Universidad Central "Marta Abreu" de Las Villas.

Facultad de Matemática, Física y Computación.

Departamento de Ciencia de la Computación.

Validación estadística del algoritmo ACO-RST-FSP y su variante en contexto distribuido.

Autora: Dainerys Pérez Pérez

Tutor: MSc. Yudel R. Gómez Díaz

Santa Clara, 2010

"Año 52 de la Revolución"

A todas las personas que de una manera u otra creyeron en mí, en especial a toda mi familia por estar pendientes en todo momento.

A los que empezaron conmigo este proyecto y no lograron ver su fin, a mi tía Cary.

Agradezco a todas las personas que de cualquier manera apoyaron la realización de este proyecto:

A mis padres por toda su confianza y apoyo incondicional, a mis abuelos, en especial a mi abuela Angelina por ofrecerme la serenidad necesaria a la hora de afrontar nuevos retos.

A toda mi gran familia por estar de mi lado en todo momento, en especial al "batallón" con el que vivo.

A mi pareja en la vida sentimental, por su paciencia y apoyo en todos estos años.

A mi tutor de tesis por ser mi guía en este último período.

A todos los profesores y personas que han contribuido a mi formación a lo largo de mi vida como estudiante y como ser humano.

A todos mis amigos, a los que no nombro por temor a que me falte alguno, en especial a los que han estado conmigo en esta última etapa, tanto en la UCLV como fuera de ella.

A todo el que le interese este trabajo.

RESUMEN

El presente trabajo aborda la temática de selección de rasgos. Es el estudio de este trabajo la validación estadística de dos algoritmos, uno en contexto local, ampliamente difundido en esta área del conocimiento, y el otro en contexto distribuido, tema novedoso en la rama de la selección de rasgos. El fundamento del contexto distribuido está basado en la cooperación entre subsistemas que comparten algún tipo de información acerca de los subconjuntos de datos sobre los que operan, para llegar a un mejor resultado, al algoritmo se le especifica el parámetros de nivel de intercambios de información entre los subconjuntos de datos para cooperar entre ellos.

Tanto en el caso del primer algoritmo como en el segundo, se efectúa un estudio estadístico para determinar los parámetros que brinden mejor resultado, además se establece una comparación entre estos dos algoritmos para arribar a conclusiones sobre su optimalidad dependiendo de los parámetros que se varíen, además de su aplicación o no a determinados problemas, dependiendo de las características de sus datos. En este proceso utilizamos las pruebas no paramétricas, por la naturaleza del problema, en especial el test de Friedman cuando se está trabajando con k muestras dependientes, con ayuda del test de Holm para las comparaciones entre muestras.

Se realiza una comparación entre los dos algoritmos, con los mejores parámetros resultantes de la validación estadística, aplicados a problemas de la vida real; específicamente al problema de predicciones meteorológicas y al de predicciones de infartos cardiacos.

Abstract

This work deals with feature selection problem. Two algorithms have been statistically proved. The first one solving the feature selection problem in traditional way, the second one is the same but running in a distributed environment, which is novelty in this field. The foundation of distributed environment is the collaboration in between across subsystems based on some kind of metadata interchange. The later algorithm is sensitive to the number of interchanges, and it that is studied as a parameter.

The role of its associated parameters for both algorithms is also studied. A comparison between original and distributed variant of the algorithm in terms of quality of results is done. Nonparametric test were used for statistical analysis. Friedman test was used for k related samples and also Holms test was useful helping to take decisions.

The algorithms are applied to two real world problems: Weather forecasting and Heart attack prediction. Using the best set of parameters the results are prominent.

Índice de Contenidos

RESUMEN i

Abstract ii

Índice de contenidos iii

Introducción 3

Capítulo I. De la minería de datos, el preprocesamiento de información y la validación de modelos. 6

1.1 Minería de Datos 6

1.1.1 Minería de Datos Distribuida 8

1.2 Reducción de dimensionalidad vertical. 9

1.2.1 Estrategias de búsqueda ó métodos de generación de subconjuntos. 13

1.3 Validación de algoritmos de selección de rasgos 14

1.3.1 Métodos estadísticos 14

Capítulo II. Algoritmos de selección de rasgos. 24

2.1 Herramientas 24

2.1.1 Optimización basada en Colonia de Hormigas 25

2.1.2 La Teoría de Conjuntos Aproximados 29

2.2 Solución al problema de selección de rasgos. Algoritmo ACO-RST-FSP. 33

2.3 Evaluación del algoritmo 37

2.3.1 Características de los conjuntos de datos. 37

2.3.2 Ajuste de parámetros 38

2.3.3 Comparación con otros métodos 46

2.4 Solución al problema de la selección de rasgos en contexto distribuido. El algoritmo D.ACO-RST-FSP 47

2.4.1 Ajuste de parámetros 48

2.5 Consideraciones parciales. 51

Capítulo III. Algoritmos ACO-RST-FSP y su variante distribuida D.ACO-RST-FSP aplicados a problemas reales 52

3.1 El problema del pronóstico climático 52

3.1.1 Ingeniería del conocimiento. Transformar la serie de tiempo en un conjunto de datos para ML. 53

3.1.2 Selección de rasgos aplicado a datos meteorológicos 58

3.2 Algoritmos de selección de rasgos aplicados al problema de cardiopatías. 59

3.3 Consideraciones parciales 62

CONCLUSIONES 64

RECOMENDACIONES 65

REFERENCIAS BIBLIOGRAFICAS 66

Introducción

En la actualidad existe tecnología capaz de adquirir y almacenar grandes volúmenes de datos. Teóricamente, mientras mayor sea la información a procesar, mayor será la precisión a la hora de emitir un criterio o resultado. En los algoritmos de aprendizaje los tiempos de ejecución, los atributos redundantes o irrelevantes y la degradación en el error de clasificación hacen esta teoría poco aceptada.

Para dar solución a los problemas planteados, y para que el clasificador pueda tener mejores resultados, se lleva a cabo el proceso de selección de rasgos, que consiste en escoger un subconjunto de atributos del conjunto original, los cuales sean relevantes y preserven la calidad de la información y lograr un buen rendimiento.

En resumen, el resultado de un proceso de selección de atributos sería:

Menos datos → los algoritmos pueden aprender más rápidos.

Mayor exactitud → el clasificador generaliza mejor.

Resultados más simples → más fácil de entender.

Menos atributos → evitar obtenerlos posteriormente.

Por las razones expuestas podemos concluir que la selección de rasgos es efectiva para eliminar atributos irrelevantes y redundantes, incrementando así la eficiencia en las tareas de minería de datos, mejorando el rendimiento y la comprensión de los resultados.

En una de las bases de datos creadas para guardar información, recopilada de lo ocurrido en años anteriores, encontramos los datos de meteorología de cuatro de las estaciones meteorológicas de la provincia, en esta se pueden encontrar los valores de distintos parámetros meteorológicos ocurridos a través de los años, en estas localidades, los cuales son de gran importancia para realizar pronósticos de lo que puede ocurrir con el clima en estos lugares, lo cual es de gran ayuda, pues el clima tiene una vital importancia en distintos sectores de la economía del país, fundamentalmente en la agricultura. Estos pronósticos pueden ser utilizados en la detección de determinadas anomalías como son las sequías y los temporales de muchas lluvias, de esta

manera los agricultores pueden tomar las medidas pertinentes con anticipación, para reducir las afecciones a fin de aumentar la seguridad de sus producciones.

Planteamiento del problema

El algoritmo ACO-RST-FSP tiene distintos parámetros, que se especifican para la corrida del mismo, estos pueden variar, influyendo en los resultados.

A su vez se tiene el algoritmo D.ACO-RST-FSP, el cual es semejante al primero solo que trabaja en un ambiente distribuido con colaboración entre los distintos subsistemas, lo cual indica que va a existir un intercambio de información entre estos, un número determinado de veces.

Se pretende realizar un estudio que permita conocer el comportamiento de estos dos algoritmos, para establecer el valor de los parámetros para los cuales estos llegan a resultados satisfactorios, además de definir cuál de los dos es el óptimo cuando se quiere llegar a una conclusión determinada.

Objetivo

Estudiar el comportamiento del algoritmo ACO-RST-FSP y su variante en contexto distribuido.

Objetivos específicos

1. Determinar la influencia de los parámetros de ACO en el algoritmo ACO-RST-FSP con relación a la calidad de los resultados.
2. Comparar los resultados del algoritmo ACO-RST-FSP con respecto a otros métodos de selección de rasgos.
3. Determinar la eficacia del algoritmo D.ACO-RST-FSP y establecer una regla para determinar el valor del parámetro que indica la cantidad de intercambios de meta-datos entre los subsistemas.

Preguntas de investigación

- ¿Cuáles son los valores de los parámetros del algoritmo ACO-RST-FSP mediante los cuales se pueden inferir resultados óptimos?

- ¿Es el algoritmo ACO-RST-FSP un buen algoritmo de selección de rasgos respecto a otros algoritmos del mismo tipo?

D.ACO-RST-FSP es un algoritmo similar al anterior, pero en un ambiente distribuido, pues realiza un intercambio de información entre los subsistemas que componen el conjunto de datos. Basado en esto se plantean las siguientes preguntas de investigación:

- ¿Cuántos intercambios de meta-información se deben realizar para lograr de D.ACO-RST-FSP su mejor comportamiento?

Capítulo I. De la minería de datos, el preprocesamiento de información y la validación de modelos.

Desde la década de los años 90, del pasado siglo, se vienen aplicando técnicas de minería de datos con diversos fines: apoyo a la toma de decisiones, gestión de procesos industriales, investigación científica, soporte al diseño de bases de datos y mejora de la calidad de los datos, entre otros.

Existe en la actualidad un conjunto de herramientas y técnicas que soportan la extracción de conocimiento útil a partir de los datos disponibles, y que se agrupan bajo el calificativo de "minería de datos".

1.1 Minería de Datos

Una definición tradicional es la siguiente: Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos.

El concepto de minería de datos también se solapa con los conceptos de aprendizaje automático y de estadística. En general, la estadística es la primera ciencia que históricamente extrae información de los datos, mediante metodologías procedentes de las matemáticas. Con el comienzo del uso de los ordenadores y como apoyo para esta tarea, surgió el concepto de *Machine Learning* (Aprendizaje Automatizado). Posteriormente, con el incremento del tamaño y con la estructuración de los datos es cuando se comienza a hablar de minería de datos.

En esencia la minería de datos (data mining), es un mecanismo de explotación y análisis, consistente en la búsqueda y extracción de información valiosa, patrones y reglas ocultos en grandes volúmenes de datos (Fayyad et al., 1996, Sangüesa and Molina, 2000)

La minería de datos se distingue de otros procesos de análisis de datos pues como resultado no obtiene datos sino conocimiento. Estos son conjuntos de reglas, ecuaciones, árboles de decisión, redes neuronales, grafos probabilísticos, los que pueden usarse para responder a interrogantes sobre los datos.

La minería de datos resulta muy útil en situaciones donde el volumen de datos es muy grande o complejo por la cantidad de variables que se manipulan. El proceso consta de varias fases:

1 **Filtrado de datos**: El formato de los datos contenidos en la fuente de datos (base de datos, Data Warehouse...) nunca es el idóneo, y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos "en bruto". Mediante el preprocesado, se filtran los datos (de forma que se eliminan valores incorrectos, no válidos, desconocidos... según las necesidades y el algoritmo a usar), se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso), o se reducen el número de valores posibles (mediante redondeo, clustering,...).

2 **Selección de Variables**: Aún después de haber sido preprocesados, en la mayoría de los casos se tiene una cantidad inmensa de datos. La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería. Los métodos para la selección de características son básicamente dos:

3 **Extracción de Conocimiento**: Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada una obliga a un pre-proceso diferente de los datos.

4 **Interpretación y Evaluación**: Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos resultados.

En muchísimos dominios de aplicación, los datos se encuentran distribuidos en varios nodos ubicados en sitios distantes. Los avances en la informática y las comunicaciones han favorecido el desarrollo de este tipo de aplicaciones. En estos casos, por lo general, no es posible o factible centralizar toda la información del sistema distribuido en un único repositorio, con el propósito

de realizar tareas de minería de datos, debido, por ejemplo, a restricciones económicas, técnicas o legales. Por lo que, es necesario aplicar técnicas de minería de datos sobre múltiples fuentes o almacenes de datos. La minería de datos sobre fuentes de datos distribuidas se denomina Minería de Datos Distribuida (Distributed Data Mining)[1].

1.1.1 Minería de Datos Distribuida

La DDM[2] tiene como objetivos: hacer sistemas escalables mediante el desarrollo de mecanismos que distribuyan las cargas de trabajo de manera flexible y el análisis realmente distribuido de datos inherentemente diseminados en diversas fuentes de datos, esto solventa el problema del ineficiente procesamiento centralizado y la seguridad para este tipo de datos (Arévalo, 2006).

DDM puede ser útil en ambientes con múltiples nodos de cómputo conectados por una red de alta velocidad. Incluso si los datos pueden ser rápidamente centralizados usando una red relativamente rápida, un apropiado balance de carga a través de los clúster de nodos puede requerir una solución distribuida. El asunto de la privacidad está creciendo en importancia en un rol protagónico en las aplicaciones emergentes de minería de datos.

Los sistemas DDM pueden ser capaces de aprender modelos desde fuentes de datos distribuidas sin intercambiar los datos en sí, esto permitiría ambas cosas, la prevención de fraudes y la preservación de la privacidad de los datos (KARGUPTA, 2003).

El primer paso en el desarrollo de una solución de minería de datos distribuida, es identificar como están distribuidos los datos. La mayoría de los algoritmos DDM son diseñados para el modelo relacional de datos, pero en este trabajo no se aborda el modelo relacional, los datos se representan en forma tabular.

En Minería de Datos Distribuida las fuentes de datos se pueden clasificar en homogéneas o heterogéneas. La mayoría de los trabajos consideran diseños homogéneos a través de los diferentes sitios, estos diseños contienen el mismo conjunto de atributos a través de los sitios de

---

[1] http://www.revistaesalud.com/index.php/revistaesalud/article/view/358/707 22/03/2010

[2] Minería de Datos Distribuida.

datos distribuidos. Este modelo de datos distribuidos usualmente ocurre en la misma organización o en dominios similares. Algunos algoritmos DDM consideran diseños heterogéneos de datos que definen diferentes conjuntos de atributos a través de un conjunto de datos distribuido; sin embargo el esquema heterogéneo es generalmente restringido a un simple escenario donde cada tabla participante comparte una llave común que vincula filas correspondientes a través de las tablas. Un paso importante es preparar los datos y en la minería de datos y DDM esto no es excepción. El pre-procesamiento de datos debe funcionar de forma distribuida. Muchas de las técnicas de pre-procesamiento de datos centralizados pueden ser directamente aplicadas sin descargar todos los conjuntos de datos hacia un solo sitio.

1.2 Reducción de dimensionalidad vertical.

La utilidad de la información que nos puede brindar una base de datos está dada en gran medida por la calidad de los mismos. Antes de realizar un análisis a los datos estos pasan por una fase de preparación en la cual se realizan una serie de tareas como son: recopilación, limpieza, transformación y reducción. (Ver **Figura 1**)

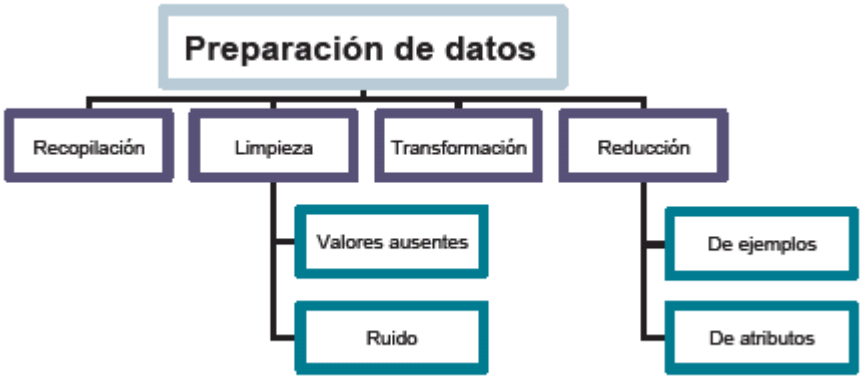

Figura 1. Proceso de preparación de los datos

Recopilación

Para poder comenzar a analizar y extraer algo útil en los datos es preciso, en primer lugar, disponer de ellos. Esto en algunos casos puede parecer trivial, partiendo de un simple archivo de datos, sin embargo en otros, es una tarea muy compleja donde se debe resolver problemas de

representación, de codificación e integración de diferentes fuentes para crear información homogénea.

Limpieza

En esta fase se resuelven conflictos entre datos, comprobando problemas de ruido, valores ausentes y valores fuera de rango.

Transformación

En ocasiones, la forma en que está representada la información originalmente no es la más adecuada para adquirir conocimiento a partir de ella. En esas situaciones se hace necesaria la aplicación de algún tipo de trasformación para adecuar los datos al posterior proceso de aprendizaje, como por ejemplo normalización o cambio de escala, discretización, generalización o extracción de atributos.

La Transformación de datos intenta transformar el formato de los datos a uno esperado por la aplicación. Las transformaciones afectan tanto el esquema de las tuplas como los dominios de sus valores. El proceso de transformación de los esquemas es generalmente resuelto junto con el proceso de limpieza de los datos. Los datos de las diferentes fuentes son transformados de acuerdo a un esquema común que satisfaga las necesidades de la aplicación. La corrección de los valores debe ser realizada solamente en los casos que los datos no estén en correspondencia con las exigencias del esquema común, conduciendo a fallas en el proceso de transformación. Esto hace que la limpieza de los datos y las transformaciones sean tareas complementarias.

Reducción

Los investigadores dedicados al Aprendizaje Automático Supervisado, y concretamente, al estudio de algoritmos que produzcan conocimiento en alguna de las representaciones usuales (listas de decisión, árboles de decisión, reglas de asociación, etc.) suelen realizar las pruebas con bases de datos estándares y accesibles a toda la comunidad científica (la gran mayoría de ellas de tamaño reducido), con objeto de verificar los resultados y validarlos con independencia. No obstante, y una vez asentadas estas propuestas, algunos de estos algoritmos sufren modificaciones orientadas a problemas específicos, los cuales, contienen una cantidad de información muy superior (decenas de atributos y decenas de miles de ejemplos) a la de los

conjuntos de datos de prueba. La aplicación de tales técnicas de minería de datos es por tanto una tarea que consume una enorme cantidad de tiempo y memoria, aún con la potencia de las computadoras actuales, que hace intratable la adaptación del algoritmo para solucionar el particular problema. Es conveniente, pues, aplicar técnicas de reducción de la información, estando orientadas fundamentalmente hacia dos objetivos: técnicas de reducción vertical o editado (reducción del número de ejemplos) y técnicas de reducción horizontal o selección de atributos (eliminar aquellos atributos que no sean relevantes para la información inherente a la base de datos). La **Figura 2** es un ejemplo donde se puede observar los dos tipos de reducción.

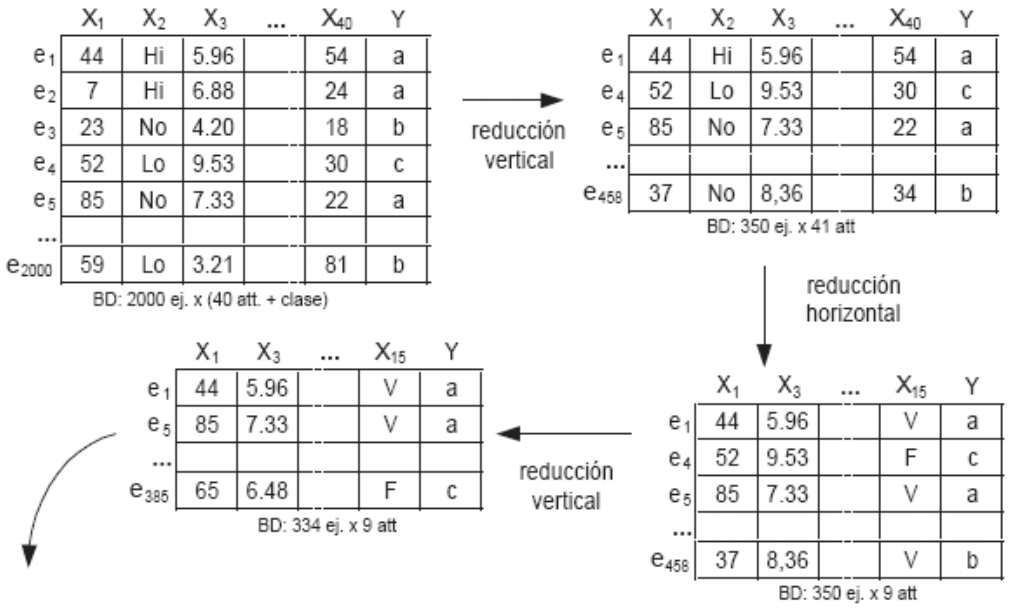

Figura 2. Tipos de reducción

Selección de rasgos

La selección de rasgos o atributos se ha convertido en el foco de muchas investigaciones en áreas de aplicación para las cuales conjuntos de datos con decenas de miles de variables están disponibles, apareciendo en diferentes tareas asociadas con el análisis de datos. Estas áreas incluyen con especial relevancia el Reconocimiento de Patrones, el Aprendizaje Automatizado (Liu and Yu, 2005, Ruiz et al., 2004), y la Bioinformática(Yvan Saeys et al., 2007). La selección de rasgos consiste en encontrar el subconjunto de atributos 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(Huan and Hiroshi, 2007). Esto es hecho eliminando rasgos irrelevantes y redundantes, proveyendo así una mejor representación de la información original, lo que reducirá significativamente el costo computacional y resultará una mejor generalización para el algoritmo de aprendizaje.

Dos tipos de enfoques pueden ser considerados para reducir la dimensionalidad: extracción de rasgos y selección de rasgos. El primero consiste en combinar las variables originales para construir nuevos rasgos, mientras que la última reduce la dimensión de la entrada mediante la eliminación de algunas variables irrelevantes o débilmente relevantes.

Definición 1 (**Selección de atributos**) Si A es el conjunto de todos los atributos en un conjunto de datos, hacer selección de atributos es escoger un subconjunto $S \in P(A)$. Donde $P(A)$ es el conjunto potencia de A, es decir, el conjunto formado por todos los subconjuntos de elementos de A y $|S| < |A|$.

Los procedimientos de selección de rasgos constan de dos componentes principales: la función de evaluación (una función heurística) y el método de generación de subconjuntos: el método de búsqueda. Según la naturaleza de la función de evaluación los algoritmos de selección de rasgos pueden ser divididos en tres categorías: filtros, envolventes (del inglés "wrapper") y empotrados. En la primera categoría se incluyen los 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 irrelevantes y redundantes. Por otro lado, en los métodos de tipo envolvente, la selección de atributos y el algoritmo de aprendizaje no son elementos independientes, ya que la selección hace uso del proceso de inducción para evaluar la calidad de cada conjunto de atributos seleccionados en cada momento. Los métodos empotrados al igual que los envolventes involucran al algoritmo de aprendizaje como parte del proceso de selección. Estos últimos realizan la selección durante el proceso de entrenamiento. Estos cuentan con su propio algoritmo de selección de atributos, como ocurre en los algoritmos que generan árboles de decisión, que utilizan sólo aquellos atributos necesarios para obtener una descripción consistente con el conjunto de aprendizaje.

Tanto filtros y envolventes hacen uso de estrategias de búsqueda para explorar el espacio de todas las posibles combinaciones de rasgos, que normalmente es demasiado grande para ser explorado exhaustivamente. Según estas estrategias los métodos de selección obtienen otra clasificación: completa, heurística y aleatoria. Dentro de las primeras se encuentran aquellas que tienen complejidad $O(n^2)$ pero aseguran la obtención del subconjunto óptimo bajo un criterio dado. Las heurísticas a diferencia de las completas recorren sólo una porción del espacio de búsqueda, y por tanto no aseguran la obtención del óptimo aunque el coste computacional es polinomial. Las estrategias aleatorias se basan en visitar diferentes regiones del espacio de búsqueda sin un orden predefinido.

Los métodos ávidos con selección hacia adelante o la eliminación dirigida hacia atrás son los más populares. En un método con selección hacia adelante se comienza con un conjunto vacío y progresivamente se agregan rasgos que mejoran la selección. En un procedimiento con eliminación dirigido hacia atrás se empieza con todos los rasgos y progresivamente elimina el menos útil. Ambos procedimientos son bastante rápidos y robustos. Sin embargo, como veremos, ellos pueden obtener subconjuntos diferentes y, dependiendo de la aplicación y los objetivos, puede preferirse un enfoque u otro.

1.2.1 Estrategias de búsqueda ó métodos de generación de subconjuntos.

Habiendo reconocido la necesidad de seleccionar rasgos en el contexto de otros rasgos y eliminar la redundancia, se dispone de una variedad amplia de algoritmos para escoger.

Existen diferentes enfoques y técnicas para seleccionar rasgos relevantes marcando las diferencias fundamentalmente por la función de evaluación o la estrategia de búsqueda utilizada. La diversidad de métodos incluye desde novedosas estrategias a clásicos como el enfoque lógico combinatorio, el estadístico, las de computación evolutiva, Búsqueda Tabú (TS), las basadas en la Teoría de los Conjuntos Aproximados; y entre los clásicos deben incluirse "best-first", "branch and bound", recocido simulado y algoritmos genéticos. Para una revisión sobre estas estrategias se recomienda ver (Kohavi and John, 1997), para una estudio en orden cronológico puede seguirse la propuesta de Martin Sewell, 2007 donde puede observarse además el sumario de métodos de selección de rasgos de Dash y Liu (DASH and Liu, 1997).

A continuación se mencionan brevemente algoritmos que siguen algunos de estos enfoques y técnicas.

Probablemente la técnica más ampliamente utilizada en extracción de rasgos es Análisis de Componentes Principales (Principal Component Analysis, PCA) (Duda et al., 2001) el cual construye nuevos rasgos no correlacionados llamados factores maximizando la varianza. La eficiencia de estos métodos ha sido demostrada en un rango amplio de dominios de aplicaciones, pero la interpretación de los nuevos rasgos no es obvia y requiere un esfuerzo importante del usuario.

Otra tendencia de métodos de búsqueda está basada en el uso de Inteligencia Colectiva (Swarm Intelligence), particularmente Optimización con Enjambres de Partículas (Particle Swarm Optimization, PSO) y las técnicas de Optimización con Colonias de Hormigas (Ant Colony Optimization, ACO) (Gómez et al., 2009, Wang, 2007, Al-Ani, 2005, Bello and Nowé, 2005, Firpi and Goodman, 2004, Jensen and Shen, 2005).

1.3 Validación de algoritmos de selección de rasgos

1.3.1 Métodos estadísticos

Una investigación bien planificada debe incluir en su diseño referencias precisas acerca de las técnicas estadísticas que se utilizan en el análisis de los datos.

El análisis estadístico es el procedimiento por medio del cual se puede aceptar o rechazar un conjunto de datos como confirmatorios de una hipótesis, conocido el riesgo que se corre -en función de la probabilidad- al tomar tal decisión. En las últimas décadas, el desarrollo de las pruebas estadísticas se ha incrementado a tal grado que en la actualidad se cuenta con varias pruebas alternativas, las cuales se pueden usar para casi todo diseño experimental, de modo que el investigador se encuentra ante el dilema de seleccionar la más apropiada y económica, para las preguntas que, mediante la investigación, desea contestar.

Ante esa situación, es necesario tener una base racional, por medio de la cual se seleccione la prueba más apropiada. Esta selección constituye el punto crítico del análisis estadístico. En la elección de una prueba estadística, se deben aplicar criterios como el tipo de escala, hipótesis,

potencia y eficacia de la prueba, características muéstrales y las tendencias rectilíneas o curvilíneas del fenómeno.

De las mediciones que en el terreno de la investigación se hayan realizado, puede inferirse el tipo de escala, de modo que éste es el primer paso para elegir un procedimiento estadístico: *la prueba paramétrica* y *la no paramétrica.*

Tipo de escala

En las observaciones de una investigación se puede dar una medición que en este campo consiste en asignar números a objetos y eventos de acuerdo con reglas de la lógica aceptables.

El sistema numérico es una creación altamente lógica, que ofrece múltiples posibilidades, para manifestaciones también de carácter lógico. Si se puede, de manera legítima, asignar números al describir características, objetos y eventos, será factible operar con ellos en todos sus modos permisibles y, de esas operaciones, derivar conclusiones aplicables a los fenómenos observados y medidos. Entonces, se justifica describir cosas reales por medio de números, siempre y cuando exista un grado de isomorfismo (semejanza de propiedades) entre las cosas reales y el sistema numérico, es decir, ciertas propiedades de los números deben tener paralelismo con los fenómenos observados, para que confiadamente se pueda asignar los números. Implícito en cada caso hay un conjunto de reglas para asignar números o valores: son estas reglas las que dan significado a las cantidades. Los objetos pueden ser perceptuales o conceptuales.

La escala de medida de una característica tiene consecuencias en la manera de presentación de la información y el resumen. La escala de medición -grado de precisión de la medida de la característica- también determina los métodos estadísticos que se usan para analizar los datos. Por lo tanto, es importante definir las características por medir. Las escalas de medición más frecuentes son las siguientes:

Escala Nominal.- No poseen propiedades cuantitativas y sirven únicamente para identificar las clases. Los datos empleados con las escalas nominales constan generalmente de la frecuencia de los valores o de la tabulación de número de casos en cada clase, según la variable que se está estudiando. El nivel nominal permite mencionar similitudes y diferencias entre los casos particulares. Los datos evaluados en una escala nominal se llaman también "observaciones

cualitativas", debido a que describen la calidad de una persona o cosa estudiada, u "observaciones categóricas" porque los valores se agrupan en categorías. Por lo regular, los datos nominales o cualitativos se describen en términos de porcentaje o proporciones. Para exhibir este tipo de información se usan con mayor frecuencia tablas de contingencia y gráficas de barras o de pastel.

Escala Ordinal.- Las clases en las escalas ordinales no solo se diferencian unas de otras (característica que define a las escalas nominales) sino que mantiene una especie de relación entre sí. También permite asignar un lugar específico a cada objeto de un mismo conjunto, de acuerdo con la intensidad, fuerza, etc.; presentes en el momento de la medición. Una característica importante de la escala ordinal es el hecho de que, aunque hay orden entre las categorías, la diferencia entre dos categorías adyacentes no es la misma en toda la extensión de la escala. Algunas escalas consisten en calificaciones de múltiples factores que se agregan después para llegar a un índice general. Debe mencionarse brevemente una clase espacial de escala ordinal llamada "escala de posición", donde las observaciones se clasifican de mayor a menor (o viceversa). Al igual que en las escalas nominales, se emplean a menudo porcentajes y proporciones en escalas ordinales.

Escala de Intervalo.- Refleja distancias equivalentes entre los objetos y en la propia escala. Es decir, el uso de ésta escala permite indicar exactamente la separación entre 2 puntos, lo cual, de acuerdo al principio de isomorfismos, se traduce en la certeza de que los objetos así medidos están igualmente separados a la distancia o magnitud expresada en la escala.

Los tipos de escala (Nominal, Ordinal e Intervalo) son fundamentales para el buen desempeño de las pruebas estadísticas, sirviendo ellos, como criterio, para optar por uno de los procedimientos estadísticos, pruebas paramétricas o pruebas no paramétricas.

Pruebas Paramétricas vs. No Paramétricas

Las pruebas estadísticas se dividen en dos grandes grupos: paramétricas y no paramétricas. Las primeras son aquellas cuyo modelo especifica ciertas condiciones o premisas que debe tener la población, de la cual se ha derivado la muestra bajo análisis; además se requiere expresar las observaciones en escala de intervalo o tasa. Por otra parte, las pruebas no paramétricas, como su nombre lo indica, no requieren satisfacer esas condiciones o premisas.

Las pruebas paramétricas son las más eficaces y de uso común en la investigación, estas deben cumplir las premisas siguientes:

1 Las observaciones deben ser independientes. Al seleccionar un caso, para incluirlo en la muestra, no se deben prejuiciar las probabilidades de selección de ningún otro caso de la población, asimismo, la puntuación que se dé a una observación no debe prejuiciar a ninguna otra.

2 Las poblaciones deben provenir de universos cuya distribución siga una curva normal.

3 Las variables consideradas en el estudio deben ser medidas por lo menos en escala de intervalo, para que sea posible hacer operaciones aritméticas.

Se puede concluir que: las poblaciones que no cumplen con los requisitos anteriores, son estudiadas con el grupo de pruebas no paramétricas.

Pruebas paramétricas

Se llaman así porque su cálculo implica una estimación de los parámetros de la población con base en muestras estadísticas. Mientras más grande sea la muestra más exacta será la estimación, mientras más pequeña, más distorsionada será la media de las muestras por los valores raros extremos.

Suposiciones que subyacen a la utilización de las pruebas paramétricas.

1 El nivel de medición debe ser al menos de intervalo. Debemos tomar una decisión a cerca de nuestra variable dependiente. ¿Es realmente un nivel de intervalo? Si es una escala no estandarizada, o si se basa en estimaciones o calificaciones con humanos. Frecuentemente aparecen como intervalo pero lo reducimos a nivel ordinal al darles rango.

2 Los datos de la muestra se obtienen de una población normalmente distribuida. Este principio suele mal entenderse como: la muestra debe distribuirse normalmente, "no es así". La mayoría de las muestras son demasiado pequeñas para siquiera parecerse a una distribución normal, la cual solo obtiene su característica en forma de campana con la acumulación de muchas puntuaciones.

Las pruebas paramétricas poseen ventajas como son: más poder de eficiencia, mayor sensibilidad a los rasgos de los datos recolectados, menos posibilidades de errores, además de dar estimaciones probabilísticas bastante exactas (robustas). Estas pruebas también presentan desventajas, siendo las mismas, más complicadas de calcular, además de poseer limitaciones con respecto a los tipos de datos que puede evaluar.

Entre las pruebas paramétricas existentes se pueden citar las siguientes:

1 Prueba del valor Z de la distribución normal

2 Prueba T de Student para datos relacionados (muestras dependientes)

3 Prueba T de Student para datos no relacionados (muestras independientes)

4 Prueba de ji cuadrada de Bartlett para demostrar la homogeneidad de varianzas

5 Prueba F (análisis de varianza o ANOVA)

Pruebas no paramétricas

Las pruebas no paramétricas nos permiten analizar datos en escala nominal u ordinal a pesar de que no se conozcan los parámetros de una población, utilizada para hacer un contraste de hipótesis. Estas pruebas se utilizan fundamentalmente cuando se presentan una de las situaciones siguientes:

- Cuando los datos puntualizan a las escalas nominal u ordinal.
- Se utiliza solo la frecuencia.
- Poblaciones pequeñas.
- Cuando se desconocen los parámetros media, moda, etc.
- Cuando se quiere contrastar o comparar hipótesis.
- Cuando se requiere de establecer el nivel de confianza o significatividad en las diferencias.
- Cuando la muestra es seleccionada no probabilísticamente.

Las pruebas no paramétricas se aplican en correspondencia a las características de las muestras que se quieran procesar, los datos que pertenecen a la escala nominal se pueden procesar con los siguientes test:

1 Leyes de la probabilidad y prueba binomial

2 Prueba ji^2 de Pearson para una muestra

3 Prueba ji_2 de Pearson para dos y más muestras independientes

4 Prueba de bondad del ajuste mediante ji^2

5 Prueba ji^2 de proporciones para tres o más muestras independientes

6 Prueba de probabilidad exacta de Fischer y Yates

7 Prueba de McNemar para muestras dependientes

8 Prueba Q de Cochran para tres o más muestras dependientes

9 Análisis secuencial

En otro caso, los que pertenecen a la escala ordinal y de intervalos se pueden procesar con los siguientes test:

1 Prueba de Kolmogorov-Smirnov para una muestra

2 Prueba de U Mann-Whitney para dos muestras independientes

3 Prueba de Wilcoxon de rangos señalados y pares igualados para dos muestras dependientes

4 Análisis de varianza de una entrada de Kruskal-Wallis para más de dos muestras independientes

5 Análisis de varianza de doble entrada por rangos de Friedman para más de dos muestras dependientes

En el caso particular de esta investigación se proceden a aplicar los test no paramétricos para muestras en escala ordinal y de intervalos, pues las poblaciones a procesar son pequeñas, dependientes y con ellas se pretende contrastar hipótesis, para inferir conclusiones sobre el comportamiento de los datos. Para robustecer el criterio anterior las poblaciones fueron

sometidas a la prueba Sapiro-Wilk (se escoge este test por contar, para el experimento con poblaciones pequeñas –menores de 100–, para poblaciones con más de 100 objetos se recomienda el Test de Kolmogorov-Smirnov) para conocer el comportamiento de la distribución en las poblaciones, concluyendo de esta manera que las poblaciones no presentan una distribución normal. Para efectuar el estudio entre las poblaciones se escoge en particular el *Análisis varianza de doble entrada por rangos de Friedman para más de dos muestras dependiente,* pues es el indicado para trabajar con los datos que siguen las características descritas anteriormente, este test se emplea para determinar la significación entre las muestras, pero cuando se obtienen diferencias significativas entre ellas, hace falta saber entre cuales de ellas se plantea la diferencia. Para lograr ese propósito se aplica el *Test de Nemenyi* para comparaciones múltiples, el que a su vez utiliza el *Test de Holm* para comparar con un algoritmo de control. A continuación se hace una breve descripción de los test estadísticos utilizados para la validación.

Análisis de varianza de doble entrada por rangos de Friedman para más de dos muestras dependientes

La prueba de Friedman sirve para comparar *J* promedios poblacionales cuando se trabaja con muestras relacionadas. La situación experimental que permite resolver esta prueba es similar a la del ANOVA de un factor con medidas repetidas: a *n* sujetos (o a *n* bloques, cada uno de tamaño *J*) se le aplican *J* tratamientos o se le toman *J* medidas con intención de averiguar si los promedios de esos *J* tratamientos o medidas son o no iguales.

Las ventajas de esta prueba frente al estadístico *F* del ANOVA (Este test es del conjunto de las pruebas paramétricas) es palpable pues no es necesario establecer los supuestos tan exigentes del ANOVA (normalidad, igualdad de varianzas) y permite trabajar con datos ordinales y de intervalo. La prueba de Friedman, por tanto, constituye una alternativa al estadístico *F* cuando no se cumplen los supuestos paramétricos del ANOVA.

El diseño está formado por *J* muestras o tratamientos relacionados y por una muestra aleatoria de *n* sujetos o bloques independientes entre sí e independientes de los tratamientos. Las puntuaciones originales deben ser transformadas en rangos R_{ij}. Esos rangos se asignan

independientemente para cada sujeto o bloque; es decir, se asignan rangos de 1 a *J* a las observaciones del sujeto o bloque1; lo mismo con el resto de los bloques por separado.

Los rangos asignados a cada sujeto o bloque suman, en todos los casos, *J (J+1)/2* (pues en cada sujeto o bloque estamos asignando rangos desde 1 a *J*). Llamaremos R_{ij} al rango asignado al sujeto o bloque *i* en el tratamiento o muestra *j* y R_j a la suma de los rangos asignados a las *n* observaciones de la muestra *j*:

$$R_j = \sum_{i}^{n} R_{ij} \quad \Rightarrow \quad \overline{R_j} = \frac{R_j}{n} \qquad \textbf{(1. 1)}$$

Obviamente, si los promedios poblacionales son iguales, los R_j serán parecidos. Tomando como punto de partida estas sumas de rangos, Friedman ha diseñado un estadístico con distribución muestral conocida capaz de proporcionar información sobre el parecido existente entre las *J* poblacionales

$$X_r^2 = \frac{12}{nJ(J+1)} \sum_j R_j^2 - 3n(J+1) \qquad \textbf{(1. 2)}$$

El estadístico de Friedman plantea la hipótesis de igualdad de promedios poblacionales. En el caso que el nivel crítico sea menor que 0,05 se rechaza la hipótesis, determinando, que la calidad de la variable que se está comparando no es la misma en los distintos niveles considerados.

Test de Iman-Davenport

Test de Iman and Davenport(Lunacek et al., 2005): Se trata de una medida derivada de la de Friedman a causa del efecto conservador indeseado que produce éste. El estadístico se muestra en la expresión **(1. 3)**

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} \qquad \textbf{(1. 3)}$$

y se distribuye acorde a una distribución F con k – 1 y (k - 1)(N - 1) grados de libertad.

Test de Nemenyi

El test de Nemenyi es un test de comparación múltiple (versión no paramétrica del método Tukey). El estadístico usado se muestra en la expresión **(1. 4)**. Este sigue la distribución Q "Studentized range static".

$$\frac{R_B^- - R_A^-}{\sqrt{\frac{k(k+1)}{12n}}} = q_{\alpha,\infty,k} \qquad \textbf{(1. 4)}$$

Con este test podemos comparar cada uno de los algoritmos del grupo con el resto. Para ello se calcula la distancia crítica (DC) según la expresión **(1. 5)** y se consideran significativas las diferencias entre los métodos cuyos rangos medios difieran en una calidad mayor a DC.

$$DC = q_{\alpha,\infty,k} * \sqrt{\frac{k(k+1)}{12n}} \qquad \textbf{(1. 5)}$$

Test de Holm

Test de Holm(Holm, 1979): Para contrastar el procedimiento de Bonferroni-Dunn, se dispone de un test que prueba secuencialmente las hipótesis ordenadas según su significancia. Denominaremos a los valores de p ordenados por $p1, p2,...,$ de tal forma que $p1 \leq p2 \leq ... \leq pk-1$. El método de Holm compara cada pi con $\alpha/(k-i)$ comenzando desde el valor de p más significante. Si $p1$ es menor que $\alpha/(k-1)$, la correspondiente hipótesis se rechaza y nos permite comparar $p2$ con $\alpha/(k-2)$. Si la segunda hipótesis se rechaza, continuamos el proceso. En cuanto una determinada hipótesis no puede ser rechazada, todas las restantes se mantienen como aceptadas. El estadístico para comparar el algoritmo i-ésimo con el j-ésimo se muestra en la expresión **(1. 6)**.

$$Z = (R_i - R_j) \Big/ \sqrt{\frac{k(k+1)}{6N}} \qquad \textbf{(1. 6)}$$

El valor z se utiliza para encontrar la probabilidad correspondiente a partir de la tabla de la distribución normal, la cual es comparada con el correspondiente valor de α. El test de Holm es más potente que Bonferroni-Dunn y no hace ninguna suposición adicional sobre las hipótesis chequeadas.

1.4 Consideraciones parciales

1. La calidad de un método de selección de rasgo puede ser medida por varios indicadores como la calidad de la clasificación en los envolventes, u otras como la longitud del subconjunto más corto encontrado, e incluso la cantidad de subconjuntos encontrados cuando el método proporciona una variedad de soluciones.
2. En la validación de un nuevo método de selección de rasgos debe tomarse en cuenta el resultado al aplicarlo a conjuntos de datos para pruebas publicadas internacionalmente. Y cuando el método sea sensible a varios parámetros debe hacerse una sugerencia de sus valores, posiblemente tomando en cuenta las características de los datos sobre los cuales se aplica.

Capítulo II. Algoritmos de selección de rasgos.

En la temática de selección de rasgos se puede afirmar casi categóricamente que los métodos existentes actualmente no han resuelto totalmente la selección de atributos con un costo computacional adecuado y el problema en contexto distribuido ha sido muy poco estudiado. La importancia que presta la comunidad científica es apreciada por el número de investigadores que publica en los talleres, congresos, y números especiales de revistas dedicados al tema como son: los dos números especiales del "*Journal of Machine Learning Research*" en 2003 y 2007, el libro "*Computational Methods of Feature Selection*" de Huan Liu e Hiroshi Motoda en 2008, los talleres "*The NIPS 2003 Workshop on feature selection*", "*Workshop on Feature Selection for Data Mining*" en 2005 y 2006, "*Workshop on New challenges for feature selection in data mining and knowledge discovery*" 2008 en el marco de la "*European Conference on Machine Learning*" 2008, "*Workshop on feature selection in Bioinfromatics*" en 2009, y "*Workshop on Feature Selection in Data Mining*" en 2010. Esto motivó la obtención de los algoritmos presentados en este capítulo y en el trabajo descrito más que presentarlos como una novedad se realiza un estudio de algunos de sus parámetros más importantes.

2.1 Herramientas

Como se mencionó en el capítulo 1, todo método de selección de rasgos debe contar al menos con dos componentes claves: el método de búsqueda y la función de evaluación. La variedad de técnicas de selección de rasgos está dada precisamente por la diversidad de algoritmos utilizados como métodos de búsqueda en la generación de los subconjuntos candidatos o la exploración del espacio de búsqueda como otra interpretación y las disímiles variantes de evaluación de estos subconjuntos. Los algoritmos estudiados en esta tesis utilizan la combinación de la Optimización basada en Colonias de Hormigas *(Ant Colony Optimization)*, como método de búsqueda, y una medida de la Teoría de Conjuntos Aproximados *(Rough Sets Theory)*, como función de evaluación. En tópicos sucesores se abordarán los temas relacionados con estas herramientas de manera precisa.

2.1.1 Optimización basada en Colonia de Hormigas

La familia de algoritmos ACO(Dorigo and Stutzle, 2004b, Dorigo and Caro, 1999, Dorigo et al., 1999, Dorigo and Stutzle, 2003) se basan en una colonia de hormigas artificiales, son modelos inspirados en el comportamiento de las colonias de hormigas reales. Estudios realizados explican como las hormigas son capaces de seguir la ruta más corta en su camino de ida y vuelta entre la colonia y una fuente de alimento. Esto es debido a que las hormigas pueden "transmitirse información" entre ellas gracias a que cada una de ellas, al desplazarse, va dejando un rastro de una sustancia llamada feromona a lo largo del camino seguido. Así, mientras una hormiga aislada se mueve de forma esencialmente aleatoria, los "agentes" de una colonia de hormigas detectan el rastro de feromona dejado por otras hormigas y tienden a seguir dicho rastro. Éstas a su vez van dejando su propia feromona a lo largo del camino recorrido y por tanto lo hacen más atractivo, puesto que se ha reforzado el rastro de feromona. Sin embargo, la feromona también se va evaporando con el paso del tiempo provocando que el rastro de feromona sufra, por otro lado, cierto debilitamiento. En definitiva, puede decirse que el proceso se caracteriza por una retroalimentación positiva, en la que la probabilidad con la que una hormiga escoge un camino aumenta con el número de hormigas que previamente hayan elegido el mismo camino.

Los algoritmos ACO son procesos iterativos. En cada iteración se "lanza" una colonia de *m* hormigas y cada una de las hormigas de la colonia 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.

Modo de funcionamiento general

El modo de operación básico de un algoritmo de ACO(Dorigo and Stutzle, 2004b, Dorigo and Stutzle, 2003, Dorigo et al., 2006) es como sigue: las *m* hormigas (artificiales) de la colonia se mueven, concurrentemente y de manera asíncrona, a través de los estados adyacentes del problema (que puede representarse en forma de grafo con ponderaciones o sin ellas). Este movimiento se realiza siguiendo una regla de transición que está basada en la información local disponible en las componentes (nodos). Esta información local incluye la información heurística y memorística (rastros de feromona) para guiar la búsqueda. Las hormigas construyen incrementalmente soluciones al moverse por el grafo de construcción. Opcionalmente, las

hormigas pueden depositar feromona cada vez que crucen un arco (conexión) mientras que construyen la solución (*actualización en línea paso a paso de los rastros de feromona*). Una vez que cada hormiga ha generado una solución, ésta se evalúa y el agente puede depositar una cantidad de feromona en dependencia de la calidad de su solución (*actualización en línea de los rastros de feromona*). Esta información guiará la búsqueda de las otras hormigas de la colonia en el futuro. Además, el modo de operación genérico de un algoritmo de ACO incluye dos procedimientos adicionales, la evaporación de los rastros de feromona y las acciones del demonio. La evaporación de feromona la lleva a cabo el entorno y se usa como un mecanismo que evita el estancamiento en la búsqueda y permite que las hormigas busquen y exploren nuevas regiones del espacio. Las acciones del demonio constituyen una funcionalidad opcional (que no tiene un contrapunto natural) para implementar tareas desde una perspectiva global que no pueden llevar a cabo las hormigas por la perspectiva local que ofrecen. Ejemplos son: observar la calidad de todas las soluciones generadas y depositar una nueva cantidad de feromona adicional sólo en las componentes asociadas a algunas soluciones, o aplicar un procedimiento de búsqueda local a las soluciones generadas por las hormigas antes de actualizar los rastros de feromona. En ambos casos el demonio reemplaza la actualización en línea a posteriori de feromona y el proceso pasa a llamarse *actualización fuera de línea de rastros de feromona*.

En ACO el significado de los rastros de feromona y la función heurística o de visibilidad dependen totalmente del problema a resolver. Los rastros de feromona, cuando se está en presencia del problema del Viajante de Comercio(Dorigo and Gambardella, 1997a) asociados a los arcos del grafo, con el objetivo de premiar las buenas secuencias. Por otra parte, en problemas de asignación (Selección de Rasgos(Bello and Nowé, 2005)) donde los cambios de posición entre componentes de una solución no influyen en la calidad de la misma, los rastros de feromona son asociados a los nodos del grafo.

Dentro de los algoritmos de ACO las diferencias fundamentales radican en la regla de transición que utilizan para la construcción de las soluciones y en el tratamiento que le dan a los rastros de feromona. Debido a esto, aparecen en la literatura distintos algoritmos ACO.

Principales algoritmos basados en colonia de hormigas

Entre los algoritmos de ACO disponibles para problemas de optimización combinatoria(Dorigo and Blum, 2005) se encuentran: el Sistema de Hormigas (Ant System, AS)(Dorigo et al., 1996), el Sistema de Colonia de Hormigas (Ant Colony System, ACS) (Dorigo and Gambardella, 1997b), el Sistema de Hormigas Máximo-Mínimo (Max-Min Ant System, MMAS) (Stützle and Hoos, 2000), el Sistema de Hormigas con Ordenación Jerárquica (Rank-Based Ant System) (Bullnheimer et al., 1999), entre otros. Esta técnica comienza a tener la madurez tecnológica adecuada para su utilización en problemas reales, como puede apreciarse en la publicación del libro (Dorigo and Blum, 2005).

A continuación se presenta una pequeña descripción del Sistema de Colonia de Hormigas debido a que fue seleccionado para llevar a cabo este trabajo.

El **Sistema de Hormigas** (Dorigo et al., 1996, Heinonen and Pettersson, 2007), desarrollado por Dorigo en su tesis doctoral en 1992 (Dorigo and Stutzle, 2004a), fue el primer algoritmo de ACO. Su versión actual (Ant Cycle) apareció conjuntamente con otras variantes de este, como el Sistema de Hormigas Densidad (Ant Density) y Sistema de Hormigas Cantidad (Ant Quantity). El AS se caracteriza por el hecho de que, la actualización de feromona se realiza una vez que todas las hormigas han completado sus soluciones, y se lleva a cabo como sigue: primero, todos los rastros de feromona se reducen en un factor constante, implementándose de esta manera la evaporación de feromona según la ecuación **(2. 1)**,

$$t_{ij} \leftarrow (1-\rho)\cdot t_{ij},\ \rho \in (0,1] \qquad \textbf{(2. 1)}$$

donde ρ se conoce como constante de evaporación y es la encargada de reducir los rastros de feromona para evitar el estancamiento de las soluciones y t_{ij} la cantidad de feromona asociada al arco a_{ij}.

A continuación, cada hormiga de la colonia deposita una cantidad de feromona en función de la calidad de su solución, según la ecuación **(2. 2)**,

$$t_{ij} \leftarrow t_{ij} + \Delta t^k \quad \forall a_{ij} \in S^k \qquad \textbf{(2. 2)}$$

donde $\Delta t^k = f(C(S^k))$, representa la cantidad de feromona a depositar por la hormiga *k* en cada arco a_{ij} de su solución encontrada (S^k). Este valor depende de la calidad de la solución $C(S^k)$.

Las soluciones en el AS se construyen como sigue. En cada paso de construcción, una hormiga *k* escoge ir al siguiente nodo con una probabilidad que se calcula como:

$$p_{ij}^{k} = \frac{(\tau_{ij})^{\alpha} \cdot (\eta_{ij})^{\beta}}{\sum_{j \in N_i^k} (\tau_{ij})^{\alpha} \cdot (\eta_{ij})^{\beta}} \qquad si \quad j \in N_i^k \qquad \textbf{(2. 3)}$$

donde N_i^k es el vecindario alcanzable por la hormiga *k* cuando se encuentra en el nodo *i*. Los parámetros α y beta β controlan el proceso de búsqueda. Para $\alpha = 0$ se tiene una búsqueda heurística estocástica clásica, mientras que para $\beta = 0$ sólo el valor de la feromona tiene efecto. Un valor de $\alpha < 1$ lleva a una rápida situación de convergencia (stagnation)(Dorigo et al., 1999). El vector P_{ij}^k contiene las probabilidades de movimiento calculadas para los nodos de la vecindad (N_i^k) de la hormiga *k*. El valor τ_{ij} representa el elemento (*i, j*) en la matriz de feromona y η_{ij} se denomina función de visibilidad o función heurística y mide la calidad de un nodo *j* a partir del *i*.

El **Sistema de Colonia de Hormigas** (Dorigo and Gambardella, 1997b, Liu et al., 2008) es uno de los primeros sucesores del AS que introduce tres modificaciones importantes con respecto a dicho algoritmo de ACO:

1. Utiliza una regla de transición distinta y más agresiva, denominada regla proporcional pseudo-aleatoria. Sea *k* una hormiga situada en el nodo *r*, $q_0 \in [0,1]$ un parámetro y *q* un valor aleatorio en el mismo intervalo, el siguiente nodo *j* se elige como:

$$j = \max_{j \in N_i^k} \left\{ \tau_{ij} \cdot \eta_{ij}^{\beta} \right\} \quad \text{si} \quad q \le q_0 \qquad \textbf{(2. 4)}$$

En caso contrario se utiliza la regla probabilística del AS (ecuación

$$p_{ij}^{k} = \frac{(\tau_{ij})^{\alpha} \cdot (\eta_{ij})^{\beta}}{\sum_{j \in N_i^k} (\tau_{ij})^{\alpha} \cdot (\eta_{ij})^{\beta}} \qquad si \quad j \in N_i^k \qquad \textbf{(2. 3)(2. 5)}$$

Como puede observarse, la regla tiene una doble intención: cuando $q \le q_0$, utiliza en gran medida el conocimiento disponible (explotar), eligiendo la mejor opción con respecto a la información

heurística y los rastros de feromona. Sin embargo, si $q > q_0$ se aplica una exploración controlada, tal como se hacía en el AS.

2. Las hormigas aplican una actualización en *línea paso a paso* de los rastros de feromona que favorece la generación de soluciones distintas a las encontradas.

   Cada vez que una hormiga viaja por una arista a_{ij}, aplica la regla:

$$t_{ij} \leftarrow (1-\varphi)\cdot t_{ij} + t(0) \qquad \textbf{(2. 6)}$$

donde $\varphi \in (0,1]$ es un segundo parámetro de decremento de feromona. Como puede verse, la regla de actualización en *línea paso a paso* incluye tanto la evaporación de feromona como la actualización de la misma.

3. Se realiza una actualización *fuera de línea* de los rastros de feromona (acción del demonio), donde el ACS sólo considera una hormiga concreta, la que generó la mejor solución global, $S_{mejor\text{-}global.}$

La actualización de la feromona se lleva a cabo evaporando primero estos rastros en todas las conexiones utilizadas por la mejor hormiga global (es importante recalcar que, en el ACS, la evaporación de la feromona sólo se aplica a las conexiones de la solución, que es también la usada para depositar feromona) tal como sigue:

$$t_{ij} \leftarrow (1-\rho)\cdot t_{ij} \quad \forall a_{ij} \in S_{mejor-global} \qquad \textbf{(2. 7)}$$

A continuación se deposita feromona a los arcos que pertenecen a la mejor solución encontrada hasta el momento usando la regla:

$$t_{ij} \leftarrow t_{ij} + \Delta t \quad \forall a_{ij} \in S_{mejor-global} \qquad \textbf{(2. 8)}$$

donde $\Delta t = f(C(S_{mejor-global}))$, es decir la cantidad de feromona está en dependencia de la calidad de la mejor solución encontrada hasta el momento $C(S_{mejor-global})$.

2.1.2 La Teoría de Conjuntos Aproximados

La Teoría de Conjuntos Aproximados (Rough Sets Theory, RST) fue introducida por Z. Pawlak en 1982(Pawlak, 1982). Se basa en aproximar cualquier concepto, un subconjunto duro del

dominio como por ejemplo, una clase en un problema de clasificación supervisada, por 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, generar las relaciones entre ellos (en forma de reglas), entre otras (Choubey, 1996, Greco and Inuiguchi, 2003, Tsumoto, 2003, Sugihara and Tanaka, 2006).

En este epígrafe se describirán los conceptos fundamentales de los Conjuntos Aproximados, tanto para el caso clásico como para el enfoque basado en relaciones de similitud.

Principales definiciones de la Teoría de los Conjuntos Aproximados

La filosofía de los conjuntos aproximados se basa en la suposición de que con todo objeto x de un universo U está asociada una cierta cantidad de información (datos y conocimiento), expresado por medio de algunos atributos que describen el objeto (Komorowski and Pawlak, 1999, Bazan et al., 2003). En la Teoría de los Conjuntos Aproximados la estructura de información básica es el Sistema de Información.

Definición 2. Sistema de Información y sistema de decisión

Sea un conjunto de atributos $A=\{a_1, a_2,...,a_n\}$ y un conjunto U no vacío llamado universo de ejemplos (objetos, entidades, situaciones o estados) descritos usando los atributos a_i; al par (U,A) se le denomina Sistema de información (Komorowski and Pawlak, 1999)[3]. Si a cada elemento de U se le agrega un nuevo atributo d llamado decisión, indicando la decisión tomada en ese estado o situación, entonces se obtiene un Sistema de decisión $(U, A \cup \{d\})$, donde $d \notin A$.

Un aspecto importante en la Teoría de los Conjuntos Aproximados es la reducción de atributos basada en el concepto de reductos. Un reducto es un conjunto reducido de atributos que preserva la partición del universo (Komorowski and Pawlak, 1999, Zhong et al., 2001). El uso de reductos en la selección y reducción de atributos ha sido ampliamente estudiado (Kohavi and Frasca,

[3] Esta definición es independiente a la definición de Sistema de Información de Shannon

1994, Carlin, 1998, Komorowski and Pawlak, 1999, Pal and Skowron, 1999, Ahn, 2000, Lazo et al., 2001, Zhong et al., 2001, Santiesteban and Pons, 2003, Caballero, 2005).

Medidas de inferencia clásicas de la Teoría de los Conjuntos Aproximados

La Teoría de los Conjuntos Aproximados ofrece algunas medidas para analizar los sistemas de información (Skowron, 1999, Arco et al., 2006). A continuación se muestra la medida utilizada en los algoritmos presentados.

Calidad de la clasificación.

Este coeficiente describe la inexactitud de las clasificaciones aproximadas:

$$\gamma(DS) = \frac{\sum_{i=1}^{l} \left| R'_*(X_i) \right|}{\left| U \right|} \qquad \textbf{(2. 9)}$$

Las expresiones *R'*-inferior ($R'_*(X)$) y *R'*-superior ($R'^*(X)$) de *X*, están definidas en los trabajos previamente citados. La medida calidad de la clasificación expresa la proporción de objetos que pueden estar correctamente clasificados en el sistema. Si ese coeficiente es igual a 1, entonces el sistema de decisión es consistente, en otro caso es inconsistente (Skowron, 1999).

En el epígrafe 2.2 de esta tesis, se describe un método qué toma la medida de Calidad de la clasificación como heurística, y cuyo objetivo es obtener conjuntos de atributos relevantes a través de los Conjuntos Aproximados.

Los Conjuntos Aproximados y la selección de atributos

Es un hecho reconocido la posibilidad de usar los reductos como selección y reducción de atributos, no solamente en la definición de las relaciones de inseparabilidad que son la base de la teoría de los conjuntos aproximados, sino en otras aplicaciones como el descubrimiento de reglas. Sin embargo, esta beneficiosa alternativa se ve limitada por la complejidad computacional del cálculo de los reductos.

En (Bell and Guan, 1998) se muestra que el costo computacional de encontrar un reducto en un sistema de información está acotado por $n^2 \cdot m^2$, donde *n* es la cantidad de atributos y *m* es la cantidad de objetos en el universo del sistema de información. Mientras que la complejidad en

tiempo de encontrar todos los reductos es $O\left(2^n \cdot J\right)$, donde *n* es la cantidad de atributos, y *J* es el costo computacional requerido para encontrar un reducto.

La reducción de atributos a través de los conjuntos aproximados se basa en comparar las relaciones de equivalencia o similitud generadas por conjuntos de atributos. Se seleccionan atributos de manera sucesiva hasta que se obtenga un conjunto reducido tal que provea la misma calidad de la clasificación supervisada que el original. Un conjunto de datos puede tener varios conjuntos de reductos. Un objetivo importante en el cálculo de reductos es encontrar un subconjunto mínimo de estos, o sea, un subconjunto reducto con mínima cardinalidad.

Se han reportado varios trabajos para encontrar reductos, basados en la Teoría de los Conjuntos Aproximados (Korzes and Jaroszewicz, 2005, Jeon and Jeong, 2006, Mi et al., 2003, Kuo and Yajima, 2003, Caballero and Bello, 2006b). Un método que ha sido muy referenciado es el QuickReduct (Chouchoulas, 1999), que trata de calcular un reducto sin generar exhaustivamente todos los posibles subconjuntos de reductos. El algoritmo comienza con un conjunto vacío y adiciona por turno, uno cada vez, a aquellos atributos que resultan tener el mayor grado de dependencia hasta que este produce el máximo valor posible para el conjunto de datos; de esta manera, para una dimensionalidad *n* atributos, se requieren $\left(n^2 + n\right)/2$ evaluaciones de la función de dependencia en el peor de los casos. Este proceso, sin embargo, no garantiza encontrar un conjunto mínimo porque al usar la función de dependencia para discriminar candidatos puede conducir la búsqueda hacia un subconjunto que no sea mínimo. Es por eso que sobre esta misma base, otros nuevos enfoques pudieran surgir para tratar de encontrar un buen reducto.

En (Caballero and Bello, 2006a) se propone el algoritmo RSReduct para el cálculo de un buen reducto. Este es un "goloso" que comienza por un conjunto vacío de atributos y a través de heurísticas llega a formar un reducto mínimo. Para la construcción de las heurísticas se siguen criterios con respecto a la relevancia (Piñero et al., 2003), la entropía y la ganancia (Mitchell, 1997) de los atributos, así como dependencia entre atributos mediante los Conjuntos Aproximados, y la manipulación de atributos con costos diferentes. Con este algoritmo se obtienen resultados satisfactorios; sin embargo, es un algoritmo costoso. La forma en que se construye la matriz de distinción provoca que la complejidad de ir chequeando la condición de reducto sea un $O\left(m^2 \cdot n^2\right)$, para *m* objetos y *n* atributos. Por otra parte, el costo computacional de

la mejor heurística del algoritmo, es elevado debido a la formación de combinatorias entre atributos.

2.2 Solución al problema de selección de rasgos. Algoritmo ACO-RST-FSP.

El problema de selección de rasgos es un ejemplo de un problema discreto difícil. Cuando se usa ACO para resolver el problema de selección de rasgos la representación del grafo es ligeramente diferente al TSP. Este problema puede ser modelado de la siguiente forma. Sea $A = \{a_1, a_2, ..., a_{nf}\}$ un conjunto de *nf* rasgos. Este conjunto puede ser representado como un grafo bidireccional fuertemente conexo en el cual los nodos simbolizan rasgos. Los valores de la feromona τ_i están asociados a los nodos a_i. Esta es una diferencia y entre nuestro enfoque y el propuesto por Jensen and Shen [12]. En la propuesta de Jensen y Shen la feromona es asociada con los arcos lo cual es común en ACO. La cantidad de feromona en el arco $a_i - a_j$ está en función del grado de dependencia del atributo a_j sobre a_i. En el enfoque planteado en esta tesis la feromona es asociada a los nodos. La cantidad de feromona está en función de la dependencia del rasgo asociado a este nodo en correspondencia con todos los otros rasgos. Como resultado la feromona asociada al nodo a_i representa la contribución absoluta de este rasgo a un subconjunto, en lugar de la contribución de a_i dado el hecho de que a_j ha sido el rasgo previo incluido en el subconjunto.

En el primer paso, cada ant *k* es asignado a un nodo, luego esta puede moverse a cualquier nodo en el grafo ($B^k \leftarrow \{a_i\}$, donde B^k es el subconjunto que la hormiga *k* construye). Las hormigas realizan una selección hacia adelante (forward selection) en la cual cada ant *k* expande su subconjunto B^k paso a paso adicionando nuevos rasgos; para realizar esto, cada ant *k* busca todos los rasgos en $A - B^k$ (donde A es el conjunto de rasgos) y selecciona el próximo rasgo entre estos para incluirlo en B^k de acuerdo con la regla del modelo ACO en uso. Esta regla es la regla proporcional seudoaleatoria en ACS. La calidad de la aproximación de la clasificación, una medida de RST (dada por la expresión (1)) es utilizada como función heurística en el modelo ACO. Este valor también es usado para determinar si el subconjunto B^k es un reducto.

El algoritmo propuesto, (nombrado ACS-RST-FS) está basado en la variante "Ant Colony System". A continuación se establecen varios aspectos del mismo:

a) Modo de funcionamiento de las hormigas.

En esta solución cada hormiga construye un subconjunto de rasgos paso a paso. Estos subconjuntos son denotados por B^k, donde *k* denota la hormiga *k*. En el paso inicial de cada ciclo las hormigas son asociadas a nodos de acuerdo a la regla descrita en el siguiente punto b). En cada próximo paso, las hormigas seleccionan otro rasgo para incluir en sus subconjuntos. Cada nodo a_i tiene asociado un valor de feromona τ_i.

b) Distribución (posicionamiento) inicial de las hormigas.

La distribución de hormigas en cada ciclo depende de la relación entre el número de hormigas (*m*) y la cantidad de rasgos (*nf*). En (Bello and Nowé, 2005) se propone el siguiente conjunto de reglas para establecer esta relación:

i) Si *m*<*nf*, realizar una distribución inicial aleatoria de las hormigas.

ii) Si *m*=*nf*, cada hormiga es asociada a cada nodo (rasgo).

iii) Si *m*>*nf*, las primeras *nf* hormigas son distribuidas según (ii), y el resto según (i).

c) Regla proporcional seudoaleatoria.

El algoritmo ACO usa como función heurística (η) para evaluar un subconjunto *B* la calidad de aproximación de la clasificación ($\eta(B)=\gamma_B(Y)$), buscando subconjuntos *B* tales que $\gamma_B(Y)=\gamma_A(Y)$. La siguiente regla proporcional pseudoaleatoria usa este valor y el rastro de feromona en la forma:

$$i=\begin{cases}\max\left(\tau_i\cdot\left(\gamma_{B^k\cup\{a_i\}}(Y)\right)^{\beta}\right) & si \quad q\le q0\\ I & en\ otro\ caso\end{cases} \qquad \textbf{(2. 10)}$$

Donde I es seleccionado según la expresión:

$$p_k\left(B^k, a_j\right) = \begin{cases} 0 & if \quad a_j \in B^k \\ \dfrac{\left[\tau_j\right]^{\alpha} * \left[\gamma_{B^k \cup \{a_j\}}(Y)\right]^{\beta}}{\sum_{a_j \in A - B^k} \left[\tau_j\right]^{\alpha} * \left[\gamma_{B^k \cup \{a_j\}}(Y)\right]^{\beta}} & if \quad a_j \in A - B^k \end{cases} \qquad \textbf{(2. 11)}$$

d) Valor inicial del rastro de feromona.

Para calcular el valor inicial de feromona se considera la alternativa de asignar un valor aleatorio $\tau_i(0)$, i=1,…,*nf*

e) Criterio de parada de las hormigas.

Una hormiga alcanza el criterio de parada cuando termina su actividad en un ciclo. Cada hormiga adiciona un rasgo cada vez a su subconjunto parcial B^k hasta alcanzar la condición $\gamma_B(Y) = \gamma_A(Y)$.

f) Criterio de parada del algoritmo.

Para los métodos de ACO existen distintos criterios de parada (Dorigo and Stutzle, 2003). El proceso de buscar subconjuntos *B* es una secuencia de ciclos (*NC*=1, 2, …$NC_{máx}$). En cada ciclo todas las hormigas construyen su conjunto *B*. El proceso termina cuando *NC* >= $NC_{máx}$.

Algoritmo ACS-RST-FS:

Entrada: Conjunto de datos que describen mediante rasgos distintos objetos e incluye un atributo clase.

Salida: Colección de reductos.

P0:

PSC ← *Falso*

NC ← 1

Calcular $\tau_i(0)$ *i*=1,…,*nf* (valor inicial del rastro de feromona)

Repetir

P1: Estado inicial para cada ciclo.

Cada hormiga k es asociada con un atributo a_i, k=1,…,m, y $B^k \leftarrow \{a_i\}$ de acuerdo con las reglas establecidas en 2.2 b) .

$ASC_k \leftarrow Falso$, k=1,…,m

P2: Repetir

para k=1,…,m hacer

si $ASC_k = Falso$ entonces

Seleccionar el nuevo rasgo a_i^* para adicionar a B^k

$B^k \leftarrow B^k \cup \{a_i^*\}$ (de acuerdo la regla proporcional pseudoaleatoria dada en 3(c), donde a_i^* fue el último rasgo adicionado)

$\tau_i \leftarrow (1-\xi)\cdot \tau_i + \xi \cdot \tau_i(0)$ (actualización local de feromona)

Actualizar ASC_k criterio de parada de la hormiga k.

fin_si

fin_para

Hasta que $ASC_k = verdadero$ para todas las hormigas.

P3: Después que todas las hormigas han terminado:

Seleccionar la mejor solución B^k , y para todo $a_i \in B^k$ hacer $\tau_i \leftarrow (1-\rho)\cdot \tau_i + \rho \cdot \gamma_B^k(Y)$

$NC \leftarrow NC + 1$

Actualizar *PSC*

Until *PSC* = *verdadero*

Considerando que el esfuerzo computacional para calcular las aproximaciones inferior o superior es $O(nf \cdot l^2)$, donde l es el número de atributos según [1] y [8], se estima que la complejidad de este algoritmo como $O(nc \cdot m \cdot nf^2 \cdot n^2)$ donde n es la cantidad de objetos en el sistema de información; cuando la cantidad de hormigas es igual al número de rasgos la complejidad es $O(nc \cdot nf^3 \cdot n^2)$.

2.3 Evaluación del algoritmo

El objetivo de este estudio experimental es evaluar el algoritmo y determinar reglas para fijar los parámetros. Se ha evaluado el funcionamiento del algoritmo con diferentes combinaciones de parámetros. Los resultados experimentales son presentados debajo. El funcionamiento del algoritmo fue comparado según los resultados en cuanto a longitud del reducto más corto, cuantas veces se encontró el reducto más corto, cantidad de reductos encontrados y la longitud promedio de los mismos; la longitud de un reducto es definida como el número de rasgos incluidos en el reducto. En la validación del algoritmo a estos resultados se les ha denominado indicadores de salida.

2.3.1 Características de los conjuntos de datos.

Los experimentos fueron aplicados a un grupo de conjuntos de datos. Estos conjuntos presentan determinadas características, detalladas en la **Tabla 1**. Esta tabla muestra las características originales de estos conjuntos de datos, en ella se muestran los nombres, el número de clases en las que se pueden clasificar los objetos que pertenecen a cada uno de los conjuntos, el número de objetos que los componen, el total de rasgos que representa a cada objeto, los tipos de rasgos si son numéricos o nominales y los atributos que deben ser eliminados del conjunto de datos.

Conjuntos de Datos.	Número Clases	Número Objetos	Rasgos			Atributos a eliminar
			Total	Numéricos	Nominales	
Breast-w	2	699	10	9	1	1(6-Bare_Nuclei)
Dermatology[4]						
Mushroom	2	8124	23	0	23	1(11-stalk-root)
Segment	7	2310	20	19	1	0

[4] No se encontró información verdadera sobre los datos originales, se publican las características de los datos con que se trabajó.

Vowel	11	990	14	13	1	0
Splice	3	3190	62	0	62	1(1-Instance_name)
LedJen 4						
Vehicle	4	846	19	18	1	0
Ionosphere	2	351	35	34	1	0
Breast Cancer Jenk_modf 4						
HeartJen	2	294	14	0	14	0
LungJen 4						

Tabla 1 Características de los conjuntos de datos.

Las características originales de los conjuntos de datos expuestas anteriormente no cumplen con algunos de los requerimientos del algoritmo ACO-RST-FSP, pues en estos conjuntos existen rasgos de tipo numérico y este algoritmo solo trabaja con rasgos de tipo nominal, además algunos rasgos presentan falta de información de determinados objetos (estos son los propuestos en la tabla para ser eliminados). Para dar solución a estos problemas se le aplicó un proceso de discretización a los rasgos de tipo numérico, convirtiéndolos en nominales y se eliminaron los rasgos que presentaban falta de información.

2.3.2 Ajuste de parámetros.

Se estudiaron los parámetros que intervienen en el algoritmo ACO-RST-FSP con influencia de manera determinante en los indicadores que se obtienen de aplicar dicho algoritmo. Estos resultan de gran importancia cuando se desea obtener algún valor específico o resultados óptimos en los indicadores de salida, pues de la forma de variar los disímiles parámetros se derivan los resultados en los indicadores de salida. Según las conclusiones que se arribe del estudio se pueden hacer sugerencias al usuario sobre los parámetros que se deben utilizar para obtener los resultados esperados. Un listado de los parámetros fundamentales de este algoritmo, con una breve descripción de su significado se puede observar en la **Tabla 2**.

Parámetros	Breve Descripción
$\varphi \in (0,1]$	parámetro de decremento de feromona, en la regla de actualización en *línea paso a paso*
ρ	constante de evaporación, es la encargada de reducir los rastros de feromona
τ	valor inicial de la feromona
k	número de hormigas
NC	número de ciclos
α	controla el proceso de búsqueda basado en la fuerza del rastro de feromonas.
β	controlan el proceso de búsqueda basado en la fuerza de la heurística
q_0	determina el carácter del algoritmo, exploratorio o de explotación.

Tabla 2 Parámetros del algoritmo ACO-RST-FSP.

Dentro de los parámetros que conforman el algoritmo se hallan β y q_0, estos serán estudiados basados en los resultados de los indicadores de salida, arrojados al ser ejecutado el algoritmo. Se escogen estos parámetros en específico porque al comienzo de la investigación se sospechaba tenían relevante influencia en los resultados óptimos para cada uno de los indicadores. Por una parte si el algoritmo sigue una política de mayor exploración se esperan buenos resultados en los indicadores referidos a cantidades de reductos encontrados, pues el algoritmo va a ser mayor fuerza en encontrar mayor cantidad de soluciones lo que se traduce en más reductos, mientras que el parámetro β influye sobre la heurística del algoritmo haciendo que este siga más el conocimiento aportado por hormigas en recorridos anteriores, por lo que esto debe influir mayormente en el indicador referente a la longitud de los reductos, que es un indicador de calidad de la solución.

A partir de esta decisión se realizaron experimentos variando el valor de los parámetros β y q_0. Para estos se escogen los valores 1 y 5 para β, y 0.3 y 0.9 para q_0, son designados estos valores porque en el caso de la heurística se necesitan dos valores uno que mantenga la función sin alteración y otro que le proporcionara más fuerza. Para determinar el carácter del algoritmo se escogen esos valores pues el parámetro oscila en el rango de 0 a 1, durante el funcionamiento del algoritmo se escogen números al azar en este mismo rango y se comparan con valor fijado para q_0, para valores bajo se establece un carácter exploratorio y para valores altos de explotación; siguiendo estas ideas se fijan valores a los extremos del intervalo. En todos los caso se fijó el valor de alpha ($\alpha = 1$) y para los valores iniciales de $\tau_i(0)$ (la feromona) se asignaron valores aleatorios. Los algoritmos con las combinaciones de valores son aplicados a conjuntos de datos del repositorio "Repositorio UCI" (Blake and Merz, 1998), cada experimento es repetido 10 veces (debido al fuerte carácter probabilístico del algoritmo), de estos valores hallamos la mediana (es un estadístico más potente que el promedio) y el resultado de las medianas es utilizado en las comparaciones.

En (Bello and Nowé, 2005) se proponen las siguientes reglas para determinar el número de hormigas como una función de la cantidad de rasgos $k = f(nf)$. En este caso se usó la clasificación del problema de selección de rasgos dada en (Kudo and Sklansky, 2000), donde se clasifican en tres categorías: pequeña escala si $nf \in [0,19]$, mediana escala si $nf \in [20,49]$ y escala grande si $nf \geq 50$.

R1: Si $nf \in [0,19]$ entonces $k \leftarrow nf$

R2: Si $nf \in [20,49]$ entonces

[Si $0.666 \cdot nf \leq 24$ entonces $k \leftarrow 24$, sino $k \leftarrow Round(0.66 \cdot nf)$]

R3: Si $nf \geq 50$ entonces

[Si $0.5 \cdot nf \leq 33$ entonces $k \leftarrow 33$, sino $k \leftarrow Round(0.5 \cdot nf)$]

donde $Round(x)$ denota el entero más cercano a x.

La **Tabla 3** presenta el comportamiento del algoritmo para diferentes conjuntos de datos del repositorio, en esta se describen: los nombres de los conjuntos de datos con los que se realizó el experimento, así como el número de rasgos de cada objeto, el número de hormigas que se utilizaron en cada uno de los conjuntos y los resultados para las variantes de los parámetros del algoritmo. La simbología n1(n2) denota n1 reductos con n2 rasgos (sólo se muestra información sobre el reductos más cortos), TR es la cantidad resultante de reductos y TPR la longitud promedio de los reductos obtenidos.

Data base	**na**	**k**	**ACS-RST-FS (β=5, q_0=0.9)**	**ACS-RST-FS (β=5, q_0=0.3)**	**ACS-RST-FS (β=1, q_0=0.3)**	**ACS-RST-FS (β=1, q_0=0.9)**
Breast Cancer Jenk_modf	9	9	5(4) TR=10.5 TPR=4.45	6(4) TR=17 TPR=4.76	6(4) TR=18.5 TPR=4.75	6(4) TR=14 TPR=4.62
breast-w	8	8	8(5) TR= 9 TPR= 5.1	9(5) TR= 11 TPR= 5.23	9(5) TR= 12 TPR= 5.25	9(5) TR= 10 TPR= 5.18
Dermatology discreto MLP	34	24	8.5(9) TR= 184.5 TPR=11.03	30.5(10) TR= 352.5 TPR=13.46	4(9) TR= 343.5 TPR=13.99	6(9) TR= 167 TPR=11.14
heartJen	13	13	1(6) TR= 11.5 TPR= 7	1(6) TR= 16.5 TPR= 7.37	1(6) TR= 19.5 TPR= 7.5	1(6) TR= 12 TPR= 7.07
ledJen discreto	24	24	1(5) TR= 1.5 TPR= 5.5	1(5) TR= 6 TPR= 11.89	1(5) TR= 13.5 TPR= 10.48	1(5) TR= 2.5 TPR= 9.15

lungJen discreto	56	33	12(4) TR= 308.5 TPR= 5.91	11.5(4) TR= 629,5 TPR= 6,74	7.5(4) TR= 604.5 TPR= 7.29	13.5(4) TR= 323.5 TPR= 6.05
ionosphere	34	24	15(7) TR=129 TPR=8.70	9(7) TR=355 TPR=10.66	6(7) TR=342 TPR=11.39	19(7) TR=160 TPR=8.59
mushroom	21	24	3(4) TR=54 TPR=5.16	3(4) TR=128 TPR=5.97	2(4) TR=117 TPR=6.08	3(4) TR=51 TPR=5.26

Tabla 3 Comparación entre algoritmos usando conjuntos de datos del "Repositorio UCI".

A los resultados mostrados en la tabla anterior, se le aplica una validación estadística. En dicha validación se aplica el test de Friedman, pues las muestras no siguen una distribución normal, además están relacionadas, para las peculiaridades que presentan las poblaciones este es el test recomendado, luego de aplicar el test de Friedman se aplica el test de Iman-Davenport para resaltar la significación, en caso que aparezca, luego para realizar una comparación entre las poblaciones se utilizo el test de Nemenyi utilizando este, el test de Holm para las comparaciones.

Para realizar la validación estadística y llegar a conclusiones de la relación existente entre los parámetros y los indicadores de salida, se llevó a cabo una comparación entre los datos de cada una de las muestras para los distintos indicadores. En esta comparación primero se nota si existe significación entre las variables, luego se muestran dos tablas. En una se muestra los rangos de las poblaciones, estos rangos destacan el comportamiento de las muestras, resultando mejor la que menor rango tenga. En la otra se realiza una comparación entre todas las muestras par-a-par para resaltar entre cuales existen diferencias significativas. En el caso que no exista significación entre las muestras la primera de las tablas puede mostrar tendencias a seguir por las poblaciones con respecto al indicador de salida correspondiente. Al realizar la validación estadística se obtuvieron los resultados que se muestran en las siguientes tablas.

El indicador que describe el total de reductos calculados presenta un valor de 1.867E-4 en la significación al aplicar el test de Friedman, esto expresa la existencia de diferencias significativas entre las muestras, para ser más precisos se aplica el test de Iman-Davenport con el cual se obtiene 3.9099E-8 como valor de significación, por tanto según este resultado también existe diferencias significativas. En la **Tabla 4** se muestran los valores de los promedios de los rangos de las poblaciones.

Algoritmos	**Rangos**
Total de reductos calculados (B=5 q0=0.9)	3.75
Total de reductos calculados (B=1 q0=0.9)	3.25
Total de reductos calculados (B=1 q0=0.3)	1.5
Total de reductos calculados (B=5 q0=0.3)	1.5

Tabla 4 Promedio de rangos de las muestras

En la **Tabla 5** se muestra la comparación entre las poblaciones dos a dos, utilizando el test de Nemenyi, que a su vez utiliza el test de Holm para las comparaciones con el valor p, por medio de este se puede saber si se acepta o rechaza la hipótesis de igualdad entre las muestras (si éste es menor o igual que el margen de error permitido (dado por el test de Holm), la hipótesis es rechazada, en caso contrario se acepta)

i	Algoritmos	p	Holm
1	Total de reductos calculados (B=5 q0=0.9) vs. Total de reductos calculados (B=1 q0=0.3)	4.9088E-4	0.0083

2	Total de reductos calculados (B=5 q0=0.9) vs. Total de reductos calculados (B=5 q0=0.3)	4.9088E-4	0.01
3	Total de reductos calculados (B=1 q0=0.9) vs. Total de reductos calculados (B=1 q0=0.3)	0.0067	0.0125
4	Total de reductos calculados (B=1 q0=0.9) vs. Total de reductos calculados (B=5 q0=0.3)	0.0067	0.0167
5	Total de reductos calculados (B=5 q0=0.9) vs. Total de reductos calculados (B=1 q0=0.9)	0.4386	0.025
6	Total de reductos calculados (B=1 q0=0.3) vs. Total de reductos calculados (B=5 q0=0.3)	1.0	0.05

Tabla 5 Nemenyi / Holm

Con las combinaciones de valores para los parámetros $\beta = 1, q_0 = 0.3$ y $\beta = 5, q_0 = 0{,}3$, se encuentra una diferencia significativa entre el total de los reductos encontrados entre estos valores y los demás, lo que hace notar que si el valor de q_0 es bajo, se encuentran más reductos, sin tener en cuenta el valor de β, este comportamiento es atribuido al carácter exploratorio de ACO cuando q_0 es bajo con lo cual se encuentran mayor diversidad en las soluciones encontradas.

El indicador que describe la longitud de reductos más cortos obtiene de la aplicación de los tests de Friedman e Iman-Davenport los valores 0.225 y 0.973 respectivamente, lo que indica que no existe diferencia significativa entre las muestras; por tanto sólo se puede plantear estas van a seguir una tendencia descrita en la **Tabla 6**. En esta tabla se muestran los valores de los promedios de los rangos de las poblaciones, pero en esta ocasión para el indicador longitud de reducto más corto.

Algoritmos	Rangos
Total de reductos calculados (B=5 q0=0.9)	2.4375
Total de reductos calculados (B=1 q0=0.9)	2.4375
Total de reductos calculados (B=1 q0=0.3)	2.4375
Total de reductos calculados (B=5 q0=0.3)	2.6875

Tabla 6 Promedio de rangos de los muestras

Con esta combinación de parámetros el algoritmo tiende a encontrar reductos más pequeños cuando el valor de q_0 es alto o el de β es bajo, es decir el algoritmo va a trabajar más con la explotación por lo que va a considerar más la experiencia acumulada por las hormigas.

El indicador que describe la cantidad de reductos más cortos obtiene de la aplicación de los tests de Friedman e Iman-Davenport los valores 0.327 y 0.34292 respectivamente, lo que indica que no existe diferencia significativa entre las muestras por lo que estas van a seguir una tendencia descrita en la **Tabla 7**. En esta tabla se muestran los valores de los promedios de los rangos de las poblaciones, pero en esta ocasión para el indicador cantidad de reducto más corto.

Algoritmos	**Rangos**
Total de reductos calculados (B=5 q0=0.9)	2.625
Total de reductos calculados (B=1 q0=0.9)	2.0
Total de reductos calculados (B=1 q0=0.3)	3.125
Total de reductos calculados (B=5 q0=0.3)	2.25

Tabla 7 Promedio de rangos de los algoritmos.

Con esta combinación de parámetros el algoritmo tiende a encontrar mayor cantidad de reductos más pequeños cuando el valor de q_0 es alto y el de β es bajo.

Para concretar, entre las cuatro variaciones de parámetros con la que se obtienen mejores resultados es con la combinación de parámetros $\beta = 1$, $q_0 = 0.3$.

2.3.3 Comparación con otros métodos

La temática de selección de rasgos es muy estudiada en la actualidad, pues no se ha encontrado aún un algoritmo óptimo para esta tarea, por lo que existe toda una gama de estos tipos de algoritmos. En la **Tabla 8** se lleva a cabo una comparación entre dos de estos incorporados en WEKA con respecto al ACO-RST-FSP, en esta tabla se muestra el nombre de los conjuntos de datos con los que se llevo a cabo el experimento, el número de rasgos que describe a cada objeto, la cantidad de hormigas utilizadas para aplicar el algoritmo a cada conjunto de datos y los resultados de cada método con dos funciones de evaluación de subconjuntos (i- CfsSubsetEval, ii- ConsistencySubsetEval).

Conjuntos de datos	**na**	**k**	**ACO-RST-FSP**	**Best-First**		**Greedy Stepwise**	
				i	**ii**	**i**	**ii**
Breast Cancer Jenk	9	9	6(4)	1(9)	1(8)	1(9)	1(8)
breast-w	8	8	9(5)	1(8)	1(5)	1(8)	1(5)
Dermatology	34	24	4(9)	1(19)	1(11)	1(19)	1(11)
heartJen	13	13	1(6)	1(6)	1(8)	1(6)	1(8)
ledJen discreto	24	24	1(5)	1(12)	1(5)	1(12)	1(5)
lungJen discreto	56	33	7(4)	1(12)	1(4)	1(12)	1(4)
ionosphere	34	24	6(7)	1(12)	1(4)	1(12)	1(4)
mushroom	21	24	2(4)	1(4)	1(5)	1(4)	1(5)

Tabla 8 Comparación entre algoritmos de selección de rasgos

2.4 Solución al problema de la selección de rasgos en contexto distribuido. El algoritmo D.ACO-RST-FSP.

El problema de la selección de rasgos en contexto distribuido, en general, es similar al problema clásico de selección de rasgos, pero en este hay varias fuentes de datos donde lo algoritmos son aplicados con el mismo objetivo. El modelo que se presenta trata con conjuntos de datos homogéneos.

Como solución se propone un nuevo algoritmo inspirado en la idea del ACS multi-tipo. La principal idea es realizar la tarea de selección de rasgos en los conjuntos de datos en la misma forma que lo hace el modelo ACS-RST-FSP, pero ahora tomando en cuenta la experiencia ganada en el resto de los subsistemas realizando la misma tarea. La experiencia ganada se expresa mediante el rastro de feromona y se trasmite a las demás colonias. De esta manera, el algoritmo tiene su propio rastro de feromona y conoce el rastro dejado por hormigas de colonias colaborativas. Adicionalmente, las colonias intercambian su rastro "frecuentemente".

De manera similar al ACS multi-tipo se extiende la fórmula de transición probabilística (2.3) como sigue en (2.10) y la regla seudo-aleatoria proporcional (2.4) como sigue en (2.9).

En estas fórmulas, ϕ_j^s representa la cantidad promedio de rastro de feromona perteneciente a otras colonias en el nodo *j*. Este es el promedio de todos los rastros de feromona dejado otras hormigas en diferentes colonias realizando la misma tarea para otros conjuntos de datos. La potencia γ indica la sensitividad de la hormiga para seguir su propia experiencia o la experiencia ganada por el resto de las colonias. Este es un parámetro a ser estudiado, pero claramente si a γ es asignado valor cero las hormigas calcularán la probabilidad basada en la heurística del problema y la feromona de su propia colonia, tal como el algoritmo original siendo ignorado el rastro de feromona de las otras colonias. Si γ es incrementado, la probabilidad de seleccionar un nodo se convertiría cada vez más dependiente de la experiencia del resto de las colonias.

2.4.1 Ajuste de parámetros

En el algoritmo D.ACO-RST-FSP, al igual que en la versión en contexto local, existe la influencia de distintos parámetros, varios de ellos explicados anteriormente. En este se incluyen nuevos parámetros propios de contexto al que pertenece el algoritmo (Ver**Tabla 9** Parámetros del algoritmo D.ACO-RST-FSP **Tabla 9**).

Parámetros	Breve Descripción
NI	Número de intercambios del grafo de feromona
γ	Controla el promedio de las feromonas del resto de las colonias

Tabla 9 Parámetros del algoritmo D.ACO-RST-FSP

Estos parámetros se adicionan a los presentados con anterioridad en el algoritmo ACO-RST-FSP. De estos nuevos parámetros, resulta de gran interés el estudio del parámetro que interviene en el intercambio de información, expresando cada cuantas iteraciones va a ocurrir un intercambio de información. Este parámetro resulta interesante pues se necesita conocer la influencia que presenta en un conjunto de datos, el conocimiento de otros de los subsistemas. Con este fin se llevaron a cabo distintos experimentos, en los cuales, el parámetro que representa los números de iteraciones a los cuales se efectúa un intercambio, toma valores de 5, 10 y 20, lo que significa que existe un intercambio al 20%, 10% y 5% del total de los ciclos.

Para realizar estos experimentos no fueron encontrados conjuntos de datos preparados para trabajar en el ambiente distribuido, por esto se escogieron varios conjuntos del repositorio UCI para ser particionados y de esta manera simular el ambiente distribuido. Al igual que la versión en ambiente local, este algoritmo tiene un carácter estocástico, por lo cual se realiza el mismo proceso que en la validación del epígrafe anterior. De forma similar se aplica la validación estadística de las muestras para cada uno de los parámetros, aplicando los mismos tests que en el tópico anterior. Las tablas que se describen a continuación contienen el mismo tipo de información que las mostradas para el algoritmo en contexto local.

El indicador que describe el total de reductos calculados presenta un valor de 0.01102 al aplicar el test de Friedman, esto expresa la existencia de diferencias significativas entre las muestras, para ser más precisos se aplica el test de Iman-Davenport con el cual se obtiene 0.00903 como valor de significación, por tanto para este también existe diferencias significativas. En la **Tabla 10** Promedio de rangos de las muestras. se muestran los valores de los rangos de las poblaciones.

Algoritmos	**Rangos**
Total de reductos calculados cada 5 ciclos	2.125
Total de reductos calculados cada 10 ciclos	1.578125
Total de reductos calculados cada 20 ciclos	2.296875

Tabla 10 Promedio de rangos de las muestras.

En la **Tabla 11**se muestra la comparación entre las poblaciones, utilizando el test de Nemenyi el cual utiliza el test de Holm para las comparaciones.

i	Algoritmos	p	Holm
1	Total de reductos calculados cada 10 ciclos vs. Total de reductos calculados cada 20 ciclos	0.00404	0.01666
2	Total de reductos calculados cada 5 ciclos vs. Total de reductos calculados cada 10 ciclos	0.02871	0.025
3	Total de reductos calculados cada 5 ciclos vs. Total de reductos calculados cada 20 ciclos	0.49177	0.05

Tabla 11 Nemenyi / Holm

La muestra resultante del algoritmo con intercambio al 10% del total de ciclos es significativamente mayor encontrando reductos que la resultante del algoritmo con intercambio al 5% del total de ciclos, mientras que con respecto a la variante calculada al 20% del total de ciclos no ofrece diferencias significativas, lo que demuestra la importancia del intercambio de información de forma periódica.

El indicador que describe la longitud de reductos más cortos obtiene de la aplicación de los tests de Friedman e Iman-Davenport los valores 0.55658 y 0.56389 respectivamente, lo que indica que no existe diferencia significativa entre las muestras por lo que estas van a seguir una tendencia descrita en la **Tabla 12**.

Algoritmos	**Rangos**
Total de reductos calculados cada 5 ciclos	1.921875
Total de reductos calculados cada 10 ciclos	1.921875
Total de reductos calculados cada 20 ciclos	2.15625

Tabla 12 Promedio de rangos de los algoritmos

El algoritmo tiende a encontrar reductos más pequeños cuando el intercambio ocurre al 10% y al 20% del total de ciclos. Esto puede estar dado por el resultado del indicador anterior pues, si son encontrados más reductos, existe la posibilidad de encontrar los más cortos.

El indicador que describe la cantidad de reductos más cortos obtiene de la aplicación de los tests de Friedman e Iman-Davenport los valores 0.94678 y 0.94836 respectivamente, lo que indica que no existe diferencia significativa entre las muestras por lo que estas van a seguir una tendencia descrita en la **Tabla 13**.

Algoritmos	**Rangos**
Total de reductos calculados cada 5 ciclos	2.03125
Total de reductos calculados cada 10 ciclos	1.953125
Total de reductos calculados cada 20 ciclos	2.015625

Tabla 13 Promedio de rangos de los algoritmos.

El algoritmo tiende a encontrar una mayor cantidad de los reductos más cortos cuando el intercambio ocurre al 10% del total de ciclos.

Como conclusión se puede observar que para obtener los mejores resultados debemos mantener un intercambio al 10% del total de ciclos.

2.5 Consideraciones parciales.

1. Como una cuestión importante en esta investigación se ha explicado y validado como extender ACO convirtiéndolo en un ACO multicolonias mediante intercambios de feromona; donde cada colonia representa un algoritmo ACO resolviendo un problema con un comportamiento colaborativo entre hormigas de otras colonias mediante intercambios "frecuentes" de feromona.
2. La principal ventaja de este modelo es encontrar varios subconjuntos de rasgos, como soluciones a la misma vez; a diferencia de otros enfoques propuestos.
3. El trabajo de investigación muestra que la relación entre los parámetros beta (β) y q_0 tiene el impacto más fuerte en la cantidad de subconjuntos generados por el algoritmo ACS-RST-FS. También esta relación influye sobre el número de rasgos seleccionados.
4. Medir la calidad de un subconjunto dado como solución en un algoritmo de selección de rasgos cuando este es un filtro es aún un problema abierto en la temática.

Capítulo III. Algoritmos ACO-RST-FSP y su variante distribuida D.ACO-RST-FSP aplicados problemas reales.

Las prestaciones de las computadoras en la actualidad con su gran capacidad de almacenamiento, han sido de gran ayuda para acumular toda la información que el usuario determine que es más importante, de este modo para algunas instituciones la prioridad ha sido almacenar lo ocurrido a través del tiempo en las distintas esferas. Este compendio de datos almacenados es de gran importancia para las instituciones que de una manera u otra necesitan acceder a datos de las mismas, ocurridos anteriormente, este es el caso de las recopilaciones de datos que llegan a nuestras manos para ser procesadas, específicamente en el área de meteorología. En ocasiones estos datos no se encuentran en su forma óptima por lo que es necesario realizar algunas modificaciones, para así, tener los datos en condiciones favorables para su procesamiento, lo que influye de manera considerable tanto en los algoritmos que se puedan aplicar como en las conclusiones a las que se quiera llegar, después del análisis de los mismos.

La preparación de los datos es un paso importante en minería de datos y DDM no es una excepción. Lo idóneo es que en DDM el preprocesamiento se realice en un estilo distribuido. Varias de las técnicas de preprocesamiento de datos centralizados pueden ser aplicadas directamente sin descargar todos los conjuntos de datos a un sitio central. Estandarizar los datos entre los diferentes conjuntos es un proceso importante en DDM; el primer paso es intercambiar el esquema de información, es decir, la estructura del conjunto de datos (atributos, unidades de medida de estos, cantidad de objetos, etc.). Típicamente esto involucra bajo costo de comunicación. En ocasiones se intercambia información adicional con información considerando el significado físico de los rasgos, y otra información específica del dominio de aplicación con el objetivo de una mejor comprensión de las fuentes de datos distribuidas.

3.1 El problema del pronóstico climático

El pronóstico meteorológico o pronóstico del tiempo es una estimación a corto plazo, a escala cronológica utilizado para predecir valores de variables meteorológicas. Por otra parte el pronóstico climático se encarga de estimar las variaciones del clima en un período de tiempo con

el que se esté trabajando; por ejemplo sequía o exceso de humedad, temperaturas medias del invierno, etc. Estas son estimaciones a mediano o largo plazo, pues en este caso se infiere a razón de decenas, meses, años y décadas.

Los pronósticos climáticos son de tipo probabilístico, por tal razón no son exactos en cuestiones de predecir el comportamiento de los valores que pueden alcanzar las variables, solo infieren sobre la permanencia por encima o por debajo, del valor de la media, esta media corresponde a lo ocurrido en este mismo período de tiempo en 5 años anteriores, por lo planteado anteriormente se hace referencia a un problema de clasificación, pues el valor clasifica en un rango los superiores y los inferiores a la media.

En el caso de los pronósticos climáticos, como se está haciendo referencia a series de tiempo, se necesita contar con el valor alcanzado por ciertas variables en períodos de tiempo anteriores, estos valores influyen en las condiciones actuales del tiempo y tendrán repercusión en el futuro, en distintos grados. A los datos originales suministrados se aplica un algoritmo de selección de rasgos para determinar la que verdaderamente influyen en el comportamiento del rasgo objetivo manteniendo una información precisa de los valores alcanzados en el pasado, sin hacer alusión a todas las variables que conforman el conjunto de datos. Este tema de la selección de rasgos aplicada a los pronósticos climáticos será abordado posteriormente de manera más amplia en este capítulo.

3.1.1 Ingeniería del conocimiento. Transformar la serie de tiempo en un conjunto de datos para ML.

En el caso específico de los "conjuntos de datos" de meteorología, se trabaja con la información adquirida de cuatro estaciones que coleccionan estos datos en la provincia Villa Clara: Caibarien, INIVIT, Sagua y Yabú. De estas se obtienen los valores de 12 variables climatológicas en un período de tiempo de alrededor de 10 años, desde enero de 1977 hasta noviembre del 2007, con la excepción del recopilado en INIVIT, el cual está recogido a partir de 1979.

En este trabajo se emplean las decenas (10 días) como período para acumular los datos, ya que este es el espacio de tiempo en el que se recogen los valores alcanzados, por los factores climatológicos en las estaciones meteorológicas dedicadas a coleccionar estos datos. Para cada

una de las variables los datos se obtienen de manera distinta, en su mayoría el valor que reflejan es el valor de la media en la decena correspondiente, solo en la variable lluvias no se sigue este mismo procedimiento, el valor que se refleja es el total de lluvias en la decena. Luego de obtener los datos con este formato, las conclusiones sobre los pronósticos a los que se quieren llegar se harán sobre la base de estimaciones decenales.

Los datos recopilados en las estaciones meteorológicas registran los valores alcanzados por las variables meteorológicas en una decena determinada, y se detallan en la **Tabla 14** . Dentro de los atributos especificados se encuentra la variable que recoge el total de lluvias ocurridas en la decena, el pronóstico de esta variable es el objeto de estudio en este trabajo, se pretende estimar el comportamiento de las lluvias para las decenas venideras, esto con ayuda de los valores alcanzados con anterioridad por esta variable, y por las otras que conforman el conjunto de datos.

Variable	Nombre de la variable climática	U/M
X_1	Temperatura media (promedio en la decena)	°C
X_2	Temperatura máxima (promedio en la decena)	°C
X3	Temperatura mínima (promedio en la decena)	°C
X4	Humedad relativa máxima (media en la decena)	%
X5	Humedad relativa mínima (media en la decena)	%
X6	Déficit de saturación (media en la decena)	mm
X7	Total de lluvias en la decena	mm
X9	Tensión de vapor de agua (promedio en la decena)	mm
X10	Humedad relativa media (promedio en decenas)	%
X11	Nubosidad (Promedio en octavos en la decena)	Octavos
X12	Velocidad meda del viento (promedio en la decena)	m/s
X13	Presión atmosférica (promedio en milímetros en la decena)	mmHg

Tabla 14. Descripción de las variables climáticas.

Los datos meteorológicos recopilados de este modo, no brindan la información deseada, por lo cual deben enfrentar un preprocesamiento, para de esa manera quedar en forma óptima para adquirir de ellos la información más exacta posible. En este preprocesamiento transitan por las fases de eliminar los datos que presenten falta de información, arreglar los datos exagerados y realizar las transformaciones necesarias para que los datos estén listos para ser procesados.

Falta de información

La ausencia de información es un problema real en casi la totalidad de las aplicaciones. La mayoría de las técnicas más simples como remplazo por algún valor constante o valor esperado podría o no trabajar dependiendo del dominio de aplicación. Con frecuencia estos datos introducidos pueden parcializar o predisponer la solución con posible repercusión en el rendimiento de alguna tarea de minería de datos. Técnicas más sofisticadas para la manipulación de valores ausentes requieren modelos predictivos de datos como árboles de decisión u otros modelos inductivos para predecir los valores perdidos. Sin embargo, la aplicación de tales técnicas aún no garantiza la integridad de la calidad del valor obtenido, tan solo con una fiabilidad como la del modelo inductor.

En este trabajo en particular se encuentra falta de información en todos los conjuntos de datos. La técnica por la que se apuesta en esta situación consiste en sustituir la falta de información (se hace referencia a la falta de valores en un lugar determinado del conjunto de datos) por un valor esperado, este valor esperado es la media aritmética. El valor de la media es calculado a partir de los valores de la variable meteorológica de la decena especificada en los años que se almacenan en el conjunto, y este valor es el que situamos donde se encuentra la falta de información, este procedimiento se ejecuta en las decenas donde se encuentra la ausencia de valores, en toda la extensión del conjunto de datos que se esté analizando, de esta manera se completan todos los datos y se obtiene un conjunto sin falta de información.

Limpieza de los datos

Al trabajar con datos recopilados anteriormente, se corre el riesgo de que estos datos no estén en su totalidad en buenas condiciones. En estos puede existir inflación, ruido, u otros fenómenos que influyan en la comprensión del conjunto de datos de manera desfavorable. Además datos falsos contribuyen a la obtención de predicciones erróneas en los nuevos objetos. En particular

cuando se realizan labores de minería de datos con conjuntos de datos de problemas prácticos reales se debe tener especial cuidado con los datos erróneos debido a que estos influyen de manera negativa en el proceso que se quiera realizar. Por tal motivo en la fase de limpieza se eliminan los datos falsos, que pueden traer problemas a la tarea que se quiera llevar a cabo.

En el proceso de limpieza se debe tener especial cuidado con los valores que pueden llegar a obtener las variables. En esta aplicación en particular es posible encontrar, por ejemplo, temperaturas excesivamente bajas, lluvias con valores negativos o con valores por encima del límite razonable, vientos con valores negativos o con valores exagerados, entre otras situaciones. Esto solo consigue deteriorar de manera considerable la obtención de un buen resultado, para la solución de estos problemas se deben fijar límites lógicos en los que oscilen los valores de las variables en cuestión, cuando se hace referencia a mediciones de temperaturas se reconocen como valores no válidos aquellos que estén 5 grados por encima o por debajo del record histórico de la variable en esa categoría. Al encontrar algún valor erróneo de esta índole este valor se sustituye de la misma forma que si se tratara de un valor ausente, se sustituye por el promedio.

Transformaciones a la estructura de los datos.

En diversas ocasiones se encuentran datos ya procesados con los procedimientos explicados anteriormente, pero frecuentemente estos no son suficientes para que los datos cumplan con los requerimientos que se necesitan para trabajar con ellos; en casos determinados, para dar solución a este problema se llevan a cabo transformaciones a los datos originales. Las transformaciones pueden ser variadas con el fin de adaptar los datos a la situación en la que se trabaja.

Hay varios dominios que no se ajustan fácilmente al modelo *atributo-valor-clase* común en la familia de métodos de aprendizaje automatizado. Ejemplos de estos lo constituyen las series de tiempo multivariadas, reconocimiento de secuencias, análisis de canasta y de bitácoras de páginas web. Por esta razón hay que realizar algunas acciones para aplicar algoritmos de esta familia sin embargo se tienen pocas opciones: aplicar preprocesamiento hecho a mano, escribir algoritmos específicamente diseñados para el domino o utilizar algoritmos con una representación más potente (Kadous and Sammut, 2005). En esta investigación se ha utilizado la primera opción a pesar de ser consumidora de tiempo y requerir conocimiento profundo del dominio de aplicación.

El clima es un factor en el que intervienen varios agentes, casi todos medibles; entre los que existe una marcada dependencia en su comportamiento, con respecto a los valores que alcanzan en un tiempo determinado. Además, el comportamiento de las variables en un período de tiempo en particular se ve influenciado por los valores obtenidos en etapas anteriores por esta misma variable y por otras que estén vinculadas a ella. En consecuencia para aplicar con éxito técnicas de aprendizaje automatizado se hace necesario incluir en cada tupla de datos los valores de las variables, en otros períodos de tiempo, que influyen sobre el rasgo objetivo. A este grupo de valores adicionados se le llama *datos de retardo*. Sobre los conjuntos de datos de meteorología que se estudian fueron aplicadas técnicas estadísticas que permitieron determinar los datos de retardo, quedando detallados en la **Tabla 15**.

NOMBRE	DESCRIPCIÓN	TIPO
Atributos asociados al valor de lluvias		
$X_{7\text{-}36}$, $X_{7\text{-}72}$, $X_{7\text{-}108}$, $X_{7\text{-}144}$	Atributos diferentes, correspondientes a los valores reales del total de lluvias acumulados en las decenas con 1, 2, 3 y 4 años de anterioridad.	real
Atributos asociados al valor de la nubosidad		
$X_{11\text{-}1}$, $X_{11\text{-}2}$, $X_{11\text{-}3}$, $X_{11\text{-}4}$, $X_{11\text{-}36}$	Atributos que describen el valor de la nubosidad en las últimas cuatro decenas y en la propia decena 1 año atrás.	real
Atributos asociados al valor de la presión atmosférica		
$X_{13\text{-}1}$, $X_{13\text{-}2}$, $X_{13\text{-}3}$, $X_{13\text{-}4}$, $X_{13\text{-}36}$	Atributos que describen el valor de la presión atmosférica en las últimas cuatro decenas y en la propia decena 1 año atrás.	real

Atributos asociados al valor de la temperatura máxima		
X_{2-1}, X_{2-2}, X_{2-3}, X_{2-4}, X_{2-36}	Atributos que describen el valor de la temperatura máxima en las últimas cuatro decenas y en la propia decena 1 año atrás.	real
Atributos asociados al valor de la humedad relativa		
X_{10-1}, X_{10-2}, X_{10-3}, X_{10-4}, X_{10-36}	Atributos que describen el valor de la humedad relativa media en las últimas cuatro decenas y en la propia decena 1 año atrás.	real
X_{2-1}, X_{2-2}, X_{2-3}, X_{2-4}, X_{2-36}	Atributos que describen el valor de la humedad relativa mínima en las últimas cuatro decenas y en la propia decena 1 año atrás.	real

Tabla 15 Atributos asociados a la variable objetivo

3.1.2 Selección de rasgos aplicado a datos meteorológicos.

En este problema en específico, se toman los mejores valores de los parámetros, resultados de las conclusiones a las que se arriban después de realizar la validación estadística en el capítulo anterior, $\beta=1$ y $q_0=0.3$, para la variante en contexto local mientras que en contexto distribuido, se realizan intercambios al 10% del total de ciclos, de estos dos experimentos se obtienen los siguientes resultados (Ver **Tabla 16**). En la tabla se detallan los indicadores de salida con los valores alcanzados por los dos algoritmos, uno en contexto local y el otro en contexto distribuido, en las determinadas estaciones de las que se obtuvieron los datos.

Indicadores de Salida	**Conjuntos de Datos**	**ACO-RST-FSP**	**D.ACO-RST-FSP**
Longitud del reducto	Caibarién	8	8

más corto	INIVIT	8	8
	Sagua	7	7
	Yabú	8	8
Cantidad de reductos más cortos	Caibarién	203	214
	INIVIT	159	200
	Sagua	2	1
	Yabú	141	180
Total de reductos calculados	Caibarién	1071	1949
	INIVIT	1096	1958
	Sagua	1105	2014
	Yabú	1096	2045

Tabla 16 Resultados de los algoritmos en contexto local y distribuido aplicados al caso de estudio Meteorología

Los resultados en general favorecen al algoritmo aplicado en contexto distribuido. Aunque en la mayoría de los indicadores no es de forma significativa se observa una ventaja en la mayoría. Esto indica la superioridad del algoritmo aplicado con colaboración entre lo subsistemas, sobre el aplicado a estos mismos datos de manera local.

3.2 Algoritmos de selección de rasgos aplicados al problema de cardiopatías.

Otro caso de estudio mostrado en este trabajo de investigación, es el referente a enfermedades de cardiopatías, específicamente predicción de pacientes propensos a sufrir un infarto agudo del miocardio (IMA). Contar con un software que permita hacer un pronóstico efectivo de personas propensas a sufrir un IMA resulta de mucha utilidad y si se suma que este pronóstico se puede

lograr con un grupo pequeño de preguntas al paciente y apenas un reducido grupo de pruebas clínicas elementales como la medición de la presión arterial el beneficio que reporta es considerable.

Posiblemente cualquier tarea de aprendizaje que se desarrolle puede ganar mayor calidad si se toman en cuenta datos de pacientes obtenido en otras clínicas que colaboran ofreciendo información sobre casos clínicos anteriores. Concretamente en esta investigación se cuentan con datos de 3 clínicas. En este tópico se describe como se realiza un proceso de selección de rasgos a partir de los datos de la clínica donde arribe el paciente o a partir de la colaboración entre las 3 clínicas. En cualquier alternativa los datos almacenados tienen la misma estructura descrita. Y cada municipio cuenta con los datos de su clínica en una base de casos disponible con un grupo de pacientes y su cuadro clínico descrito por las variables detalladas en la **Tabla 17**. Por tanto se trata de aplicar el algoritmo ACO-RST-FSP y su variante distribuida D.ACO-RST-FSP sobre estos datos.

Variable	**Nombre de la variable**	**Valores posibles**
X1	Grupo	clase
X2	Número	eliminado manualmente
X3	Iniciales	eliminado manualmente
X4	Sexo	F o M
X5	Edad	un valor entero
X6	Raza	B o N
X7	Antecedentes de IMA	Sí o No
X7	Angina de pecho	Sí o No
X8	Otras formas	Sí o No
X9	Fuma	Sí o No
X10	Tiempo de fumador	“No fuma” o “1-5 años” o “6-10 años” o “Más de

		10”
X11	Fuma cigarrillos	Sí o No
X12	Fuma tabacos	Sí o No
X13	Fuma pipa	Sí o No
X14	Cantidad de cigarros	"Ninguno" o "Hasta 10" o "De 11 a 20" o "Más de 20"
X15	HTA	Sí o No
X16	Grado de HTA	“Sin HTA”o “Grado I”o “Grado II”o “Grado III”
X17	Tiempo de HTA	“Sin HTA”o “10 o menos” o “11-20 anos” o “Mas de 20”
X18	Antec. de diabetes M.	Sí o No
X19	Tipo de diabetes	“Sin diabetes” o “I” o “II”
X20	Tiempo de diabetes	un entero
X21	Obesidad	Sí o No
X22	Sedentarismo	Sí o No
X23	Practica ejercicio	Sí o No
X24	Pertenece a círculo	Sí o No
X25	Hiperlipoprotenimia	Sí o No
X26	Tipo hiperlipoprotenimia	“Sin hip” o “I” o “II” o “III” o “IV” o “V”
X27	Ingestión de bebidas alcohólicas	“Nunca” o “Esporádicamente” o “Frecuentemente”
X28	Stress en el hogar	Sí o No

X29	Stress en el trabajo	Sí o No
X30	Stress en otro lugar	Sí o No

Tabla 17 Rasgos de una base de casos de cardiopatías

Al aplicar el algoritmo con sus dos variantes, con los mejores parámetros resultantes de la validación estadística, como en el caso de estudio anterior se obtuvieron los mismos resultados, favoreciendo estos al algoritmo D.ACO-RST-FSP (Ver **Tabla 18Tabla 18**)

Indicadores de Salida	**Conjuntos de Datos**	**ACO-RST-FSP**	**D.ACO-RST-FSP**
Longitud del reducto más corto	Municipio 1	5	5
	Municipio 2	6	6
	Municipio 3	4	4
Cantidad de reductos más cortos	Municipio 1	14	14
	Municipio 2	1	2
	Municipio 3	1	4
Total de reductos calculados	Municipio 1	39	84
	Municipio 2	79	87
	Municipio 3	99	133

Tabla 18 Resultados de los algoritmos de selección de rasgos aplicados al caso de estudio de cardiopatías.

3.3 Consideraciones parciales

1. Al aplicar el algoritmo de selección de rasgos ACO-RST-FSP y D.ACO-RST-FSP a dos problemas reales donde la información se encuentra esparcida en varios conjuntos de

datos la variante distribuida presentó mejores resultados para los indicadores *cantidad de reductos más cortos* y *total de reductos calculados*; sin embargo, la *longitud del reducto más corto* fue la misma en ambos casos.

2. En el problema del pronóstico climático permitió hacer una reducción de 40 rasgos a 8. Además de la reducción de dimensionalidad, significa que se evita un recálculo innecesario en varios de los rasgos que inicialmente constituían los datos de retardo.
3. Con la aplicación en los conjuntos de datos de cardiopatía se obtuvieron tres resultados trascendentales:
    a. Se eliminaron rasgos dependientes entre sí, lo que facilitó aplicar clasificadores que exigen independencia entre los rasgos.
    b. Las combinaciones de subconjuntos de variables (la colección de reductos) que resultaron facilitan a especialistas cardiólogos qué preguntas realizar.
    c. Ambas variantes encontraron un reducto (diferente en cada variante) común lo que facilita la aplicación de un algoritmo de clasificación en contexto distribuido.

CONCLUSIONES

1. Se estudió el comportamiento de los algoritmos ACO-RST-FSP y su variante en contexto distribuido D.ACO-RST-FSP.

    a. La principal ventaja de este modelo es encontrar varios subconjuntos de rasgos, como soluciones a la misma vez; a diferencia de otros enfoques.

    b. Medir la calidad de un subconjunto dado como solución en un algoritmo de selección de rasgos cuando este es un filtro es aún un problema abierto en la temática.

2. El trabajo de investigación muestra que la relación entre los parámetros beta (β) y q_0 tiene un impacto fuerte en la cantidad y calidad de los subconjuntos generados por el algoritmo ACO-RST-FS. Se hizo una recomendación de la mejor combinación.

3. Se compararon los resultados del algoritmo ACO-RST-FSP con respecto a otros métodos de selección de rasgos.

4. Se aplicaron los algoritmos estudiados a dos problemas reales realizando una reducción considerable en la dimensionalidad y repercusión en los resultados al aplicarle un método de clasificación.

RECOMENDACIONES

1. Estudiar la influencia del parámetro γ en el algoritmo D.ACO-RST-FSP. Y determinar si es conveniente utilizar valores desagregados γ_i asociados con cada colonia de manera que la *i*-ésima colonia con la mejor solución tenga mayor γ_i.
2. Estudiar los parámetros siguiendo un *método de mallas* para establecer la combinación de valores a asignar.
3. Se propone crear un algoritmo basado en ACO pero utilizar una heurística más informada en el sentido de aprovechamiento de la colaboración en el contexto distribuido.
4. Desarrollar un algoritmo con aprendizaje no supervisado con aplicación en el problema de meteorología, ya que el mismo en su forma original no presenta clases "duras".

REFERENCIAS BIBLIOGRAFICAS

AHN, B. S. (2000) *The integrated methodology of rough set theory and artificial neural networks for business failure predictions*

AL-ANI, A. (2005) Feature subset selection using ant colony optimization. *Int. Journal of Computational Intelligence,* 2**,** 53-58.

ARCO, L., BELLO, R. & GARCÍA, M. (2006) On clustering validity measures and the Rough Set Theory. *5th Mexican International Conference on Artificial Intelligence.* IEEE Computer Society Press

ARÉVALO, L. I. (2006) Línea de Investigación.

BAZAN, J., SON, N. H., SKOWRON, A. & SZCZUKA, M. (2003) A View on Rough Set Concept Approximations. IN 26392003, L. (Ed.) *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. 9th International Conference, RSFDGRC2003.* Chongqing, China.

BELL, D. & GUAN, J. (1998) Computational methods for rough classification and discovery *ASIS,* 5**,** 403-414.

BELLO, R. & NOWÉ, A. (2005) A Model based on Ant Colony System and Rough Set Theory to Feature Selection. *Genetic and Evolutionary Computation Conference (GECCO05).* Washington DC, USA.

BLAKE, C. L. & MERZ, C. J. (1998) UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html.

BULLNHEIMER, B., HARTL, R. F. & STRAUSS, C. (1999) A new rank-based version of the Ant System: A computational study. *Central European Journal for Operations Research and Economics,* 7**,** 25-38.

CABALLERO, Y. (2005) Uso de los Conjuntos Aproximados para el tratamiento de los datos. Santa Clara, Cuba, Universidad Central de Las Villas.

CABALLERO, Y. & BELLO, R. (2006a) Two new feature selection algorithms with Rough Sets Theory. IN M. BRAMER, E. (Ed.) *Artificial Intelligence in Theory and Practice.* Springer Boston.

CABALLERO, Y. & BELLO, R. (2006b) Two new feature selection algorithms with Rough Sets Theory. IN BRAMER, M. (Ed.) *Artificial Intelligence in Theory and Practice.* Springer Boston.

CARLIN, U. S. (1998) Rough set analysis of medical datasets and A case of patient with suspected acute appendicitis *ECAI 98 Workshop on Intelligent data analysis in medicine and pharmacology.*

CHOUBEY, S. K. (1996) A comparison of feature selection algorithms in the context of rough classifiers. *Fifth IEEE International Conference on Fuzzy Systems.*

CHOUCHOULAS, A. (1999) A Rough Set Approach to Text Classification.

DASH, M. & LIU, H. (1997) Feature selection for classification. *Intelligent Data Analysis,* 1**,** 131-156.

DORIGO, M., BIRATTARI, M. & STUTZLE, T. (2006) Ant colony optimization. *Computational Intelligence,* 1**,** 28-39.

DORIGO, M. & BLUM, C. (2005) Ant colony optimization theory: a survey. *Theory and Computer Science,* 344**,** 243-278.

DORIGO, M. & CARO, G. D. (1999) The Ant Colony Optimization meta-heuristic. IN CORNE, D., DORIGO, M. & GLOVER, F. (Eds.) *New Ideas in Optimization.* London UK, McGraw-Hill.

DORIGO, M., CARO, G. D. & GAMBARDELLA, L. M. (1999) Ant algorithms for discrete optimization. *Artificial Life,* 5**,** 137-172.

DORIGO, M. & GAMBARDELLA, L. M. (1997a) Ant colonies for the travelling salesman problem. *Biosystems,* 43**,** 73-81.

DORIGO, M. & GAMBARDELLA, L. M. (1997b) Ant Colony System: A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation,* 1**,** 53-66.

DORIGO, M., MANIEZZO, V. & COLORNI, A. (1996) The Ant System: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems Man, and Cybernetics,* Part B 26**,** 29-41.

DORIGO, M. & STUTZLE, T. (2003) The ant colony optimization metaheuristic algorithms, applications, and advances. IN F.GLOVER & KOCHENBERGER, G. A. (Eds.) *Handbook of Metaheuristics.* Kluwer.

DORIGO, M. & STUTZLE, T. (2004a) *Ant Colony Optimization,* Cambridge, MA, MIT Press.

DORIGO, M. & STUTZLE, T. (2004b) *Ant Colony Optimization.*, MIT Press.

DUDA, R. O., HART, P. E. & STORK, D. G. (2001) *Pattern Classification (2nd Edition)*, Wiley-Interscience.

FAYYAD, U., PIATESKY-SHAPIRO, G., SMYTH, P. & UTHURUSAMY, R. (1996) *Advance in Knowledge Discovery and Data Mining,* Cambridge,Mass, MIT Press.

FIRPI, H. & GOODMAN, E. (2004) Swarmed feature selection. *Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop (AIPR 2004).*

GÓMEZ, Y., BELLO, R. & NOWÉ, A. (2009) Two Step Swarm Intelligence to Solve the Feature Selection Problem. *Journal of Universal Computer Science.,* Vol. 14**,** pp. 2582-2596.

GRECO, S. & INUIGUCHI, M. (2003) Rough Sets and Gradual Decision Rules. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. *9th International Conference, RSFDGRC2003.* Chongqing, China.

HEINONEN, J. & PETTERSSON, F. (2007) Hybrid ant colony optimization and visibility studies applied to a job-shop scheduling problem *Applied Mathematics and Computation,* 187**,** 989-998.

HOLM, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics,* 6**,** 65- 70.

HUAN, L. & HIROSHI, M. (2007) *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*, Chapman & Hall/CRC.

JENSEN, R. & SHEN, Q. (2005) Fuzzy-Rough Data Reduction with Ant Colony Optimization. *Fuzzy Set and System.*

JEON, G. & JEONG, J. (2006) Designing a video deinterlacing system based on Rough Set attributes reduction. *The International Symposium on Fuzzy and Rough Sets. ISFUROS2006.* Santa Clara, Cuba.

KADOUS, M. W. & SAMMUT, C. (2005) Classification of Multivariate Time Series and Structured Data Using Constructive Induction. *Machine Learning,* 58**,** 179–216.

KARGUPTA, H. (2003) Distributed Data Mining: Algorithms, systems and applications.

KOHAVI, R. & FRASCA, B. (1994) Useful Feature Subsets and Rough Set Reducts. *Third International Workshop on Rough Sets and Soft Computing.*

KOHAVI, R. & JOHN, G. H. (1997) Wrappers for Feature Subset Selection. *Artificial Intelligence,* 97**,** 273-324.

KOMOROWSKI, J. & PAWLAK, Z. (1999) Rough Sets: A tutorial. *Rough Fuzzy Hybridization: A new trend in decision-making. Springer***,** 3-98.

KORZES, M. & JAROSZEWICZ, S. (2005) Finding reducts without building the discernibility matrix. *5th International Conference on Intelligent Systems Design and Applications (ISDA'05).* Wroclaw, Poland, IEEE Computer Society.

KUDO, M. & SKLANSKY, J. (2000) Comparison of algorithms that select features for pattern classifiers. . *Pattern Recognition Letters,* 33**,** 25-41.

KUO, T.-F. & YAJIMA, Y. (2003) Approximate Reducts of an Information System. IN 26392003, L. (Ed.) *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. 9th International Conference, RSFDGRC2003.* Chongqing, China.

LAZO, M., RUIZ, J. & ALBA, E. (2001) An overview of the evolution of the concept of testor. *Pattern Recognition***,** 753-762.

LIU, H. & YU, L. (2005) Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. On knowledge and data enginnering.,* 17**,** 1-12.

LIU, Y. X., XIONG, J. & SUN, B. M. (2008) Research on dynamic scheduling of job-shop production with the ant colony optimal algorithm. *Applied Mechanics and Materials,* 10**,** 109-113.

LUNACEK, M., WHITLEY, D. & KNIGHT, J. N. (2005) Measuring mobility and the performance of global search algorithms. *Genetic and Evolutionary Computation Conference (GECCO 2005)***,** 1209-1216.

MI, J.-S., WU, W.-Z. & ZHANG, W.-X. (2003) Approaches to Approximation Reducts in Inconsistent Decision Tables. IN 26392003, L. (Ed.) *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing 9th International Conference, RSFDGRC2003.* Chongqing, China.

MITCHELL, T. (1997) *Machine Learning*, McGraw Hill.

PAL, S. K. & SKOWRON, A. (1999) *Rough Fuzzy Hybridization: A New Trend in Decision-Making*.

PAWLAK, Z. (1982) Rough sets. *International Journal of Information & Computer Sciences* 11**,** 341-356.

PIÑERO, P., ARCO, L., GARCÍA, M. M. & CABALLERO, Y. (2003) Two New Metrics for Feature Selection in Pattern Recognition. *Lectures Notes in computer Science (LNCS 2905).* Springer-Verlag

RUIZ, R., AGUILAR-RUIZ, J. & RIQUELME, J. (2004) Wrapper for ranking feature selection. *5th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'04).* Springer Verlag.

SANGÜESA, R. & MOLINA, L. (2000) *Data Mining, una introducción*, Ediciones UOC.

SANTIESTEBAN, Y. & PONS, A. (2003) LEX: un nuevo algoritmo para el cálculo de los testores típicos. *Revista Ciencias Matemáticas,* 21.

SKOWRON, A. (1999) New directions in Rough Sets, Data Mining, and Granular Soft Computing. *7th International Workshop (RSFDGRC'99), Yamaguchi, Japan.* Lecture Notes in Artificial Intelligence 1711.

STÜTZLE, T. & HOOS, H. (2000) MAX-MIN Ant System. *Future Generation Computer Systems,* 16**,** 889-914.

SUGIHARA, K. & TANAKA, H. (2006) Rough Sets approach to information systems with interval decision values in evaluation problems. *The International Symposium on Fuzzy and Rough Sets. ISFUROS2006.* Santa Clara, Cuba.

TSUMOTO, S. (2003) Automated extraction of hierarchical decision rules from clinical databases using rough set model. *Expert systems with Applications,* 24**,** 189-197.

WANG, X. (2007) Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters,* 28**,** 459-471.

YVAN SAEYS, IÑAKI INZA & LARRAÑAGA, P. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatic,* 23**,** 2507-2517.

ZHONG, N., DONG, J. & OHSUGA, S. (2001) Using Rough sets with heuristics for feature selection. *Journal of Intelligent Information Systems,* 16**,** 199-214.