Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 140-155
http://rcci.uci.cu

Tipo de artículo: Artículo de revisión
Temática: Inteligencia artificial
Recibido: 7/05/2013 | Aceptado: 2/10/2013 | Publicado: 21/01/2014

Selección de atributos relevantes aplicando algoritmos que combinan conjuntos aproximados y optimización en colonias de hormigas

Feature selection applying algorithms base on rough set and ant colony optimization

Yanela Rodríguez*[1], Yumilka Fernández[1], Rafael Bello[2], Yailé Caballero[1]

[1] Departamento de Computación. Universidad de Camagüey, Cuba
[2] Universidad Central "Marta Abreu" de las Villas, Carretera a Camajuaní, km 5 ½, Santa Clara, Villa Clara, Cuba

{yanela.rodriguez, yumilka.fernandez, yaile.caballero}@reduc.edu.cu; rbellop@uclv.edu.cu

Resumen

La selección de atributos relevantes puede ser vista como uno de los problemas más importantes en el campo del aprendizaje automático. En esta investigación se hace un análisis sobre los métodos de selección de atributos; haciendo énfasis en aquellos que emplean técnicas de Optimización en Colonias de Hormigas (ACO) y la Teoría de los Conjuntos Aproximados (RST). Se propone además, un sistema que permite la generación automatizada de los subconjuntos de rasgos principales que describen los datos, a través de cualquiera de los algoritmos tratados en esta investigación. Por otro lado se implementaron e incluyeron en el sistema algoritmos como el clásico QUICKREDUCT y otros encontrados en la bibliografía. Para verificar la eficiencia de los métodos estudiados se llevaron a cabo experimentos con bases de casos internacionales y se realizaron comparaciones con otros métodos. Además estos métodos se aplicaron en el preprocesamiento de los datos para pronosticar, de forma automatizada, las

temperaturas diarias en el Centro Meteorológico de Camagüey. Los resultados demostraron que los algoritmos implementados proveen una solución eficiente al problema de selección de rasgos.
Palabras clave: Selección de atributos relevantes, aprendizaje automático, optimización en colonias de hormigas, conjuntos aproximados.

Abstract

Feature selection can be viewed as one of the most fundamental problems in the field of machine learning. An analysis on the methods of feature selection is done in this investigation; stressing those that use techniques of Ant Colony Optimization and the Rough Set Theory. Also, in this investigation it is proposed a system that allows the generation automatized of the subsets of principal features that describe the data, through any of algorithms studied in this investigation. Moreover, algorithms were implemented and included in the system, like the classical QUICKREDUCT and some others found in the bibliography. To verify the efficiency of the methods studied, experiments were carried out on some standard international datasets and comparisons with other methods were made. Also these methods were applied in the pre-processing of data to predict, automatically, the daily temperatures in Camagüey's Meteorologic Center. The results demonstrated that these algorithms can provide efficient solution to find a minimal subset of the features.
Keywords: *Feature selection, machine learning, ant colony optimization, rough sets.*

Introducción

La selección de rasgos consiste en encontrar el subconjunto de atributos del conjunto de datos original que mejor describe los objetos del dominio; tiene como meta reducir la dimensionalidad del conjunto de rasgos a través de la selección del subconjunto de rasgos de mejor desempeño bajo algún criterio de clasificación (H. Liu & Motoda, 2007). Este proceso de selección se hace eliminando rasgos irrelevantes y redundantes (Bell & Wang, 2000; Blum & Langley, 1997), proporcionando así una mejor representación de la información original reduciendo significativamente el costo computacional y contribuyendo a una mejor generalización del algoritmo de aprendizaje. Normalmente este proceso está presente en las etapas previas de las principales tareas de la minería de datos, ya sean supervisadas o no (Liu & Yu, 2005).

La selección de atributos es un campo de investigación y desarrollo productivo desde los años setenta, donde confluyen distintas áreas como el reconocimiento de patrones (Dudani, 1975; R. Kohavi & Frasca, 1994; Yang & Honavar, 1998), el aprendizaje automático (Breiman, Friedman, Olshen, &Stone, 1984; Hall, 1999; R Kohavi & John,

1997; Koller & Sahami, 1996; Kudo, Somol, Pudil, Shimbo, & Sklansky, 2000) y la minería de datos (Davies & Russell, 1994; Liu & Yu, 2002). Las técnicas de selección de características se aplican en muchos entornos diferentes, como por ejemplo la clasificación de textos, recuperación de imagen (Siedlecki & Sklansky, 1988) y bioinformática. Se hace constar, que el proceso de selección de atributos, además de preceder a la clasificación, suele estar presente en las etapas previas de las principales tareas de la minería de datos, ya sean supervisadas o no, como regresión, agrupamiento y reglas de asociación (Ruiz, 2005).

Los procedimientos de selección de rasgos constan de dos componentes principales: la función de evaluación y el método de generación de subconjuntos (basado en un proceso de búsqueda). Existen diferentes enfoques y técnicas para seleccionar atributos relevantes, tales como las técnicas de Optimización mediante Colonias de Hormigas (del inglés, *Ant Colony Optimization*, ACO) y las basadas en la Teoría de los Conjuntos Aproximados (del inglés, *Rough Sets Theory*, RST).

La RST por Z. Pawlak en 1982 (Pawlak, 1982). La filosofía de los conjuntos aproximados se basa en aproximar cualquier concepto, un subconjunto duro del dominio como por ejemplo una clase en un problema de clasificación supervisada, a un par de conjuntos exactos llamados aproximación inferior y aproximación superior del concepto. Con esta teoría es posible tratar tanto datos cuantitativos como cualitativos y no se requiere eliminar las inconsistencias previas al análisis; respecto a la información de salida puede ser usada para determinar la relevancia de los atributos y generar las relaciones entre ellos (Choubey, 1996; Chouchoulas & Shen, 1999; Greco & Inuiguchi, 2003; Grzymala-Busse & Siddhaye, 2004; Miao & Hou, 2003; Midelfart & Komorowski, 2003; Piñero & Arco, 2003; Sugihara & Tanaka, 2006; Tsumoto, 2003; Zhao & Zhang, 2003). La inconsistencia describe una situación en la cual hay dos o más valores en conflicto para ser asignados a una variable (Parsons, 2006).

Sobre los Conjuntos Aproximados se han manifestado diversos autores, los cuales ven esta teoría como la mejor herramienta para modelar la incertidumbre cuando esta se manifiesta en forma de inconsistencia, y como una nueva dirección en el desarrollo de teorías sobre la información incompleta (Grabowski, 2003; Skowron, 1999; Skowron & Peters, 2003). La principal ventaja que tiene el análisis de datos basado en RST es que para operar este no requiere parámetros adicionales además de los datos de entrada (Düntsch & Gediga, 2000).

Los algoritmos de ACO reproducen el comportamiento de las hormigas reales en una colonia artificial. Estos han sido aplicados a un gran número de problemas cuyas soluciones generan explosión combinatoria como el clásico del vendedor ambulante, problemas de ruteo en redes de telecomunicaciones, planificación de tareas, etcétera. En (Jensen

& Shen, 2003) se plantea el uso de estas técnicas para el cálculo de reductos debido a que las hormigas pueden descubrir las mejores combinaciones de atributos en la medida en que atraviesan el grafo.

Los algoritmos ACO son procesos iterativos. En cada iteración se "lanza" una colonia de m hormigas y cada una de ellas construye una solución al problema. Las hormigas construyen las soluciones de manera probabilística, guiándose por un rastro de feromona artificial y por una información calculada a priori de manera heurística. Estos algoritmos son esencialmente métodos constructivos: en cada iteración del algoritmo, cada hormiga construye una solución al problema recorriendo un grafo. Cada arista del grafo, que representa los posibles caminos que la hormiga puede tomar, tiene asociados dos tipos de información que guían el movimiento de la hormiga:

- *Información heurística*, mide la preferencia heurística de moverse desde el nodo i hasta el nodo j; es decir, la preferencia a recorrer la arista a_{ij}. Se denota por n_{ij}. Las hormigas no modifican esta información durante la ejecución del algoritmo.
- *Información de los rastros artificiales de feromona*, mide la "deseabilidad aprendida" del movimiento de i a j. Imita de forma numérica a la feromona real que depositan las hormigas naturales. Esta información se modifica durante la ejecución del algoritmo dependiendo de las soluciones encontradas por las hormigas. Se denota por τ_{ij}.

Desarrollo

Optimización en Colonias de Hormigas y conjuntos aproximados aplicados a la selección de rasgos

Algoritmo AS-RST-FS

AS-RST-FS dado en (Bello, Nowé, Caballero, Gómez, & Vrancx, 2005), está basado en el algoritmo Sistema de Hormigas (AS) desarrollado por Dorigo en su tesis doctoral en 1992 (Dorigo, 1992), el primer algoritmo de ACO. En AS-RST-FS, ACO es utilizada para generar subconjuntos de rasgos empleando una aproximación de filtro basada en selección hacia adelante. RST ofrece la función heurística para medir la calidad de un subconjunto de rasgos.

Algoritmo ACS-RST-FS

Este algoritmo dado en (Bello, *et al*., 2005), aunque está basado en el algoritmo Sistema de Colonias de Hormigas (ACS), es similar al anterior. Sin embargo para esta variante se utiliza la regla de transición probabilística de ACS. Además es diferente la forma en la que se actualiza el valor de la feromona, el valor de la feromona se actualiza de forma local y luego de forma global. Cada vez que un nodo correspondiente a un rasgo sea adicionado a un

subconjunto se actualizará el valor de la feromona y además, se actualizará también el valor de la feromona para el mejor subconjunto generado en un ciclo.

Algoritmo AFSBRSACO

AFSBRSACO dado en (Ming, 2008) es un algoritmo híbrido en el que RST se utiliza para definir la importancia de los rasgos mediante las aproximaciones superior e inferior. Estas aproximaciones son empleadas como heurísticas para guiar el proceso de selección de rasgos. Por otro lado, ACO se utiliza para implementar el método de búsqueda; generando subconjuntos de rasgos que usan una aproximación de filtro basada en la selección hacia adelante. En el caso de este algoritmo las hormigas parten del *CORE* o núcleo. Además, la feromona está asociada a los arcos denotando la posibilidad de ir a un nodo j desde un nodo i.

Algoritmo RSFSACO

En RSFSACO dado en (Chen, Miao, & Wang, 2010) la información heurística se calcula dinámicamente durante el proceso de construcción de las soluciones. La importancia de los rasgos, definida por la entropía y la información mutua, se adopta como una información heurística.

Para construir una solución cada hormiga debe comenzar a partir del núcleo de rasgos. En el siguiente paso la hormiga selecciona aleatoriamente un rasgo y luego selecciona el siguiente rasgo de aquellos que quedan sin seleccionar con una alta probabilidad. La probabilidad es calculada por la formula dada en (Dorigo, Maniezzo, & Colorni, 1996). Después de que cada hormiga haya construido una solución, se deberá actualizar la feromona de cada arista. Si el subconjunto optimo es encontrado o las iteraciones alcanzan su máximo ciclo, entonces el algoritmo para y devuelve el mínimo reducto de rasgos encontrado. Si ninguna condición se cumple la feromona se actualiza, se crea un nuevo conjunto de hormigas y el proceso itera una vez más.

Algoritmo ACO-RST-FSP

Este método que se propone en (Gómez, 2010) se clasifica como "filtro"[1], utiliza ACO como procedimiento de generación de subconjuntos y como función de evaluación de la calidad de los subconjuntos la medida calidad de la

[1] algoritmos en los que la selección de atributos se realiza como un preprocesado independiente de la fase de inducción, por lo que puede entenderse como un filtrado de los atributos.

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 140-155
http://rcci.uci.cu

clasificación de RST. Durante la ejecución del algoritmo cada hormiga construye un subconjunto de rasgos hasta que este alcance un valor de la calidad de la clasificación igual al calculado para el conjunto de todos los rasgos, es decir, hasta formar un reducto.

Propuesta del sistema

El sistema SAICCA (**S**elección de **A**tributos con **I**nteligencia **C**olectiva y **C**onjuntos **A**proximados) es una aplicación de escritorio que cuenta con una interfaz amigable y a la vez fácil de utilizar. Fue desarrollado haciendo uso de la plataforma NetBeans IDE 7.1 RC1 y el lenguaje de programación de alto nivel Java. Es un sistema automatizado que permite la selección de atributos relevantes de una base de casos, a través de cualquiera de los algoritmos tratados anteriormente. Permite al usuario modificar los parámetros necesarios para la ejecución de los algoritmos, y el registro de toda la información que interviene en este análisis. A través de esta aplicación se logra obtener de manera rápida y precisa el subconjunto de atributos relevantes que describe la base de casos, manteniendo la calidad de la clasificación. Además se brindan, al usuario, reportes donde se muestran los resultados obtenidos por cada uno de los algoritmos utilizados.

Resultados y discusión

Resultados experimentales

Para realizar los experimentos se utilizaron 10 conjuntos de datos reconocidos internacionalmente. Estos conjuntos son provenientes del repositorio, para aprendizaje automatizado, disponibles en el sitio ftp de la Universidad de Irvine, California, del sitio personal de Jensen. Las características de estos conjuntos de datos aparecen en la Tabla 1. Para el análisis estadístico de los resultados se utilizó la prueba de Friedman, la cual permite detectar diferencias estadísticamente significativas entre un grupo de resultados.

Tabla 1. Descripción de los conjuntos de datos internacionales que se usaron en los experimentos.

Bases de Casos	Cantidad de rasgos	Cantidad de instancias	Cantidad de clases	Ausencia de información
Audiology	69	226	24	Si
Biomed	8	194	2	No
Bupa	6	345	2	No
Cleveland	13	303	5	No
Heart-statlog	13	270	2	No

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 140-155
http://rcci.uci.cu

Ionosphere	34	351	2	No
Iris	4	150	3	No
New Thyroid	5	215	3	No
Wine	13	178	3	No
Zoo	16	101	7	No

Experimento 1: Demostrar que se logran reducciones significativas del conjunto de atributos cuando se aplican los métodos estudiados de construcción de reductos. Determinar cuál de estos métodos obtiene mejores resultados respecto a la longitud promedio de los reductos encontrados y al tiempo de ejecución. Los resultados experimentales están resumidos en las Tablas 2 y 3.

Tabla 2. Longitud de los reductos obtenidos por los diferentes métodos de cálculo de reductos.

Bases de Casos	**AS-RST-FS**	**ACS-RST-FS**	**AFSBRSACO**	**RSFSACO**	**ACS-RST-FSP**
Audiology	3	3	4	4	2
Biomed	2	2	3	2	2
Bupa	3	3	4	3	3
Cleveland	3	3	5	5	3
Heart-statlog	3	3	5	4	3
Ionosphere	2	2	4	5	2
Iris	3	3	4	3	3
New Thyroid	2	2	3	2	2
Wine	2	2	3	2	2
Zoo	5	5	7	7	5

Tabla 3. Tiempo de ejecución (s) obtenido por los diferentes métodos de cálculo de reductos.

Bases de Casos	**AS-RST-FS**	**ACS-RST-FS**	**AFSBRSACO**	**RSFSACO**	**ACS-RST-FSP**
Audiology	8395.469	1584.266	35.297	2088.922	4772.719
Biomed	31.219	2.875	0.703	4.734	0.22304
Bupa	62.969	18.156	3.094	3.281	1193.60002
Cleveland	875.687	191.422	17.547	19.922	3.51756
Heart-statlog	382.531	32.968	2.782	17.609	2.09064
Ionosphere	1765.453	310.953	24.047	1848.969	44.828
Iris	3.157	1.859	0.001	0.109	0.03192

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 140-155
http://rcci.uci.cu

New Thyroid	16.078	2.141	0.516	0.375	0.09928
Wine	110.172	27.687	3.234	15.718	0.47308
Zoo	198.781	43	1.688	7.047	1.77192

Se demostró que existe una reducción significativa del conjunto de atributos cuando se aplican los métodos de construcción de reductos estudiados. Se aplicó el test de Friedman y este arrojó que existen diferencias significativas entre los métodos estudiados respecto a la longitud promedio de los reductos encontrados y al tiempo de ejecución de estos métodos. Luego de este experimento se pudo concluir que el método que obtiene resultados globales óptimos es el ***ACS-RST-FSP***.

Experimento 2: Determinar si existen diferencias significativas entre los algoritmos estudiados en esta investigación y el método AttributeSelection (de la herramienta Weka con varias combinaciones entre los diferentes evaluadores de atributos y los diferentes métodos de búsqueda), de acuerdo a la longitud de los reductos obtenidos. Los resultados experimentales están resumidos en las Tabla 4.

Tabla 4. Longitud de los reductos obtenidos por los métodos estudiados y los métodos de Weka para el cálculo de reductos.

Bases de Casos	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Audiology	3	3	4	4	2	16	13	33	16
Biomed	2	2	3	2	2	7	6	6	6
Bupa	3	3	4	3	3	1	1	1	1
Cleveland	3	3	5	5	3	6	9	7	4
Heart-statlog	3	3	5	4	3	8	10	7	6
Ionosphere	2	2	4	5	2	6	7	14	4
Iris	3	3	4	3	3	2	2	2	2
New Thyroid	2	2	3	2	2	5	5	5	5
Wine	2	2	3	2	2	13	5	11	10
Zoo	5	5	7	7	5	10	5	9	6

Leyenda:
A-RSFSACO, **B**- AS-RST-FS, C-ACS-RST-FS, **D**-AFSBRSACO, **E**-ACS-RST-FSP, **F**-CfsSubsetEval-ExhaustiveSearch, **G**-ConsistencySubsetEval-BestFirst, **H**-CfsSubsetEval-GeneticSearch, **I**-SymmetricalUncertAttributeSetEval-FCBFSearch.

Se aplicó el test de Friedman y este arrojó que existen diferencias significativas entre los métodos estudiados y los métodos AttributeSelection (de la herramienta Weka con varias combinaciones entre los diferentes evaluadores de atributos y los diferentes métodos de búsqueda) respecto a la longitud promedio de los reductos encontrados y ellos. Luego de este experimento se pudo concluir que los métodos estudiados son significativamente superiores a los

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 140-155
http://rcci.uci.cu

AttributeSelection (de la herramienta Weka con varias combinaciones entre los diferentes evaluadores de atributos y los diferentes métodos de búsqueda).

Experimento 3: Comparar la calidad de la clasificación, de RST, obtenida con todos los atributos de los conjuntos de datos y con los conjuntos de atributos relevantes seleccionados por los métodos estudiados y el método AttributeSelection (de la herramienta Weka con varias combinaciones entre los diferentes evaluadores de atributos y los diferentes métodos de búsqueda). Los resultados experimentales están resumidos en las Tablas 5.

Tabla 5. Calidad de la clasificación utilizando los reductos obtenidos por los métodos estudiados y los de Weka.

Bases de Casos	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***
Audiology	1	1	0.07	0.07	0.07	1	0.95	1	1	1
Biomed	1	1	1	1	1	1	1	1	1	1
Bupa	1	1	1	1	1	1	0.22	0.22	0.22	0.22
Cleveland	1	1	1	1	1	1	0.98	1	1	0.65
Heart-statlog	1	1	1	1	1	1	1	1	1	1
Ionosphere	1	1	1	1	1	1	1	1	1	0.99
Iris	1	1	1	1	1	1	0.98	0.98	0.98	0.98
New Thyroid	1	1	1	1	1	1	1	1	1	1
Wine	1	1	1	1	1	1	1	1	1	1
Zoo	1	1	1	1	1	1	1	1	1	0.81

Se aplicó test de Friedman y arrojó que existen diferencias significativas entre los métodos estudiados y los métodos

Leyenda: A - BC Original, **B** – RSFSACO, **C** - AS-RST-FS, **D** - ACS-RST-FS, **E** – AFSBRSACO, **F**- ACS-RST-FSP, **G** - CfsSubsetEval-ExhaustiveSearch, **H** - ConsistencySubsetEval-BestFirst, **I** - CfsSubsetEval-GeneticSearch, **J**- SymmetricalUncertAttributeSetEval-FCBFSearch.

Luego de este experimento se pudo concluir que con los reductos obtenidos con los métodos estudiados se mantiene la medida calidad de la clasificación de RST. Mientras que por otro lado, con los reductos obtenidos con los métodos AttributeSelection (de la herramienta Weka con varias combinaciones entre los diferentes evaluadores de atributos y los diferentes métodos de búsqueda) esta medida disminuye considerablemente.

Para dar por concluido este estudio experimental se decidió llevar a cabo una prueba de Friedman para calcular el ranking de los algoritmos con el objetivo de determinar el mejor algoritmo, es decir, el de más alto ranking. Los

resultados de esta prueba se pueden ver en la **Tabla 6**, donde es posible observar que el mejor algoritmo de los estudiados es el **ACS-RST-FSP**.

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 140-155
http://rcci.uci.cu

Tabla 6. Ranking de los algoritmos (Friedman).

Algoritmo	Ranking
ACS-RST-FSP	3
RSFSACO	3.15
AS-RST-FS	3.15
AFSBRSACO	4.7
SymmetricalUncertAttributeSetEval-FCBFSearch	5.35
ACS-RST-FS	5.9
ConsistencySubsetEval-BestFirst	6
CfsSubsetEval-GeneticSearch	6.85
CfsSubsetEval-ExhaustiveSearch	6.9

Aporte práctico al problema de pronóstico de las temperaturas diarias en el Centro Meteorológico de Camagüey

La calidad y la precisión de las predicciones y avisos meteorológicos se imponen para el desarrollo de cualquier país, por su utilidad en distintos sectores socioeconómicos (sector agrícola, energético, salud ciudadana, entre otros) que precisan de este tipo de predicción para poder evaluar a corto y medio plazos sus políticas de actuación ante situaciones climatológicas adversas. La temperatura del aire (temperatura, como se le conoce comúnmente), como medida del contenido de calor del medio aéreo, es uno de los elementos climáticos más importantes, pues resulta un elemento indispensable para la planificación adecuada de muchas de las actividades básicas del hombre, incluyendo hasta su vestuario.

El Centro Meteorológico de Camagüey, en coordinación con el Centro Nacional de Meteorología, ha facilitado los datos correspondientes a las seis estaciones meteorológicas para las variables temperaturas máximas y mínimas y otras que inciden en las variaciones de la temperatura. Se cuenta con los valores reales diarios de estos datos en el período comprendido entre los años 2007-2011, así como los valores de las temperaturas pronosticados por el Departamento de Pronósticos de Camagüey.

El comportamiento de las temperaturas máximas y mínimas de un día específico está estrechamente relacionado con los siete días anteriores (Lecha & Florido, 1989); por este motivo para construir las bases de casos, se tuvieron en cuenta los valores de las variables temperatura máxima y temperatura mínima de los siete días que le anteceden al que se desea pronosticar entre otros factores de impacto. De esta manera, se tienen 26 atributos predictores y un atributo

Revista Cubana de Ciencias Informáticas
Vol. 8, No. 1, enero-marzo, 2014
ISSN: 2227-1899 | RNPS: 2301
Pág. 140-155
http://rcci.uci.cu

objetivo para cada caso, ver Tabla 7. Este es un típico problema de clasificación supervisada, donde se quiere predecir la clase correspondiente a los valores de las temperaturas máximas y mínimas, dado un nuevo objeto.

Tabla 7. Descripción de los conjuntos de datos de Meteorología.

Bases de Casos	Cantidad de rasgos	Cantidad de instancias	Cantidad de clases
Camagüey07	26	366	13
Camagüey08	26	365	13
Camagüey09	26	357	13
Camagüey10	26	358	13
Camagüey11	26	359	13

Teniendo en cuenta las características de los datos, es viable la aplicación de los algoritmos descritos en esta investigación con el objetivo de eliminar aquellos atributos irrelevantes que pueden influir de manera nociva en la posterior asertividad del clasificador a la hora de predecir los valores de temperaturas diarias. Al aplicar los algoritmos se obtuvieron resultados favorables, pues se logra reducir significativamente el número de atributos de la base de casos sin afectar la calidad de la clasificación. En la **Tabla 8** se muestran el tiempo de ejecución (TE) y el tamaño de los reductos (TR) obtenido con dichos métodos.

Tabla 8. Resultados de los algoritmos de selección de rasgos estudiados aplicados al caso de estudio de pronóstico de las temperaturas diarias.

	AS-RST-FS		***ACS-RST-FS***		***AFSBRSACO***		***RSFSACO***		***ACS-RST-FSP***	
BC	**TE (s)**	**TR**	**TE (s)**	**TR**	**TE (s)**	**TR**	**TE (s)**	**TR**	**TE (s)**	**TR**
Camagüey07	4153.73	2	1885.88	2	22.09	3	404.19	4	18.24	2
Camagüey08	4478.93	2	2155.70	3	21.84	3	670.36	3	18.95	2
Camagüey09	4506.35	3	1748.03	3	106.50	4	781.02	3	19.11	3
Camagüey10	4581.62	3	2487.92	3	14.50	3	614.50	3	15.98	3
Camagüey11	5752.31	3	2131.69	3	14.17	3	703.53	3	25.30	3

Teniendo en cuenta los resultados experimentales obtenidos en la tabla anterior; se llegó a la conclusión que el algoritmo con mejores resultados globales era el algoritmo **ACS-RST-FSP**. Finalmente, se decidió utilizar este en el preprocesamiento de los datos.

Conclusiones

Como resultado de esta investigación se realizó un estudio detallado del comportamiento de los algoritmos que combinando ACO y RST, proporcionan a los investigadores otras alternativas que permitan encontrar subconjuntos reducidos de atributos, capaces de representar la información necesaria en problemas de aprendizaje supervisado. Particularmente se encontraron buenas soluciones al aplicar estos algoritmos en el preprocesamiento de los datos para optimizar el pronóstico de las temperaturas diarias en el Centro Meteorológico de Camagüey. Los métodos implementados en la investigación para el cálculo de reductos logran reducciones altamente significativas de la cantidad de atributos respecto al conjunto original de datos, mientras que la medida calidad de la clasificación de RST no se vio afectada por los conjuntos de atributos reducidos. Los algoritmos incluidos en el sistema SAICA, en la mayoría de los casos, superan el desarrollo de los métodos de selección de rasgos de la herramienta Weka. Estos resultados están apoyados por las pruebas estadísticas no paramétricas realizadas. En la solución al problema del pronóstico automatizado de las temperaturas máximas y mínimas del Centro Meteorológico de Camagüey, se realizó un preprocesamiento de los datos para seleccionar los atributos relevantes. Cualquiera de los reductos encontrados permiten a especialistas meteorólogos determinar qué variables observar con potencia suficiente para un buen pronóstico. Se desarrolló una herramienta informática que permite la selección de rasgos relevantes de una base de casos, a través de algoritmos basados en ACO y RST; opción esta que no encontraremos en herramientas que implementan técnicas para el aprendizaje automático y la minería de datos como KEEL y WEKA.

Referencias

- BELL, D., & WANG, H. A Formalism for Relevance and its Application in Feature Subset Selection. *Machine Learning*. 2000.
- BELLO, R., NOWÉ, A., CABALLERO, Y., GÓMEZ, Y., & VRANCX, P. *A Model Based on Ant Colony System and Rough Set Theory to Feature Selection*. Paper Presented at the Genetic and Evolutionary Computation Conference (GECCO05). 2005.
- BLUM, A., & LANGLEY, P. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence*. 1997.
- BREIMAN, L., FRIEDMAN, J., OLSHEN, R., & STONE, C. *Classification and Regresion Trees*. Paper Presented at the Wadsworth Int. Group. 1984.

- CHEN, Y., MIAO, D., & WANG, R. A Rough set Approach to Feature Selection Based on Ant Colony Optimization. *Pattern Recognition Letters, 31*, p. 226–233. 2010.
- CHOUBEY, S. K. *A Comparison of Feature Selection Algorithms in the Context of Rough Classifiers*. Paper Presented at the Fifth IEEE International Conference on Fuzzy Systems. 1996.
- CHOUCHOULAS, A., & SHEN, Q. A Rough Set-Based Approach to Text Classification. *Lectures Notes in Artificial Intelligence, 1711*, p. 118-127. 1999.
- DAVIES, S., & RUSSELL, S. *Np-Completeness of Searches for Smallest Possible Feature Sets*. Paper Presented at the AAAI Fall Symposium on Relevance. 1994.
- DORIGO, M. *Optimization, Learning and Natural Algorithms.* Politecnico di Milano. 1992.
- DORIGO, M., MANIEZZO, V., & COLORNI, A. The Ant System: Optimization by a Colony of Cooperating Agents. *IEEE Trans. Syst. Man Cybernet, Part B 26 (1)*, 29-41. 1996.
- DUDANI, S. The Distance-Weighted k-Nearest-Neighbor Rule. *Man and Cybernetics*. 1975.
- DÜNTSCH, I., & GEDIGA, G. ROUGH Set Data Analysis: A Road to Non-Invasive Knowledge Discovery. *Methodos Publishers*. 2000.
- GÓMEZ, M. Y. *Algoritmos que combinan conjuntos aproximados y optimización basada en colonias de hormigas para la selección de rasgos. Extensión a múltiples fuentes de datos.* Unpublished Tesis Doctoral, Universidad Central "Marta Abreu" de Las Villas, Santa Clara. 2010.
- GRABOWSKI, A. Basic Properties of Rough Sets and Rough Membership Function. *Journal of Formalized Mathematics, 15*. 2003.
- GRECO, S., & INUIGUCHI, M. *Rough Sets and Gradual Decision Rules. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. Paper Presented at the 9th International Conference RSFDGRC 2003.
- GRZYMALA-BUSSE, J. W., & Siddhaye, S. *Rough Set Approaches to Rule Induction from Incomplete Data*. Paper Presented at the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Bases systems IPMU 2004.
- HALL, M. *Correlation-Based Feature Selection for Machine Learning.* University of Waikato, Hamilton, New Zealand. 1999.

- JENSEN, R., & SHEN, Q. *Finding Rough Set Reducts with Ant Colony Optimization.* Paper presented at the UK Workshop on Computational Intelligence. 2003.
- KOHAVI, R., & FRASCA, B. *Useful Feature Subsets and Rough Set Reducts*. Paper Presented at the 3rd Int. Workshop on Rough Set and Soft Computing. 1994.
- KOHAVI, R., & JOHN, G. Wrappers for Feature Subset Selection. *Artifcial Intelligence*. 1997.
- KOLLER, D., & SAHAMI, M. *Toward Optimal Feature Selection*. Paper Presented at the 13th Int. Conf. on Machine Learning. 1996.
- KUDO, M., SOMOL, P., PUDIL, P., SHIMBO, M., & SKLANSKY, J. Comparison of Classifier Specific Feature Selection Algorithms. p. 677-686. 2000.
- LECHA, L., & FLORIDO, A. *Principales características climáticas del régimen térmico del archipiélago cubano*. La Habana, Cuba. 1989.
- LIU, H., & MOTODA, H. Computational Methods of Feature Selection. 2007.
- LIU, H., & YU, L. *Feature Selection for data Mining* (Technical report). Temp, Arizona: Arizona State University. 2002.
- LIU, H., & YU, L. Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Trans. On Knowledge and data Enginnering, 17,* p. 1-12. 2005
- MIAO, D., & HOU, L. *An Application of Rough Sets to Monk's Problems Solving. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. Paper presented at the 9th International Conference, RSFDGRC 2003.
- MIDELFART, H., & KOMOROWSKI, J. Learning Rough Set Classifiers from Gene Expression and Clinical data. *Fundamenta Informaticae, 53*, 155-183. 2003.
- MING, H. *Feature Selection Based on Ant Colony Optimization and Rough Set Theory*. Paper presented at the International Symposium on Computer Science and Computational Technology. 2008.
- PARSONS, S. Current Approaches to Handling Imperfect Information in Data and Knowledges Bases. *IEEE Transaction On knowledge and data enginnering*. 2006.

- PAWLAK, Z. ROUGH Sets. *International Journal of Computer and Information Sciences, 11*, 341-356. 1982.
- PIÑERO, P., & ARCO, L. Two New Metrics for Feature Selection in Pattern Recognition. *Lectures Notes in Computer Science LNCS 2905*, p. 488-497. 2003.
- RUIZ, D. R. *Selección de Atributos mediante proyecciones.* Unpublished Tesis Doctoral, Universidad de Sevilla, Sevilla. 2005.
- SIEDLECKI, W., & SKLANSKY, J. On Automatic Feature Selection. *Int. Journal of Pattern Recognition and Artifcial Intelligence, 2*, p. 197-220. 1988.
- SKOWRON, A. *New Directions in Rough Sets*. Paper Presented at the 7th International Workshop (RSFDGRC'99). 1999.
- SKOWRON, A., & PETERS, J. F. *Rough Sets: Trends and Challenges*. Paper presented at the Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing 9th International Conference, RSFDGRC 2003.
- SUGIHARA, K., & TANAKA, H. *Rough Sets Approach to Information Systems with Interval Decision Values in Evaluation Problems*. Paper presented at the The International Symposium on Fuzzy and Rough Sets ISFUROS 2006.
- TSUMOTO, S. Automated Extraction of Hierarchical Decision Rules from Clinical Databases Using Rough Set Model. *Expert systems with Applications, 24*, p. 189-197. 2003.
- YANG, J., & HONAVAR, V. Feature Extraction, Construction and Selection. In K. A. Publishers Ed. p. 117-136. 1998.
- ZHAO, Y., & ZHANG, H. *Classification Using the Variable Precision Rough Set. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. Paper presented at the 9th International Conference, RSFDGRC 2003.