaj M, GuruRao C. Data mining–past, present and future–a typical survey on data streams. Procedia Technology. 2014; 12: 255-263.
11. Aggarwal C, Bhuiyan M, Al Hasan M. Frequent pattern mining algorithms: A survey. In Frequent pattern mining. 2014: 19-64 Springer, Cham.
12. Hussain K, Salleh M, Cheng S, Shi Y. Metaheuristic research: a comprehensive survey. Artificial Intelligence Review. 2018; 1-43.
13. Talbi E. Hybrid metaheuristics for multi-objective optimization. Journal of Algorithms & Computational Technology. 2015; 9(1): 41-63.
14. Nesmachnow S. An overview of metaheuristic: accurate and efficient methods for optimization. International Journal of Metaheuristics. 2014; 3(4): 320-347.
15. Li M, Yang S, Liu X. Diversity Comparison of Pareto Front Approximation in Many-Objective Optimization. IEEE Transaction on Cybernetics. 2014; 44(12): 2568-2584.
16. Ishibuchi H, Setoguchi Y, Masuda H, Nojima Y. Performance of decomposition-based many-objective algorithms strongly depends on Pareto front shapes. IEEE Transactions on Evolutionary Computation. 2017; 21(2): 169-190.
17. Tušar T, Filipič B. Visualization of Pareto front approximations in evolutionary multiobjective optimization: A critical review and the prosection method. IEEE Transactions on Evolutionary Computation. 2015; 19(2): 225-245.
18. Giagkiozis I, Fleming P. Pareto front estimation for decision making. Evolutionary computation .2014; 22(4): 651-678.
19. Hancer E, Xue B, Zhang M, Karaboga D, Akay B. Pareto front feature selection based on artificial bee colony optimization. Information Sciences. 2018; 422: 462-479.
20. Chapman J, Lu L,Anderson-Cook C. Process optimization for multiple responses utilizing the Pareto front approach. Quality Engineering. 2014; 26(3):253-268.
21. Xiong N, Molina D, Ortiz M, Herrera F. A walk into Metaheuristics for Engineering Optimization: Principles, Methods and Recents Trends. International Journal of Computational Intelligence Systems. 2015;8 (4):606-636.
22. Infante A, Andre M, Rampesaud L. Conformación de equipos de proyectos de software aplicando algoritmos metaheurísticos de trayectoria multiobjetivo. Inteligencia Artificial, 2015 17 (54), 1-16.
23. Joyce T, Herrmann M. A review of no free lunch theorems, and their implications for metaheuristic optimization. In Nature-Inspired Algorithms and Applied Optimization. 2018; 27-51. Springer.
24. Alabert A, Berti A, Caballero R, Ferrante M. No-free-lunch theorems in the continuum. Theoretical Computer Science. 2015; 600: 98-106.
25. Yu W, Li B, Jia H, Zhang M, Wang D. Application of multi-objective genetic algorithm to optimize energy efficiency and thermal comfort in building design. Energy and Building. Elsevier. 2015, 88: 135-143.
26. Zuo X. Vehicle Scheduling of an Urban Bus Line via an Improved Multiobjective Genetic Algorithm. IEEE Transaction on Intelligent Transportation Systems. 2015; 16(2).

AUTORES

Orenia Lapeira Mena, recibió el Máster en Informática Aplicada en la Universidad Tecnologica de la Habana "José A Echeverría" (CUJAE), en la Habana en el año 2016. Profesora Asistente del Departamento de Ingeniería de Software (DIS) en la CUJAE. Profesora del Grupo de Investigación SCoDA (SoftComputing and Data Analysis). Sus áreas de interés incluyen la minería de datos, los predicados en forma normal, la extracción de conocimiento basado en metaheurísticas y los sistemas difusos. E-mail: olapeira@ceis.cujae.edu.cu

Taymi Ceruto Cordovés, recibió el Máster en Informática Aplicada en la Universidad Tecnológica de la Habana "José A Echeverría" (CUJAE); en el año 2010 y el título de Doctor en Ciencias en el 2014, respectivamente. Profesora Asistente. Pertenece al Grupo de Investigación SCoDA (SoftComputing and Data Analysis). Sus áreas de interés incluyen la minería de datos, los predicados en forma normal, la extracción de conocimiento basado en metaheurísticas y los sistemas difusos. E-mail: taymiceruto@gmail.com

Alejandro Rosete Suárez, recibió el Máster en Informática Aplicada en la Universidad Tecnológica de la Habana "José A Echeverría" (CUJAE); en el año 1995 y el título de Doctor en Ciencias en el 2000, respectivamente. Profesor Titular del Departamento de Inteligencia Artificial e Infraestructura de Sistemas Informáticos (DIAISI). Es co-jefe del Grupo de Investigación SCoDA (SoftComputing and Data Analysis). Sus áreas de interés incluyen las metaheurísticas, ingeniería de software basada en agentes inteligentes, modelos de decisión, minería de datos, extracción de conocimiento basado en metaheurísticas, entre otros. E-mail: rosete@ceis.cujae.edu.cu