RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
# Reconocimiento de Patrones

## Reconocimiento de patrones: conceptos y metodología

**José Ruiz-Shulcloper,**
**Jesús Ariel Carrasco-Ochoa**
**y José Francisco Martínez-Trinidad**

**RT_054** **octubre 2013**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
**www.cenatav.co.cu**

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
**Reconocimiento de Patrones**

# Reconocimiento de patrones: conceptos y metodología

José Ruiz-Shulcloper,
Jesús Ariel Carrasco-Ochoa y
José Francisco Martínez-Trinidad

**RT_054** **octubre 2013**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
**www.cenatav.co.cu**

# Reconocimiento de patrones: conceptos y metodología

José Ruiz-Shulcloper[1], Jesús Ariel Carrasco-Ochoa[2] y José Francisco Martínez-Trinidad[2]

[1] Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV),
La Habana, Cuba
jshulcloper@cenatav.co.cu
[2] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Puebla, México
{ariel, fmartine}@ccc.inaoep.mx

RT_054, Serie Azul, CENATAV
Aceptado: 18 de septiembre de 2013

**Resumen.** En el trabajo se exponen los conceptos básicos del Reconocimiento de Patrones y una conceptualización de lo que se podría entender por esta disciplina científica que tiene un fuerte carácter interdisciplinario y aplicado. También se evidencian los estrechos vínculos de esta disciplina con la denominada Minería de Datos. Se expone la necesidad de partir de una metodología que rija el proceso de modelación matemática de problemas reales de reconocimiento de patrones, en su espectro más amplio, para lograr una efectiva introducción de sus resultados en la práctica profesional de diferentes disciplinas no matemáticas. A partir de esta necesidad se propone una metodología para la modelación matemática de problemas de reconocimiento de patrones, que pudiera ser extendida a otras áreas del conocimiento.

**Palabras clave:** metodología, modelación matemática, reconocimiento de patrones.

**Abstract.** In this paper are exposed the basic concepts of Pattern Recognition and a conceptualization that what could be understood by this scientific discipline, which has a strong interdisciplinary character. Also, it is showed the tight links of this discipline with the so called Data Mining. It is exposed the necessity of a methodology that governs the mathematical modeling process of real world problems of pattern recognition, in its widespread sense, in order to obtain an effective introduction of its results in the professional practice of different no-mathematical disciplines. Starting from this requirement, it is proposed a methodology for the mathematical modeling of pattern recognition problems, which could be extended to others knowledge areas.

**Keywords:** methodology, mathematical modeling, pattern recognition.

## 1 Introducción

Aunque reconocer patrones es una actividad intrínseca de muchos seres vivos, no es sino hasta finales de los años 50 que el Reconocimiento de Patrones se empieza a conformar como una disciplina científica. La publicación del libro “Principles of Neurodynamics” de Rosenblatt en 1962, puede considerarse como el inicio de una disciplina que aún está en plena formación y que tiene como bases fundamentales a la Matemática, la Computación y las Ingenierías.

El surgimiento de las computadoras y con ellas las ideas relacionadas con la sustitución de algunas actividades humanas con estos dispositivos, hizo que de manera natural se abordarán, entre otros, una serie de proyectos científicos tales como la construcción de autómatas lectores (dispositivos capaces de leer textos impresos). De estos proyectos surgió el *Perceptrón*, la primera de un conjunto de herramientas para enfrentar estas tareas, al cual siguieron muchos otros intentos análogos.

Es importante señalar que desde sus inicios el Reconocimiento de Patrones estuvo estrechamente vinculado a la identificación, al reconocimiento y la clasificación de imágenes. No es por ello extraño que, en muchas ocasiones, se identifique esta disciplina con todo lo relacionado exclusivamente con el procesamiento y análisis de imágenes.

Por otro lado, a partir de la década de los noventas, se observa un explosivo crecimiento de la capacidad para generar, recolectar y almacenar datos. Sin embargo, la accesibilidad a grandes volúmenes de datos, que en muchas ocasiones es un gran problema, no es el único ni el más complejo. Lo que se convierte en el mayor problema es el poder procesarlos e interpretarlos. Con este propósito se empieza a desarrollar en esos años una nueva área de investigación: la Minería de Datos, que se puede identificar como una parte del proceso de Descubrimiento de Conocimiento a partir de Datos. En esta disciplina una de las tareas más importantes es la clasificación de los datos. Esto hace que tenga un nexo ineludible con el Reconocimiento de Patrones. Y recíprocamente, reconocer patrones es una actividad que depende, tanto en humanos como en los dispositivos computacionales, del análisis y el procesamiento de datos. Luego procesar y analizar datos es un factor común entre estas dos disciplinas, pero hay muchos más.

Según nuestra concepción, el Reconocimiento de Patrones y la Minería de Datos abarcan una cantidad mucho mayor de problemas que el reconocimiento de imágenes y señales. Cuando se hable de Reconocimiento de Patrones o Minería de Datos se estará hablando de investigaciones que tienen que ver con descripciones de objetos (de naturaleza física o abstracta) las cuales deben ser clasificadas o cuyos elementos constituyentes deben ser analizados [1-3].

En muchas disciplinas (Medicina, Sociología, Geociencias, Criminología, y otras) se presentan problemas de clasificación, de diagnóstico, de pronóstico, de determinación de factores de influencias y otros, que son claramente problemas en los que los objetos de estudio no son necesariamente imágenes ni señales [4-6]. Por ejemplo, el diagnóstico diferencial de enfermedades a partir de los signos y síntomas del paciente, el pronóstico de magnitudes máximas de terremotos, el pronóstico de perspectividad de zonas geológicas respecto a un cierto mineral, la determinación de consumidores anómalos de energía eléctrica, y también la determinación de los factores de riesgo de una enfermedad o de las variables socio-económicas que más influyen en el surgimiento de la delincuencia, el perfil de un criminal, el modus operandi de una acción que se repite, las preferencias de los consumidores de un cierto mercado o de un grupo de turistas, entre muchos otros. Estos son algunos de los problemas que aparecen en estas disciplinas, que son denominadas en la literatura como ciencias poco formalizadas (*soft sciences*) y que de manera directa tienen una gran influencia en la vida cotidiana y en la economía.

Hay que subrayar que los problemas antes mencionados son problemas de reconocimiento de patrones y poseen un apreciable nivel de complejidad pues involucran, entre otras cosas, elementos de subjetividad, modelos desarrollados por especialistas de otras áreas del conocimiento y no necesariamente leyes universales, aunque éstas también están presentes.

Es común que en problemas de reconocimiento de patrones los objetos se representen en términos de un conjunto de variables (rasgos, propiedades, características) o a partir de descripciones sintáctico-estructurales de los mismos (gramáticas, grafos).

En este contexto, es también común la suposición de que el conjunto de valores de las variables, en términos de los cuales se describen los objetos, pertenecen todos al conjunto de los números reales, o exclusivamente, a los valores de una lógica, casi siempre asumida como bivalente, es decir, la Lógica Matemática Clásica. Aunque hay algunos trabajos en los que, bajo esas mismas condiciones, se emplea una lógica k-valuada o una lógica difusa (*fuzzy logic*). Sin embargo, en la práctica se encuentran problemas, como los mencionados anteriormente, en los que las descripciones de los objetos están dadas en términos de variables numéricas y no numéricas, simultáneamente. Además, en ocasiones para

algunos objetos existen variables cuyos valores se desconocen (*missing values*), es decir, las descripciones pueden ser incompletas.

Procesar datos mezclados (numéricos y no numéricos) e incompletos es un problema planteado hace mucho tiempo, y existen diferentes grupos de investigadores que lo han abordado. Entre los intentos que tratan de dar una solución al problema, se pueden mencionar los trabajos de G.S. Sebestyen, J.C. Gower; L. Goldfarb; H. Ralanbomdrainy; E. Diday; R.S. Michalski; Yu.I. Zhuravlev; y V. Valev, entre otros. Sin embargo, no es difícil encontrar en la literatura soluciones en las que se pretenden plantear estos problemas de reconocimiento de patrones con datos mezclados e incompletos en términos sólo de números que en la realidad no lo son, por no poseer ni las herramientas, ni la metodología de la modelación matemática adecuadas para resolverlos. En el mejor de los casos, se han ideado formas para eludir el tener que trabajar con datos numéricos y no numéricos de manera simultánea.

# 2 Algunos conceptos básicos

En la literatura se pueden encontrar diferentes intentos de dar una conceptualización de Reconocimiento de Patrones. Algunas de éstas tienen un carácter intuitivo [7,8], otras están basadas en la Estadística o la Teoría de la Probabilidad [9], o en la Teoría de los Lenguajes Formales [10,11], otras en operadores algebraicos [12,13] y otras en la Teoría de los Subconjuntos Difusos [14].

En esta sección se expondrán dos aproximaciones de la conceptualización de Reconocimiento de Patrones partiendo, una de ellas, de las ideas básicas de la Teoría de Conjuntos y la otra, del concepto de procesamiento de datos. Ambas aproximaciones se complementan y conforman una unidad que permite comprender con más claridad la metodología que aquí se exponen.

## 2.1 Enfoque conjuntual

En la Teoría de Conjuntos existen tres conceptos primarios (que no se definen): conjunto, elemento y pertenencia.

Existen dos formas de determinar un conjunto: A) por extensión (característico de conjuntos finitos) listando los elementos del conjunto; B) por intención (característico de conjuntos infinitos) indicando la propiedad que deben satisfacer los elementos del conjunto.

Consecuentemente existen dos formas de saber si un elemento pertenece o no a un conjunto: A) si se tiene el listado de sus elementos componentes, se comprueba si el elemento aparece en el mismo; B) si se conoce la propiedad, se verifica si ésta se cumple o no.

Sin embargo, en muchas ocasiones no se tiene el listado completo de los elementos de un conjunto ni se conoce la propiedad que caracteriza al conjunto de manera unívoca o se conoce la propiedad de una manera poco precisa como por ejemplo las personas muy altas, las personas sanas. Puede ocurrir que exista una formulación que presuponga determinados conocimientos previos de la persona que tiene que decidir en cuanto a la pertenencia de un elemento a un conjunto dado.

Por ejemplo: si un verso pertenece al conjunto de versos de Gabriela Mistral, para ello es necesario saber quién es Gabriela Mistral y cuáles son sus poemas.

En estas situaciones, para decidir acerca de la relación de pertenencia de un elemento con un conjunto dado, se debe recurrir al parecido o similaridad del elemento con los ya conocidos de ese conjunto, o al grado de cumplimiento de las propiedades que caracterizan al conjunto.

Considere el problema del diagnóstico médico, sobre la base de la información almacenada en las historias clínicas. Se tiene una sucesión de pacientes $P_1,...,P_{m_1}$ que padecen la enfermedad $E_1$ y los siguientes $P_{m_1+1},...,P_{m_2}$ padecen la enfermedad $E_2$ y así sucesivamente se tiene información de las

enfermedades $E_1,...,E_r$; sin descartar la posibilidad de que un mismo paciente padezca más de una de las enfermedades.

No se puede dar un listado completo de las personas que padecen o padecerán cada una de las enfermedades (caso extensional), en muchos casos, tampoco, se tiene una buena propiedad (o una sucesión de ellas) que caracterice a cada enfermedad. En este caso puede no resultar fácil, a partir del análisis de los datos, descubrir tales propiedades. En estas condiciones aparece el problema de decidir la o las enfermedades que padece un nuevo paciente (uno que antes no se había diagnosticado). Es claro que, en casos como éste, se tiene que recurrir de nuevo al parecido o similaridad entre el cuadro sintomatológico (conjunto de síntomas y signos) del paciente y los ya estudiados anteriormente.

De manera similar, en las Geociencias (Geología, Geofísica y otras) aparecen problemas análogos para determinar la posible existencia o no de un yacimiento de recursos minerales para su explotación industrial; hacer mapas de pronósticos de magnitudes máximas de terremotos; encontrar conjuntos más pequeños de propiedades que permitan hacer los pronósticos a un costo más bajo; y otras muchas investigaciones del mismo tipo.

También en las Ciencias Agrícolas, en la Sociología, en el análisis de las encuestas, etc., surgen problemas similares cuando se tienen que analizar, por ejemplo, los terrenos con el propósito de conocer cuántas y cuáles son los tipos de áreas perspectivas para ciertos cultivos, o determinar factores que promueven la criminalidad, o conocer las tendencias en las opiniones acerca de un tema en particular, entre muchos otros problemas de este mismo tipo, los cuales aparecen principalmente en las llamadas *ciencias poco formalizadas* (Medicina, Geociencias, Sociología, Psicología, Ecología, y otras) o *soft sciences* en la literatura en idioma inglés.

En la solución de esta clase de problemas se han utilizado técnicas de la Estadística, la Teoría de la Probabilidad, la Teoría de Conjuntos Difusos, Funciones Potenciales, Lógica Matemática Clásica y Polivalente, Lingüística Matemática, Teoría Combinatoria, Teoría de Grafos, Ecuaciones Diferenciales y otros modelos de la Matemática. Cada una de estas técnicas y modelos ha tenido ventajas y desventajas, éxitos y restricciones.

No existe un modelo omnipotente para la solución de un problema que esencialmente es el mismo: decidir en cuanto a las relaciones entre **objetos**, entre un objeto y un conjunto dado de objetos o las **variables** adecuadas para encontrar estas relaciones o **descubrir regularidades** y **propiedades implícitas** en los mismos.

Por otra parte, también es posible tener un universo de objetos y justamente lo que se quiere averiguar es qué tipos de objetos se encuentran en ese universo, es decir, cómo se estructura el universo en clases.

Sobre la base de las formas de determinación de los conjuntos, el problema consistirá en saber si un elemento debe o no estar en el mismo conjunto que otros del universo.

## 2.2 Enfoque de procesamiento de datos

El significado del reconocimiento de patrones también se puede concebir como un cierto procesamiento de datos que tiene un interés particular: *se quiere reconocer, clasificar, estructurar, caracterizar, establecer un diagnóstico, un pronóstico, una génesis, los factores que inciden en un fenómeno u objeto, etc*.

Desde este punto de vista, el **Reconocimiento de Patrones es** una ciencia con un fuerte carácter aplicado e interdisciplinario. Está relacionado con procesos (ingenieriles, físicos, matemáticos y computacionales) de datos (entendidos en una concepción general como se verá más adelante) que provienen de descripciones de objetos (fotos, hologramas, escrituras, jeroglíficos, símbolos, señales bioeléctricas, acústicas, pacientes, zonas geológicas, etc.) con el propósito de obtener (por medio de dispositivos computacionales y/o seres humanos) información que permita establecer las propiedades de ciertos subconjuntos de objetos y/o las relaciones entre ellos. También conocer de las propiedades que poseen las variables en términos de las cuales estos objetos son representados. Estas propiedades constituyen el soporte del posible conocimiento que de estos datos podemos extraer.

El Reconocimiento de Patrones (RP) es una ciencia interdisciplinaria, como se expresaba en la conceptualización anterior, cuyas fuentes integrantes son, esencialmente, las Ciencias Técnicas, la Informática, la Computación y la Matemática, entre otras. Su estructura interna ha estado históricamente fraccionada en diferentes áreas de estudio como el Procesamiento de Imágenes, el Procesamiento de Señales, el Análisis e Interpretación de Imágenes, el Análisis e Interpretación de Señales, la Visión por Computadora, la Percepción Remota, las Redes Neuronales para RP, los Algoritmos Genéticos en RP, las Técnicas de Inteligencia Artificial para RP, la Geometría Descriptiva para RP, la Morfología Matemática para RP, el Reconocimiento Estadístico, el Reconocimiento Sintáctico Estructural, el Reconocimiento Lógico Combinatorio, por mencionar algunas.

No obstante esta diversidad de áreas de investigación, consideramos que todas ellas forman parte de una misma disciplina. Al igual que ocurre en la Matemática, la Física y muchas otras disciplinas con estructuras complejas como la del Reconocimiento de Patrones, donde un especialista en Ecuaciones Diferenciales es un matemático; un especialista en Lógica Matemática también es un matemático, o un especialista en Estado Sólido es un físico; un especialista en Física Atómica es un físico también. Análogamente, un especialista en Procesamiento de Imágenes o de Señales, en Visión por Computadora o Percepción Remota es un especialista en Reconocimiento de Patrones.

El campo de trabajo que se tiene en la disciplina de Reconocimiento de Patrones es inconmensurable. Tanto desde el punto de vista de las investigaciones teóricas como de las aplicaciones. Independientemente de que existen zonas de investigación en Reconocimiento de Patrones donde la Ingeniería (genéricamente hablando), la Computación o la Matemática tienen un peso mayoritario, hay muchas zonas de investigación teórica y aplicada que tienen un carácter híbrido.

Es claro que no hemos dado una definición formal de Reconocimiento de Patrones. Y no lo haremos. Existen varios intentos de definiciones pero consideramos que es muy temprano para intentar encerrar en pocas palabras una disciplina que sólo tiene poco más de cincuenta años de *nacida*, lo que en términos de formación de una ciencia es relativamente muy poco tiempo. No es raro que algunos especialistas en Computación, Matemática o de las Ingenierías consideren a esta zona del conocimiento como parte de la Computación, o de la llamada Inteligencia Artificial, o una rama de la Ingeniería Eléctrica, o de la llamada Cibernética Matemática, etc. Internacionalmente no nos hemos puesto de acuerdo al respecto. De hecho tampoco hay un acuerdo unánime en cuanto al contenido de esta naciente disciplina, es decir, de las áreas que abarca. También se habla de la Visión por Computadora, del Procesamiento de Imágenes y Señales, y de la Percepción Remota como disciplinas independientes del Reconocimiento de Patrones. Nosotros no coincidimos con esta idea.

Es nuestro criterio e implícitamente también el de la Asociación Internacional para el Reconocimiento de Patrones (IAPR), que bajo el nombre de Reconocimiento de Patrones debemos considerar todas las manifestaciones antes citadas como elementos integrantes de un todo. Tampoco consideramos al Reconocimiento de Patrones como una parte de otra disciplina científica sino como una *naciente disciplina*.

Pero mientras nos ponemos de acuerdo y para que esto ocurra lo más pronto posible, debemos de trabajar en el desarrollo de esta zona del conocimiento y después nos ocuparemos de las denominaciones más adecuadas o convenientes. De lo que el lector puede estar seguro es que al no cerrarnos en nuestra área específica de trabajo, al interactuar de manera cooperativa con esas otras zonas de trabajo, nuestras posibilidades de desarrollo crecen apreciablemente, como lo está demostrando la historia de esta aún joven disciplina.

## 2.3 Conceptos intuitivos

En Reconocimiento de Patrones, en general, son básicos los conceptos de objeto, patrón, universo de objetos, variables, representación o descripción de objetos, espacio de representación, clases, reconocimiento, y muchos sinónimos de los anteriores.

Desde un punto de vista intuitivo, **objeto** ha sido un término empleado en diferentes disciplinas para referirse a los entes sujetos a estudio. Ejemplos de objetos pueden ser considerados un paciente de una enfermedad, una zona geográfica, un dispositivo eléctrico, un conjunto de personas, una bioseñal, una fotografía, y muchos otros.

**Patrón** es sinónimo de objeto. En algunas ocasiones es conveniente establecer diferencias entre los objetos acerca de los cuales se conocen ciertas propiedades (por ejemplo, la pertenencia a una clase, tipo, etc.) de otros que no se sabe nada acerca de dichas propiedades. En esos casos suele usarse el término "objeto" para aquellos de los que se desconocen estas propiedades y se les llama "patrones" a aquellos de los que estas propiedades son conocidas.

Las imágenes, sin lugar a dudas, pueden ser tratadas como patrones y se pueden representar de diferentes maneras, todas ellas **datos**, y constituyen objetos de investigación del Reconocimiento de Patrones y de la Minería de Datos, pero son sólo uno de los patrones que tienen esa misma característica. También las señales, como la voz, las explosiones, el sonido en general, los electrocardiogramas, electroencefalogramas, etcétera, son patrones. Pero son asimismo patrones: un hecho delictivo, un fenómeno atmosférico, un texto, una opinión acerca de algo, y muchos ejemplos más.

**Variable** (atributo, rasgo, característica, propiedad) es un factor a tener en cuenta en el estudio de los objetos. De hecho constituye la vía real para poder estudiar, procesar, analizar a los objetos. Es decir, se tiene que estudiar a los objetos a través de su descripción expresada en términos de un conjunto de variables. En ciertos problemas prácticos estas variables constituyen el objetivo central del estudio: su relevancia informacional, su influencia en un fenómeno dado, entre otros, son cuestiones que se requieren conocer por sí mismas y en otras ocasiones son necesarias para ser empleadas en el proceso de reconocimiento.

Consideraremos tres tipos fundamentales de variables: *numéricas* o *cuantitativas* (valores de un cierto dominio numérico), *no numéricas* o *cualitativas* (códigos, valores veritativos de un cierto cálculo proposicional) y *primitivas de un alfabeto* o *grafos*. Estos grupos de variables han determinado el surgimiento de diferentes enfoques en el Reconocimiento de Patrones.

**Representación o descripción de un objeto**. El estudio de los objetos se hace a través de sus variables. La existencia de los mencionados grupos de variables ha dado lugar a diferentes formas de representación de los objetos. A estas formas de representación han estado asociadas herramientas de diferentes áreas de la Matemática, las Ciencias de la Computación, la Física y las Ciencias Técnicas y por ende, han surgido distintos enfoques del Reconocimiento de Patrones.

Tradicionalmente se han considerado sólo dos formas básicas de representación de los objetos en el marco del Reconocimiento de Patrones: la representación en términos de un cierto alfabeto de partes (primitivas) o grafos de los objetos, característico del enfoque sintáctico estructural, y la representación en términos de un conjunto de variables cuantitativas o cualitativas a las que han estado asociados los restantes enfoques.

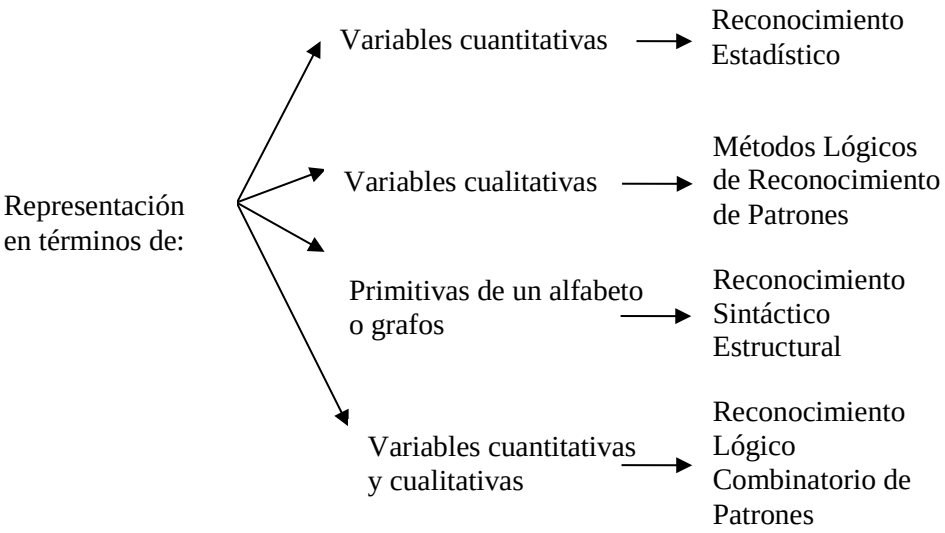


**Fig. 1.** Formas de representación de los objetos y principales enfoques asociados.

Por otra parte, históricamente se ha venido produciendo una separación muy marcada entre la forma en que se abordan los problemas que son descritos en términos de mediciones (léase variables cuantitativas) y los que no pueden ser descritos de manera exclusiva con mediciones. Los primeros han sido abordados desde la óptica del enfoque estadístico del Reconocimiento de Patrones como ya mencionamos anteriormente. Los últimos no han sido tratados de forma adecuada, según nuestro criterio. Sobre este aspecto se abunda en [15-18].

Por **universo** de objetos se entiende el conjunto de todos los posibles objetos admisibles para los propósitos del estudio en cada caso particular. Es por ello un concepto relativo a la investigación que se esté realizando. En general, aquí este concepto se manejará de manera análoga a como se hace en la Teoría de Conjuntos.

**Clase** es un conjunto de objetos que constituyen un tipo, una categoría, en el contexto de los objetos del universo. Siempre será un subconjunto propio[1] del universo de objetos. El término "conjunto" supone implícitamente el que todos los objetos de la clase satisfacen una cierta propiedad (lo que en Teoría de Conjuntos se denomina, la determinación intencional del conjunto).

Es importante subrayar la posibilidad que la conceptualización propuesta abre al uso de otros modelos de teorías de conjuntos, como es el caso de la Teoría de los Subconjuntos Difusos, también llamados "borrosos" (*Fuzzy Set Theory*) o de la de los Conjuntos Rugosos, también llamados "aproximados" (*Rough Set Theory*). Estos conceptos y herramientas constituyen medios importantes en el proceso de modelación matemática de problemas prácticos.

[1] Estrictamente incluido.

**Reconocimiento** es un proceso por el cual se determina la pertenencia de un objeto a un cierto conjunto de objetos (clase). En ocasiones estas clases son conocidas, se tienen muestras de objetos que pertenecen respectivamente a cada una de ellas o las propiedades que las caracterizan. En ocasiones no se tienen muestras o propiedades de todas estas clases aunque se sabe que existen y en otras ocasiones no se sabe ni cuántas clases pueden existir en el universo de objetos en cuestión, sólo se dispone de una muestra de los objetos. Reconocimiento es también el proceso para determinar la importancia informacional de una variable, un objeto o un conjunto de ellos, o las propiedades que en ellos pudieran estar implícitas.

## 2.4 Problemas fundamentales de reconocimiento de patrones

Por problemas fundamentales[2] de reconocimiento de patrones en este trabajo se entienden todos aquellos relacionados con la clasificación de objetos y fenómenos y con la determinación de los factores que inciden en los mismos. Estos problemas también aparecen en la Minería de Datos y en las aplicaciones de estos modelos y herramientas a problemas en ciencias poco formalizadas. Así, se consideran cuatro familias de problemas que se denominan respectivamente:

1.- selección de variables y objetos;
2.- clasificación supervisada;
3.- clasificación no supervisada;
4.- clasificación parcialmente supervisada.

A continuación se describen, a grandes rasgos, estas cuatro familias de problemas a partir de la Teoría Clásica de Conjuntos.

### *2.4.1 Selección de variables y objetos*

La selección de variables es uno de los pasos fundamentales en cualquier problema de clasificación debido a que la mayoría de los problemas de reconocimiento de patrones están basados en la descripción de los objetos en términos de un conjunto de variables. ¿Cuáles variables tomar en cuenta? ¿Cuáles de ellos son mejores para ciertos propósitos? ¿Cuál es la importancia informacional de cada uno de ellos y de algunos de sus subconjuntos? ¿Cuál es la forma en que deben compararse los valores de dichas variables? También la selección de variables tiene un interés en sí mismo, al ser un procedimiento que permite conocer con mayor profundidad las diferentes clases de objetos en términos de los cuales son descritos.

Relacionado con esto, en la literatura se identifican dos problemas diferentes pero muy vinculados: la selección de variables para la clasificación y la selección de variables para la descripción [19].

En el primer grupo de problemas lo que se enfrenta es la determinación del mejor subconjunto de variables para la clasificación de nuevos objetos (no clasificados). Esto conlleva la reducción del conjunto de todas las posibles variables (reducción de la dimensionalidad) sobre la base de las diferencias que estas variables presentan en cuanto a mejor reconocer o clasificar a los nuevos objetos y otros problemas de optimización adicionales del subconjunto de variables a emplear, que también se deben tener en cuenta en muchos problemas prácticos, como la eficiencia computacional entre otros. Además, no todas las variables son igualmente importantes para la clasificación de un nuevo objeto. Éste es uno de los aspectos que también se incluyen en la solución de la selección de variables para la clasificación.

Los problemas de selección de variables para la descripción son muy frecuentes en las ciencias poco formalizadas. Este problema conlleva la determinación de un subconjunto de variables que de una mejor manera caracteriza a los objetos de cada una de las clases. Es claro que estos problemas suponen

2 Aquí no abordaremos problemas tales como la captación, el preprocesamiento de los datos y otros problemas que también son de interés para el Reconocimiento de Patrones.

la existencia previa de las clases o en su defecto, la repercusión que estas variables tuvieron en la formación de las mismas. Problemas como la determinación de síndromes, factores de riesgos, perfiles de usuarios, características discriminantes de conjuntos de objetos o fenómenos, y muchos más, son instancias de este tipo de problemas prácticos.

De igual manera, no todos los objetos son igualmente importantes por lo cual también se debe hacer una selección de los mismos. La selección adecuada de las variables y los objetos constituye un problema esencial en la eficacia y la eficiencia de la solución de muchos problemas prácticos.

### *2.4.2 Clasificación supervisada*

Este es el problema más conocido del Reconocimiento de Patrones pues mucha de la actividad humana está vinculada, directa o indirectamente, con procesos de clasificación supervisada. Este problema también aparece de manera natural en las ciencias poco formalizadas y es tratado formalmente en primera instancia por la Estadística. Sin embargo, no ha sido la única rama de la Matemática que aborda la temática, de hecho en esta obra no se hace uso de la misma como herramienta fundamental sino de la Matemática Discreta. El problema de la clasificación supervisada aparece también en otras disciplinas muy relacionadas con el Reconocimiento de Patrones, como por ejemplo en el Aprendizaje Computacional.

Dado un universo de objetos y el conocimiento acerca de la existencia de ciertas clases y una muestra de objetos que pertenecen a cada una de ellas (conjunto de entrenamiento), o un conjunto de propiedades (reglas) que caracterizan a cada una de ellas, el problema consiste en determinar para objetos no clasificados las relaciones de pertenencia de los mismos con cada una de las clases.

Existen muchos ejemplos de este tipo de problema: el diagnóstico médico y técnico, el pronóstico de fenómenos, la identificación de señales, de sonidos, de imágenes y muchas más, son casos particulares de clasificación supervisada. Además, quizás la mayoría de los procesos de aprendizaje estén basados en un cierto tipo de problema de clasificación supervisada.

### *2.4.3 Clasificación no supervisada*

El problema de la clasificación no supervisada (*clustering*, *structuralization*) es una de las tareas más importantes en las ciencias naturales y sociales. La creación de Taxonomías es parte del trabajo científico en muchas de las áreas del conocimiento y es una de las herramientas necesarias en muchas de las investigaciones en la actualidad.

El problema principal a resolver es encontrar las relaciones entre los objetos de un universo en términos de sus características (rasgos, variables). Estas relaciones se establecen sobre la base del concepto de analogía (similaridad). Quizás uno de los conceptos más importantes en el Reconocimiento de Patrones.

El propósito es agrupar los objetos según su analogía (parecido, semejanza, cercanía si se está hablando de un espacio de representación con distancia definida). En este sentido se pueden encontrar dos situaciones diferentes, a saber, el número de grupos (*clusters*) es conocido (o impuesto) previamente o no.

En el primer caso, el problema lo denominamos *clasificación no supervisada restringida* (agrupamiento o estructuración restringida) y en el segundo, *clasificación no supervisada libre* (agrupamiento o estructuración libre).

Todos los procesos taxonómicos, en cualquier ciencia, son instancias de este tipo de problemas. La clasificación no supervisada es la piedra angular en la mayoría de los problemas de Minería de Datos y de extracción de conocimiento.

En general, la descripción de los grupos de objetos, es decir, su determinación intencional (conceptual), es también un problema común en ambos tipos de problemas de clasificación no supervisada.

### *2.4.4 Clasificación parcialmente supervisada*

Ésta es quizás la familia de problemas de reconocimiento de patrones en la que menos estudio se ha realizado. A pesar de la existencia de muchos problemas reales donde aparecen situaciones de este tipo, la clasificación parcialmente supervisada no ha recibido igual atención que las restantes familias de problemas en Reconocimiento de Patrones.

El problema consiste en una combinación de los problemas de clasificación anteriormente descritos. En el universo de objetos dado, se conoce la existencia de ciertas clases e incluso se tienen muestras de algunas de ellas (o propiedades que las caracterizan) pero no de todas. El problema central es clasificar nuevos objetos en estas circunstancias, en las que no se tienen muestras de todas las clases y en las que incluso pudieran existir clases que se desconocen.

Por ejemplo, se sabe que en una habitación hay un grupo de personas conversando. Se conocen las voces de algunas de ellas pero no de todas. Se puede incluso no conocer cuántas personas se encuentran en la habitación. El problema pudiera ser determinar qué dijo cada persona durante la conversación.

Un problema análogo aparece cuando se tiene una descripción de una cierta clase de objetos, por ejemplo de “buenos pilotos” y se busca en un conjunto de candidatos, un grupo de potenciales buenos pilotos, basado en el historial de los buenos pilotos estudiados. Aquí el problema principal es que resulta infrecuente el que se estudien los “malos pilotos”, estos se desechan y no se guardan sus características. Como se puede apreciar no se está ante un problema ni de clasificación supervisada, pues se tiene una muestra de objetos de algunas de las clases pero no de todas, ni tampoco ante un problema de clasificación no supervisada por la misma razón.

## 2.5 ¿Por qué una metodología?

Para resolver los problemas reales que el Reconocimiento de Patrones y la Minería de Datos abordan, se considera necesaria una metodología de la modelación matemática de dichos problemas. Es decir, resolver primero cómo deben abordarse estos problemas.

Como se mencionó anteriormente, uno de los objetivos centrales de esta rama del conocimiento es el de crear herramientas para la solución de problemas prácticos, en particular, en zonas poco formalizadas del conocimiento. Sin embargo, desde las primeras experiencias en este campo, que se expondrán más adelante, se llegó al convencimiento de que las soluciones a los problemas tenían que ser fruto de un proceso más sólido y fundamentado que el de *prueba y error* pues, entre otras cosas, en muchos problemas reales el especialista del área de la aplicación no tiene aún los conocimientos para juzgar si la solución propuesta es correcta o no. Lo más que podría afirmar es si coincide o no con sus expectativas. Por otro lado, lo que le daría sentido a la aplicación de modelos matemáticos y herramientas computacionales en áreas del conocimiento no matemático, es que puedan aportar elementos que le permitan a esos especialistas descubrir nuevos conocimientos.

Pudiera pensarse que el trabajo de los especialistas (no matemáticos) fuera sólo plantear el problema y lo demás es cuestión de los matemáticos y las computadoras y en ese universo de cajas negras ellos no tienen nada que buscar.

Por su parte, pudiera pensarse también que el trabajo de los modeladores del problema (matemáticos, informáticos u otros en general) comienza sólo a partir del momento en el que es formulado el problema matemático a resolver.

Con cierta razón pudiera decirse que los modeladores no se pueden convertir en médicos, geólogos, criminólogos, etc., para resolver problemas de modelación matemática y aplicación de técnicas computacionales en esas áreas del conocimiento; y que los especialistas de esas áreas no se van a convertir en matemáticos y computistas para resolver sus problemas.

Ante esta *Torre de Babel*, las soluciones que se fueron adoptando en una primera etapa por diferentes investigadores, incluidos nosotros mismos, no nos satisfacían. No nos parecía correcto el hecho resultante que, con modelación matemática o sin ella, los especialistas de las mencionadas áreas de aplicación hacían sus diagnósticos y pronósticos como lo tenían concebido antes de los resultados del sistema que se elaboraba al efecto.

¿Qué sentido tenía entonces el trabajo en la solución de muchos de esos problemas? [3]

¿En qué medida resultaba útil el procedimiento que se estaba siguiendo para la solución de problemas tales como el diagnóstico diferencial de enfermedades, el pronóstico de perspectividad de ciertos minerales, el pronóstico de ocurrencia de ciertos fenómenos naturales, del tipo de fenómeno, etcétera?

Si el modelador parte de un modelo matemático dado a priori para la solución de un problema práctico cuya esencia no le es accesible y si el especialista de esa área no conoce qué se hace con sus datos y por qué, ¿se puede esperar acaso que el especialista no matemático confíe en respuestas que no son las que él espera?, ¿pensará por un momento en esos casos que el que puede tener algún error de concepción, dato, información, es él y no la caja negra de la cual no tiene la menor idea de cómo fue construida?, ¿lo haría usted?.

No sería difícil encontrar en la literatura reciente, trabajos en los que problemas de reconocimiento de patrones son interpretados para su solución, por ejemplo, en términos de espacios normados. Sin embargo, en esos problemas carecen de sentido operaciones como la suma o el producto.

Con mucha frecuencia ocurre que no se verifican, en el problema que se quiere resolver, los presupuestos de las técnicas que se aplican, sin embargo se pretende que sean aceptadas por el especialista.

Resulta inmediato que en todo proceso de modelación matemática de problemas de Reconocimiento de Patrones existen además otros modelos no matemáticos que hay que tener en cuenta. Para eso, no hay que convertirse en científico de cada una de las ramas en las que se hacen aplicaciones, ni estos respectivos especialistas se tienen que convertir en matemáticos o especialistas en computación. Pero sí resulta imprescindible conocer la esencia del problema en términos del modelo del especialista no matemático y para éste es inevitable entender el porqué de las decisiones que se van tomando en el proceso de formalización matemática de la información por él aportada y el proceso que se aplicará a sus datos para garantizar no desvirtuar la realidad.

Dado que cada ciencia tiene sus especificidades, se expone a continuación una metodología de la modelación matemática de problemas de Reconocimiento de Patrones que se considera aplicable a cualquier zona del conocimiento poco formalizada.

## 2.6 Metodología para modelación matemática de problemas de reconocimiento de patrones

La metodología que se presenta en esta sección está basada sobre los siguientes principios:

- No se puede modelar lo que no se conoce.
- Los especialistas no matemáticos no tienen que dominar el lenguaje de la Matemática ni de la Computación.
- El problema no matemático se formula estrictamente en el lenguaje de la ciencia en particular.
- No se va a experimentar con los datos para ver si la respuesta conviene.

[3] Hay que aclarar que existe una gran cantidad de problemas prácticos para los cuales son posibles soluciones de este tipo, es decir, los sistemas realizan una serie de pasos que de antemano el especialista sabe lo que va a resultar de ellos y la computadora viene a auxiliarle en el sentido de la velocidad de procesamiento de los datos, del manejo de grandes volúmenes de información, etc.

- La frecuencia, o repetición, está en la base del conocimiento en las ciencias poco formalizadas.
- La analogía es uno de los instrumentos fundamentales para la adquisición del conocimiento en las ciencias poco formalizadas.
- Debe haber una comprensión mutua de los elementos esenciales del problema real formulado y de las herramientas que se emplearán para el procesamiento de los datos que emanen del proceso de modelación.
- Discutir el modelo no-matemático es una necesidad del matemático o especialista en computación.
- Discutir el modelo matemático ayuda a que la solución que se alcance se aplique.
- Al modelo matemático se llega, no se parte de él.
- Se pretende construir un sistema computacional que constituya una herramienta más del trabajo del especialista (no matemático). Lo que se denomina *sistema herramienta*.
- El sistema herramienta no sustituye al especialista, lo potencia en su trabajo rutinario y en sus investigaciones.
- Se cree en las cajas negras cuando nunca se equivocan, pero esto no siempre ocurre.
- Para manejar un sistema herramienta no se necesita saber Computación, tampoco Matemática.
- El sistema debe aspirar a la automatización del proceso de modelación matemática, según las concepciones metodológicas propuestas.
- La modelación matemática de un fenómeno sólo puede acometerse con un equipo multidisciplinario.
- La regla fundamental de un equipo multidisciplinario debe ser la honestidad.
- Se debe tener un lenguaje común en el equipo multidisciplinario.
- Se deben definir obligaciones y funciones específicas para cada miembro del equipo, sin menoscabo de la responsabilidad colectiva.

En la literatura se encuentran varios trabajos que le conceden importancia a la modelación matemática de los problemas de reconocimiento de patrones y en los que se afirman que muchos de los problemas del insuficiente éxito de las herramientas matemáticas y computacionales en la solución de problemas prácticos radican justamente en cuestiones de carácter metodológico [20-22].

La metodología que proponemos se muestra esquemáticamente en la Figura 2.

Se debe subrayar que lo que a continuación se analizará no es un conjunto de recetas que se deban memorizar y aplicar mecánicamente. Lo que se expondrá es una manera de pensar, una actitud ante la solución de problemas prácticos de Reconocimiento de Patrones. Como tal se debe ver.

El objetivo fundamental de esta metodología es servir de base sólida para la elaboración de los conceptos, modelos y herramientas necesarias para enfrentar problemas de reconocimiento de patrones de la realidad, sin cometer el usual error de transformarla, desvirtuarla, adecuarla a lo que le conviene a un modelo matemático en particular, sino a la inversa: *partir de modelar la realidad y encontrar el modelo matemático adecuado para enfrentarla*.

Si, por ejemplo, se quiere estudiar el movimiento de una gallina en una jaula y se presupone que es un cilindro recto de base circular y se aplican las herramientas matemáticas más avanzadas, de seguro no se obtendrá la modelación del movimiento de la gallina en la jaula. A partir de esto, las recomendaciones a los granjeros en cuanto a las dimensiones y características de las jaulas no deberían ser muy confiables. Aunque algunos podrían argumentar en su defensa que el mejor modelo de la realidad es la realidad misma y con ella no es posible trabajar y que de lo que se trata es de dar una aproximación al problema de modo tal que se puedan encontrar soluciones razonables en la práctica.

Ante esta disyuntiva, optamos por partir de algunos principios básicos y con ellos construir una metodología que permita llegar a soluciones más fundamentadas, más creíbles, más apegadas a la

realidad misma. Una de ellas es que *del modelo matemático no se parte, al mismo se llega después de un proceso de modelación de la realidad*.

En consonancia con estas ideas se aborda uno de los problemas conceptuales de mayor relevancia en el Reconocimiento de Patrones: el concepto de analogía, (semejanza, similaridad) entre objetos y entre valores de una variable. Gran parte del reconocimiento de patrones gira en torno a este concepto, de una u otra manera. A pesar de ser quizás el concepto más importante de la disciplina, el mismo ha sido subestimado al reducirlo al inverso o el opuesto de una función de distancia y por ende, resolver todos los problemas en espacios métricos. Lo que además no es cierto, ya que existen muchas semejanzas que no son el inverso o el opuesto de una función de distancia.

Convenientemente, la Matemática para estos espacios en los que hay definida una distancia, tiene una gran cantidad de herramientas con las que se pueden dar soluciones a los problemas que en la realidad satisfacen los requisitos que estas herramientas exigen, pero, como se verá más adelante, esto conlleva en muchas ocasiones a las violaciones de principios metodológicos básicos y con ello al que tales soluciones a problemas reales no sean adecuadas. Con frecuencia se cometen violaciones muy fuertes al pretender plantear problemas de reconocimiento de patrones a partir de descripciones de los objetos en términos de supuestas mediciones que en la realidad no lo son, por no poseer ni las herramientas ni la metodología de la modelación matemática adecuadas para resolverlos. En el mejor de los casos se han ideado formas para eludir el tener que trabajar con datos cuantitativos y no cuantitativos de manera simultánea. Esta es la motivación central de este trabajo y del surgimiento de dos líneas de investigación: el Reconocimiento Lógico Combinatorio de Patrones y como consecuencia de ésta, la Minería de Datos Mezclados e Incompletos (se trata, en estos casos, de hacer Reconocimiento de Patrones y Minería de Datos a partir de descripciones de objetos con datos numéricos, no numéricos y ausencia de valores en los mismos).

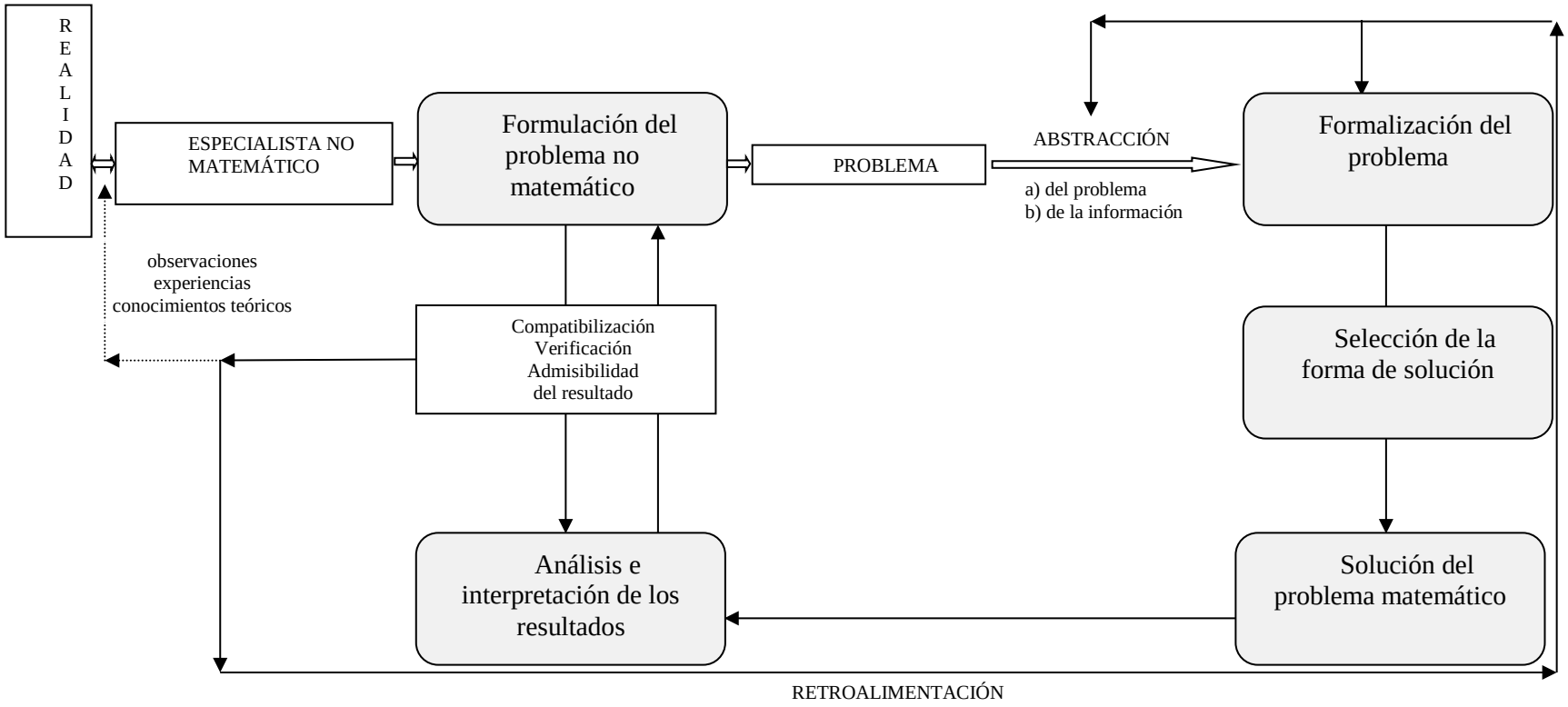


**Fig. 2.** Esquema de la metodología propuesta.

## 2.7 ¿Cómo modelar problemas de reconocimiento de patrones?

Como se mencionó anteriormente, para Tou y González [7] desde hace más de 35 años, el análisis y diseño de los sistemas de reconocimiento de patrones está relacionado con el análisis de datos y el procesamiento de la información. Con esta afirmación coincidimos totalmente, pero ¿qué se debe entender en nuestros días por análisis de datos y procesamiento de información?

Ante todo, qué son los datos. Por dato se puede entender un número, un nombre, una dirección, una cualidad. También se debe entender por datos o conjuntos de datos, dependiendo del problema en cuestión, una imagen, una foto, un símbolo, un jeroglífico, una señal de radio, un electrocardiograma, un documento, un libro, una función, una matriz, un tensor, etc. Esto es en el sentido de que esos objetos son descritos en términos de valores de ciertas variables, es decir, en términos de datos. Luego cuando se habla de procesamiento y análisis de imágenes y señales se está hablando de casos particulares de procesamiento y análisis de datos, por lo que es necesario tener en consideración esas particularidades.

Identificar un rostro en una imagen o seleccionar un pez en una banda de transportación, convertir la voz en texto impreso o establecer un diagnóstico médico a partir de la lectura de un electrocardiograma (ECG) o sólo a partir la entrevista con un paciente, son problemas intrínsecamente relacionados con datos en los cuales se tienen propósitos específicos. Esos propósitos determinan la forma en que se deben procesar los datos. Esto implica que cualquier proceso de datos es precedido por un proceso de modelación del problema que se necesita resolver.

Luego, para resolver un problema, en particular un problema de reconocimiento de patrones, se necesita ante todo modelar el problema y después procesar los datos (ver Figura 3). En muchos casos la solución final del problema es un programa computacional que el usuario emplea para resolver el problema en cuestión. En otros, se hace necesaria la construcción de un dispositivo en el cual un programa computacional trabaja. Es decir, que las etapas para la solución de un problema de reconocimiento de patrones serían: *modelación*, *procesamiento de datos* y *solución*.

En la práctica casi siempre se tienen datos que necesitan de un proceso previo para poder extraer la información que se requiere. En el caso particular del Reconocimiento de Patrones, aunque no hay una frontera clara en el proceso a partir de los datos en bruto hasta las conclusiones finales, un paradigma útil es considerar el Reconocimiento de Patrones dividido en cuatro tipos de procesos de datos: adquisición, preprocesamiento, representación/descripción y análisis.

La **etapa de adquisición** es el primer paso para procesar datos. De hecho, la forma en que se adquieran los datos es la primera transformación que se les está haciendo, es decir, la forma en que sean adquiridos los datos afectará todo el proceso posterior. Esta etapa se caracteriza por el hecho que la entrada son los datos originales tomados de las fuentes originales y la salida son los datos en bruto a partir de los cuales se debe extraer la información que se busca. Debe observarse que en este proceso se tiene una fuente, por ejemplo un electrocardiógrafo, a partir de la cual se toma una señal, el ECG del paciente. Esta señal electrocardiográfica es por lo general ruidosa, no es lo suficientemente clara o limpia, debido a las capacidades del equipo de muestreo, por lo que no es fácil de leer por un cardiólogo y extraer la información que se busca.

La **etapa de preprocesamiento** es un proceso de datos caracterizado por el hecho de que tanto la entrada como la salida son datos de la misma naturaleza y significan lo mismo. Por ejemplo, ambas son señales, imágenes, jeroglíficos, matrices, tuplas de valores de ciertas variables, etc. Filtrar las imágenes o las señales, incrementar la resolución o el contraste de las mismas, restaurar una imagen, eliminar ruidos, ajustar los valores de las variables, validar los datos, escalarlos, son ejemplos de procesamiento de datos. Observe que en la entrada de esta etapa se tiene por ejemplo una señal, el ECG de un paciente, y en la salida casi la misma señal, el mismo ECG pero quizás sin ruidos, en el cual sea menos complejo leer la información que se requiere.

La **etapa de representación/descripción** es el proceso en el cual los datos originales son transformados en una forma nueva, adecuada para el proceso posterior. Esta etapa está caracterizada por el hecho de que la entrada es diferente a la salida, al menos en su significado. Es un proceso por el

que se describen los objetos involucrados en el problema. La segmentación (particionar una imagen o una señal en regiones), la selección de variables, la representación por ondeletes (wavelets) de una imagen, son sólo ejemplos de representación/descripción de los datos. Una señal ECG por ejemplo, puede ser descrita en términos de ciertos complejos de segmentos de la misma: PQR, RS, T, y otros. Se sabe que a partir de ellos se puede establecer que el segmento PQR es normal, pero que el RS no lo es o que la T está invertida en la segunda derivación, etc. Es ese caso, la salida es la secuencia de atributos de la mencionada señal.

Finalmente, la **etapa de análisis** es un proceso en el cual se encuentra el significado de los datos originales o al menos de una parte de ellos. Se puede reconocer la ocurrencia de cierta información previamente almacenada y se puede tomar una decisión, llegar a una conclusión. Problemas de decisión, interpretación, caracterización, clasificación, reconocimiento, son ejemplos de análisis de datos. En el caso de la señal ECG, si se tienen suficientes conocimientos médicos, se puede determinar la normalidad de un paciente desde el punto de vista de su sistema cardiovascular. En el caso de una foto, se pueden detectar personas con un cierto vestuario, distinguir sus rostros e incluso si se tienen conocimientos previos se puede identificar a estas personas.

**La metodología que se propone consta en general de cinco etapas resumidas de la siguiente manera:**

1) Formulación del problema no matemático, es decir, el problema que se quiere resolver.
2) Formalización del problema, es decir, creación del problema matemático.
3) Selección de la forma de solución del problema.
4) Solución del problema matemático.
5) Análisis e interpretación de los resultados, respecto al problema no matemático original que se quiere resolver.

Cada una de las etapas antes mencionadas se divide en subetapas, las cuales ayudan a definir los objetivos que deben cumplirse en ellas. A continuación se describen brevemente estas subetapas.

**1) Formulación del problema no matemático**

En esta etapa, el especialista (o los especialistas) del área de aplicación tiene una mayor participación porque es quien expresa en su lenguaje el problema a resolver. Esta etapa incluye las siguientes subetapas:

1.1) Formulación de los objetivos, es decir, qué es lo que se espera resolver, por ejemplo, decidir si un paciente tiene o no cierta enfermedad.
1.2) Determinación de las hipótesis y presuposiciones del modelo del especialista no matemático, es decir, qué es lo que se va a suponer, por ejemplo, que todos los pacientes tienen problemas cardiacos y se quiere determinar si son propensos a un infarto.
1.3) Determinación de los objetos, propiedades, clases de objetos, es decir, cuáles van a ser los objetos de estudio y sus propiedades, por ejemplo, los objetos van a ser pacientes médicos y se le van a medir tales o cuales síntomas y signos, y lo que se quiere determinar es si tienen una determinada enfermedad.
1.4) Descripción de la información inicial y la forma en que se obtiene, es decir, cuál es la información con la que se va a trabajar y de dónde se obtuvo, por ejemplo, definir con cuántos pacientes se trabajará, cuántos son propensos a ataques cardiacos, y de dónde se obtuvieron sus síntomas y signos.
1.5) Formulación de los resultados esperados y criterios para su evaluación, es decir, qué es lo que se espera que el sistema responda y cómo se evaluará. Por ejemplo, se quiere que el sistema responda si los pacientes son, o no, propensos a ataques cardiacos, y se evaluará aplicándolo sobre un grupo de pacientes, para los cuales ya se sabe la respuesta.
1.6) Además, se debe determinar en esta etapa qué información es relevante, si esto se conoce, cómo se recolecta la información, cómo se interpreta y manipula la información, cómo se requiere que se presenten los resultados, la identificación de ruidos y distorsiones de la

información, la valoración de los errores de la información en su entrada, procesamiento y salida.

Es obvio que en esta etapa, el papel principal lo desempeña el especialista (pudieran ser varios) del área de aplicación. Sin embargo, nada tendría sentido si el papel de los modeladores (matemáticos, ingenieros, computistas) es pasivo. Se trata, por el contrario, de cuestionar, de entender la esencia del fenómeno que nos deben explicar, si bien en el lenguaje del especialista del área, sin supuestos acomodos para que entiendan y con el mayor rigor de esas ciencias en particular, pero con la intención de alcanzar un verdadero diálogo, en el que las ideas esenciales subyacentes al problema que se investiga se vean con precisión. Y para lograrlo no hay que cursar la especialidad de Geología, Geofísica, Cardiología, Criminología u otra. Esta afirmación vuelve a ser avalada por la experiencia práctica que se ha tenido en la realización de estos trabajos.

**2) Formalización del problema: creación del problema matemático**

Esta etapa es posible que se lleve a cabo a medida que el especialista formula el problema. Es compleja porque se requiere traducir del lenguaje del especialista al lenguaje formal de la Matemática, de tal manera que de la etapa anterior queden reflejados los objetivos, objetos, propiedades y su escala de medición, características, relaciones entre objetos y entre propiedades, el concepto de clase de objetos, propiedades de las mismas, los conceptos de analogía, la evaluación de los errores, etc. En esta etapa se realiza:

2.1) Formalización de los objetivos, es decir, tomar los objetivos formulados y formalizarlos, por ejemplo, el sistema computacional debe tomar los datos de un grupo de pacientes y ser capaz de determinar si otro paciente tiene o no cierta enfermedad.

2.2) Determinación de los objetos matemáticos, variables, clases de objetos, etc., es decir, determinar la forma en que se representarán los objetos de estudio, por ejemplo, cada paciente se representará con tal conjunto de síntomas y signos, los cuales serán representados por variables de tal o cual tipo, y se tendrán tales clases de objetos.

Hay una tendencia de escoger espacios matemáticamente conocidos, como los Euclidianos, métricos, etc. Esto se debe a que la realidad se ve desde la óptica del matemático que modela el problema. Se debe trasladar el problema al lenguaje de la Matemática lo más fielmente posible, sin lastrarlo con presuposiciones matemáticas, sin obligar a que las cosas sean como le hace falta a tal o cual modelo. Se necesita de un análisis imparcial, una amplia cultura matemática (o un equipo de matemáticos) y mucha honestidad científica.

2.3) Determinación de las propiedades de las variables, objetos y clases que representen el modelo del especialista no matemático, es decir, determinar cómo se manejarán las variables que se usan, por ejemplo, tales o cuales síntomas se compararán de tal manera, y los pacientes entre sí se compararán de tal o cual manera, las clases serán o no disjuntas, etc.

2.4) Formación del conjunto de datos, es decir, la obtención de todos los datos que se usarán, por ejemplo, tomar los datos de pacientes y escribirlos en una matriz de descripciones.

2.5) Análisis de la concordancia entre los resultados esperados y los objetivos, es decir, verificar que los resultados que se esperan satisfagan los objetivos.

2.6) Formalización de los criterios para la evaluación de los resultados, es decir, determinar cómo se evaluará el sistema computacional, por ejemplo, se le alimentarán datos de control y se verificarán los resultados.

Es común dar siempre la misma importancia a todas las propiedades que describen a los objetos, y suponer que son uniformes tanto los tipos de variables que los representan, sus dominios de definición y los criterios de comparación de los valores de cada una de ellas. Sin embargo, no es difícil hallar ejemplos en la práctica en lo que esto no es así. Para un médico es obvio que, al establecer un diagnóstico diferencial de cardiopatías, no tienen la misma importancia que el paciente tenga la presión

arterial alta a que tenga 46 años, o que sea un hombre o una mujer, blanco o negro, grueso o delgado. Si bien todos esos factores influyen en la aparición de cardiopatías, lo hacen de manera diferente.

Por otra parte, modelar todos estos factores de manera uniforme, por ejemplo con variables Booleanas, no siempre es admisible desde el punto de vista del especialista. Con el sexo, por ejemplo, no se tendrían problemas; con la raza ya empiezan a complicarse las cosas en la práctica, los factores genéticos influyen en estos problemas de salud y no es cierto que haya sólo dos grupos raciales; sin embargo, ya con la presión arterial las cosas cambian. La presión puede ser alta de muchas maneras, algunas de las cuales pueden ser peligrosas incluso para la vida del paciente.

Estos aspectos influyen en la búsqueda de la solución y en la selección de algoritmos para resolver el problema en cuestión, y determinan en gran medida la forma en que serán procesados los datos originales a partir de su organización en lo que se denomina conjunto de entrenamiento.

**3) Selección de la forma de solución del problema**

En esta etapa se tienen las siguientes sub-etapas:

3.1) Análisis de los datos, con el objetivo de seleccionar el aparato matemático para la solución del problema matemático, es decir, determinación de las características del problema, por ejemplo, cuántas clases se tienen, cuántas variables y de qué tipo, etc.

3.2) Planteamiento del problema matemático y determinación de la herramienta para la solución, es decir, con base en las características del problema determinar qué herramienta se usará para resolverlo.

3.3) Elaboración del esquema de procesamiento de la información, es decir, determinar cómo tomará los datos el sistema computacional, dónde dará los resultados, etc., por ejemplo, tomará los datos de un archivo, dará los resultados en pantalla, etc.

El proceso de formalización muchas veces restringe fuertemente el área de búsqueda de las técnicas de solución. En esta etapa un papel decisivo lo desempeña el análisis del conjunto de entrenamiento para, entre otros aspectos, detectar errores cometidos en su formación, analizar la calidad de los datos, variabilidad de los datos para detectar objetos anómalos, la posible necesidad de cambio de escala, el posible cambio de codificación.

Se debe tener en cuenta que: los modelos matemáticos tienen su área de aplicabilidad donde resultan confiables; no es fácil señalar el área de aplicación de un modelo; los modelos son consistentes, los errores se dan por usarlos donde no se debe; cualquier herramienta que se use dará una información de salida para una cierta información de entrada; se debe evaluar el tratamiento de la ausencia de información.

En esta etapa se puede reducir la cantidad de información requerida, al mismo tiempo que se aumenta su calidad. Se decide el enfoque o combinación de ellos para la solución del problema matemático, determinando la familia de algoritmos a la que pertenece. No existe un procedimiento unívoco para garantizar la realización de una adecuada selección. Se pueden ir eliminando los modelos revisando los requisitos básicos de cada uno e ir eliminando aquellos que incumplen los supuestos a los que se llegaron en el proceso de formalización del problema planteado.

Sobre la base del análisis realizado, se determina el o los enfoques para la solución del problema, por lo general esto significa la determinación de la(s) familia(s) de algoritmos que se usarán. Luego, se procede a seleccionar el algoritmo para la solución del problema matemático, que cumpla ciertas condiciones de optimalidad respecto a los criterios establecidos.

Esta etapa concluye con la elección del modo de solución que se aplicará y si es el caso, el esquema de procesamiento de la información.

**4) Solución del problema matemático**

Esta etapa consta de las siguientes sub-etapas:

4.1) Diseño del esquema para el procesamiento de la información y la solución del problema, esto es, diseñar el sistema computacional de solución.

4.2) Implementación del esquema de solución, es decir, codificar los algoritmos de solución, o generar el sistema computacional con base en algoritmos ya programados.
4.3) Análisis formal de los resultados, desde el punto de vista de los criterios de evaluación formulados, es decir, aplicar la forma de evaluación determinada.

El sistema computacional se elabora (si lo amerita el caso) teniendo en cuenta los datos formalizados y el tipo de algoritmo a utilizar, y se obtiene la solución matemática del problema matemático. Se analiza la concordancia del resultado matemático alcanzado con los objetivos del problema matemático, teniendo como herramienta fundamental la formalización de los criterios para la evaluación de resultados de la segunda etapa.

**5) Análisis e interpretación de los resultados, respecto al problema**

Durante esta etapa se realiza lo siguiente:

5.1) Concordancia de los resultados con el modelo del especialista no matemático, es decir, decidir si el sistema resuelve o no el problema.
5.2) Toma de decisiones respecto a las acciones a llevar a cabo en torno a la solución, es decir, decidir si se va a modificar el sistema, se analizará el modelo del especialista no matemático, se probará con otros datos, etc.

Los resultados matemáticos se interpretan traduciendo del lenguaje matemático al lenguaje del especialista, en forma similar a lo que se hizo en su contraparte en la segunda etapa. Después de la correspondencia del resultado matemático con el problema matemático en la etapa anterior, se requiere un análisis de la interpretación del resultado matemático en términos del problema a resolver. Las acciones resolutivas obtenidas son variadas y dependen de los resultados de dicho análisis.

El especialista del área de aplicación también es el máximo responsable de esta etapa la cual debe ser ejecutada en conjunto con los elementos del equipo multidisciplinario.

Es necesario decir que cada zona del conocimiento tiene sus peculiaridades y éstas añaden etapas y le introducen algunas modificaciones a las que hemos mencionado. Por ejemplo, en las Geociencias en una etapa anterior a la formulación del problema, los geocientíficos ubican la formulación del modelo geólogo-geofísico sobre el cual se hará la modelación. Y esto es así porque estos modelos pueden variar de un especialista a otro e incluso de una zona de estudio a otra. Sin embargo, en la Medicina no ocurre de esta manera, el modelo de la tuberculosis, su definición conceptual, no varía de un paciente a otro (aunque sí sus manifestaciones) ni de un médico a otro (aunque sí pueden valorar de manera diferentes algunos síntomas y/o signos).

En la aplicación de la metodología propuesta, es decir, en el proceso de modelación matemática a la luz de los preceptos aquí expuestos, se pueden encontrar diferentes situaciones. Puede ocurrir que, durante la formulación del problema, del análisis del modelo del especialista surjan deficiencias que lleven al colectivo a la decisión de corregir dicho modelo antes de proseguir. También puede ocurrir que se concluya que el problema planteado no es uno de reconocimiento de patrones y que su solución debe encaminarse por otra vía. A igual conclusión se pudiera llegar si después de terminada la primera etapa, la formalización del problema nos lleva al planteamiento de un problema matemático para el cual no haya un modelo adecuado (metodológicamente).

Por otra parte, no hay que pensar que la aplicación de la metodología siempre nos llevará de manera lineal a la solución definitiva. En ocasiones habrá que regresar a etapas anteriores para reconsiderar algunas de las decisiones tomadas, ya sea para confirmarlas o modificarlas.

Finalmente, se quiere expresar que este proceso, que puede parecerle a algunos engorroso, aburrido o innecesario, ha dado frutos aún antes de procesar los datos. Es opinión, de los médicos y geocientíficos con los que se ha trabajado [23-34], que el proceso de modelación les resultó beneficioso para sus hipótesis y concepciones, para depurar e incluso aumentar la calidad de sus modelos. Esto no significa que no haya mucho aún por hacer en el plano metodológico. Esta metodología para la

modelación matemática de problemas de Reconocimiento de Patrones para ciencias poco formalizadas, es sólo un punto de partida.

### 2.8 ¿Cómo impacta esta metodología en la elaboración y explotación de sistemas computacionales de reconocimiento de patrones?

La elaboración de sistemas computacionales y su explotación por especialistas no-matemáticos, según nuestras concepciones, no está exenta de presupuestos metodológicos, por el contrario, éstos están estrechamente vinculados con los anteriormente descritos para la modelación del problema.

Como perspectiva, de lo que se trata es de automatizar todo el proceso de modelación matemática expuesto, desde el diálogo hasta la selección de una técnica o un modelo de algoritmos. Es obvio que esto es una proyección muy ambiciosa a la que se irá aproximando paulatinamente.

Al utilizar un sistema computacional se tiene la tendencia de usarlo para ver qué se obtiene, sin mediar análisis alguno. Algunas veces se hace esto para confirmar una respuesta esperada. Lo cual se resume en darle datos a un sistema para ver si la respuesta que nos da nos conviene.

Lo que se busca es que el sistema computacional responda a los requerimientos del proceso de modelación matemática descrito anteriormente. Si el sistema no fue elaborado sobre la base de un proceso de modelación matemática, es necesario realizar la modelación matemática y después determinar en qué medida el sistema en cuestión responde a los requerimientos de dicho proceso.

Consideramos que los sistemas computacionales en ciencias poco formalizadas deben cumplir las funciones de:

- resolver el problema concreto
- resolver problemas de investigación para la solución de cuestiones metodológicas y teóricas de la disciplina del problema práctico.

Se han elaborado sistemas computacionales bajo estas normativas, por ejemplo el sistema ALISA [35] elaborado en el Instituto de Recursos Minerales de la ex-URSS y PROGNOSIS [36], elaborado en el Instituto de Cibernética, Matemática y Física de Cuba, actualmente en explotación en el Centro de Investigaciones del Petróleo de Cuba.

## 3 Ejemplo de aplicaciones: Determinación de anomalías AGE perspectivas para fosforita

Los trabajos del pronóstico de perspectividad de rocas fosfóricas de génesis sedimentarias a partir de las anomalías AGE (aerogamma-espectrométricas) y los de perspectividad de mineralización de materiales de interés en la esfera nuclear, en particular de fosforitas de génesis sedimentaria fueron desarrollados con el Centro de Estudios de Materiales Básicos de la SEAN (Secretaría de Asuntos Nucleares).

Éste es el clásico problema de clasificación de zonas geológicas respecto a determinados objetivos de búsqueda. Es el caso de la clasificación de objetos geológicos de una zona de estudio en perspectivos y no perspectivos; o en grados de perspectividad en el que se incluyen diferentes tipos de objetos geológicos perspectivos y los no-perspectivos.

Este problema es factible de ser abordado en aquellos casos en que el modelo geológico dispone de suficiente conocimiento de la región de estudio como para afirmar de un conjunto de objetos geológicos que son perspectivos o no-perspectivos.

**1) Formulación del problema no matemático**

1.1) El **objetivo fundamental** del trabajo consiste en elaborar un procedimiento que permita establecer cuándo una anomalía aerogamma-espectrométrica (AGE) está asociada a una zona de perspectividad para fosforitas de génesis sedimentaria.

1.2) El modelo conceptual geólogo-geofísico que se utilizó se fundamenta en la asociación estrecha existente, para la región de estudio, entre el fósforo y los elementos radioactivos naturales en todos los yacimientos y manifestaciones de rocas fosfóricas, así como en las cortezas de intemperismo ferro arcilloso originadas a partir de la destrucción de las rocas carbonato-fosfatizadas.

Concretamente, el modelo conceptual geólogo-geofísico de anomalías AGE perspectivas para fosforitas se conforma de los siguientes **presupuestos**:

1. Por *anomalía* se consideraron todos los segmentos de itinerario AGE donde el registro en el canal de conteo total superaba en tres veces el fondo local para este parámetro. El *itinerario* aquí se refiere al recorrido que realiza el avión portador del equipo de medición en un vuelo a altura constante (75 *m*).

2. Las *anomalías AGE perspectivas* no están asociadas directamente a los afloramientos de las capas de fosforitas granulares y de calcarenitas fosfatizadas, sino a zonas de corteza de intemperismo desarrolladas tanto en extensión superficial como en profundidad.

3. Las anomalías AGE perspectivas están localizadas espacialmente en los flancos de estructuras anticlinales y sinclinales de pendiente suave en condiciones tectónicas relativamente tranquilas.

4. Las anomalías AGE perspectivas están espacialmente asociadas a secuencias carbonato-arcillosas representadas por calizas, calcarenitas y arcillas.

5. Las anomalías AGE perspectivas están asociadas a depósitos del Mioceno Inferior y Medio preferiblemente de este último en el contacto con las rocas del Oligoceno y con las rocas del Eoceno Medio a Superior.

6. En zonas de perspectividad conocida para fosforitas se evidencia una alta densidad de anomalías AGE.

7. Asociado a cada segmento de itinerario AGE **pueden ser considerados los atributos** (variables): *ancho de anomalía* (longitud del segmento anómalo); *valores máximo y de fondo local de la intensidad sumaria de la actividad gamma*; *valores asociados y de fondo local de las concentraciones de los elementos radioactivos naturales*. Como valor asociado de las concentraciones se entienden las magnitudes de estos parámetros en los puntos donde se determina el valor máximo de la intensidad sumaria de la actividad gamma.

Sin embargo, se sabe que no sólo estos atributos geofísicos dan información en torno a la presencia del fósforo, tal y como se plantea en los presupuestos anteriores. En virtud de éstos se consideró[4] además:

8. Asociados a cada segmento de itinerario AGE definir una ventana de un kilómetro cuadrado cuyo centro se hiciese coincidir con la proyección de los epicentros anómalos (valores máximos) representados en el mapa geológico a escala 1:1,000,000.

9. En dichas áreas asociadas a las respectivas anomalías AGE evaluar los atributos: *edad de las rocas asociadas a la anomalía*; *serie genética de los suelos en la que se localiza el epicentro de la anomalía*; *desarrollo del suelo*; *longitud de los contactos perspectivos asociados a la anomalía* y *densidad anómala asociada a cada anomalía*.

1.3) De lo anterior se puede resumir que nuestros **objetos de investigación** serán ventanas de 1 kilómetro de área, centradas en los respectivos centros anómalos representados en el mapa geológico a escala 1:1,000,000.

Estas ventanas serán descritas en términos de 14 atributos, 9 de ellos geofísicos, asociados directamente a las anomalías AGE y 5 geológicos, especificados en el presupuesto 9. Es necesario precisar algunas **cuestiones en torno a dichos atributos**:

1.4) Las fuentes informativas utilizadas fueron: el catálogo de anomalías AGE del sector de Güines; el levantamiento geológico a escala 1:1,000,000 de la provincia Mayabeque; el levantamiento geológico 1:50000 de la región Güines-Pipián; los mapas de suelos a escala 1:50,000 de la región objeto de

[4] Inicialmente se había pensado trabajar con los segmentos de itinerario AGE como objetos de investigación, sin embargo, el incorporar los presupuestos contenidos en 9, nos hicieron cambiar esa formulación.

estudio. Además del catálogo de anomalías AGE, se incluyó la representación en hojas cartográficas a escala 1:50,000 de los segmentos de itinerarios AGE correspondiente a cada anomalía y la proyección de los epicentros anómalos en los mismos. Para todas estas fuentes informativas se conoce el error de medición respectivo y la calidad en general de la información con que se cuenta. Se hizo un análisis de los atributos cuya información aparece en escala 1:50,000, ya que todo el trabajo se decidió hacer en escala 1:100,000, y se concluyó que las deformaciones propias del cambio de escala son insustanciales para los propósitos que perseguimos, por lo que pueden ser utilizadas.

1.5) Lo que nos interesa es tener un **criterio de discriminación de anomalías AGE perspectivas** respecto a las fosforitas de tipo sedimentarias, por lo que es imprescindible que el resultado sea un mapa de las anomalías perspectivas de fosforitas del tipo mencionado en la región de estudio. Este mapa debe confeccionarse a partir de la determinación de la perspectividad de cada ventana asociada a las anomalías AGE (no de todas las cuadrículas de las zonas). **Partimos de la información que existen yacimientos en la región**, en particular, los de Meseta Roja y Loma Candela, **cuyas anomalías y respectivas ventanas servirán de patrones positivos** para el estudio, **en calidad de patrones negativos**, es decir, de descripciones asociadas a zonas no-perspectivas, se escogió un grupo de ventanas en las que, sobre la base de criterios geológicos, resultaba teóricamente imposible la presencia de fosforitas del tipo sedimentarias. Para la evaluación, se utilizará un grupo de ventanas para las cuales se sabe si son o no perspectivas

1.6) Desde el punto de vista geológico no sólo es necesaria la inclusión en esta investigación de atributos diferentes a los geofísicos (mencionados en el presupuesto 7, que son **numéricos**) sino también urge que el análisis de las descripciones de las ventanas asociadas a las anomalías AGE y la toma de decisiones en relación con las mismas se haga esencialmente de forma **cualitativa**. Además, a la hora de hacer comparaciones entre ventanas que sabemos que están asociadas a zonas perspectivas o no-perspectivas, con otras que queremos conocer su perspectividad, **es necesario que dicha comparación se haga sobre la base de la analogía que existe entre las mismas**, definida ésta en términos de los atributos seleccionados.

Se sabe que el trabajo de búsqueda de minerales es costoso por lo que **los errores de discriminación deben minimizarse**. **No son equivalentes los errores** de dar por perspectiva una zona no-perspectiva y dar por no-perspectiva una zona perspectiva. Se prefirió el segundo error al primero y que si se va a cometer algún error y existe la **posibilidad de abstenerse** de tomar una decisión, sería preferible, aunque tampoco es deseada una abstención.

**2) Formalización del problema: creación del problema matemático**

2.1) El sistema computacional debe tomar los datos de un mapa de anomalías y determinar cuáles son perspectivas respecto a la fosforita del tipo sedimentario.

2.2) A cada ventana asociada a las anomalías AGE se le hizo corresponder un tuplo 14-dimensional en el que cada coordenada denota el valor de la correspondiente variable en la ventana. Estas variables[5] y sus descripciones se detallan a continuación.

2.3) Variable No 1- *A*: *ancho de la anomalía aerogamma espectométrica* (AGE).

Variable No 2.- *OC*: *valor máximo de la intensidad sumaria* (*mcr/h*) *de la actividad gamma en la anomalía AGE*.

Variable No 3.- *U*(*Ra*): *valor asociado de la concentración de uranio según radio* (*ppm*) *en la anomalía AGE*.

Variable No 4.- *Th*: *valor asociado de la concentración de torio* (*ppm*) *en la anomalía AGE*.

Variable No 5.- *K*: *valor asociado de la concentración de potacio* (%) *en la anomalía AGE*.

Variable No 6.- *FOC*: *valor de fondo local de la intensidad sumaria* (*mcr/h*) *de la actividad gamma en la anomalía AGE*.

Variable No 7.- *FU*(*Ra*): *valor de fondo local de la concentración de uranio según radio* (*ppm*) *en la anomalía AGE*.

[5] Obsérvese que las denotaciones de las variables se han realizado siguiendo las que utilizan los especialistas de esta área, para facilitar la comprensión por parte de ellos del procesamiento que posteriormente se realizará con las mismas.

Variable No 8.- *FTh*: *valor de fondo local de la concentración de torio* (*ppm*) *en la anomalía AGE*.
Variable No 9.- *FK*: *valor de fondo local de la concentración de potasio*.

Las nueve variables anteriores son variables cuantitativas reales mayores o iguales a cero. Las mismas se obtienen a partir de los registros del espectómetro AGE y por ende son susceptibles de errores de medición. No hay razones para afirmar la existencia de valores umbrales de perspectividad o de intervalos de valores con un comportamiento análogo. Por ello, se considera que el **criterio de comparación de los valores de las variables** anteriores es el del denominado *error admisible*. Las magnitudes de los errores para cada variable son:

| VARIABLE | ERROR |
|---|---|
| *A* | 0.05 *km* |
| *OC* | 0.18 *mcr/h* |
| *U*(*Ra*) | 0.75 *ppm* |
| *Th* | 1.05 *ppm* |
| *K* | 0.25 % |
| *FOC* | 0.18 *mcr/h* |
| *FU*(*Ra*) | 0.75 *ppm* |
| *FTh* | 1.05 *ppm* |
| *FK* | 0.25 % |

Variable 10.- *SITEST*: *Situación Estratigráfica de las rocas asociadas a la ventana*. Al analizar la distribución de las ventanas, consideradas como unidad informacional, en relación con la información geológica disponible se determinó que, como regla, en el área informacional básica (ventana) se manifestaba la presencia de rocas en contacto de una, dos y hasta tres edades. Por esta causa se elaboró una escala nominal que incluía todas las combinaciones de una, dos y tres edades lo que condiciona la presencia de 174 situaciones estratigráficas posibles en las ventanas. Esta escala fue agrupada en seis grupos de edades cuyo comportamiento, en la relación con la perspectividad para fosforita de las ventanas, puede considerarse análogo en cada grupo y diferenciante en relación con el resto de los grupos.

Estos grupos fueron:

Grupo 1.- Rocas de edad Jurásico, Cretácico, Paleógeno, Cuaternario e intrusivos básicos y sus contactos.
Grupo 2.- Oligoceno considerado independientemente o su contacto con rocas incluidas en el primer grupo.
Grupo 4.- Mioceno Inferior considerado independientemente o en contacto con rocas incluidas en el primer y segundo grupo.
Grupo 5.- Mioceno Inferior y Medio en contacto mutuo con rocas incluidas en el primer y segundo grupo.
Grupo 6.- Contacto entre las rocas del Mioceno Inferior y las rocas del Eoceno Medio Superior y Oligoceno.

Variable No 11.- *LONPER*: *Longitud de los contactos perspectivos asociados a cada anomalía*. Esta variable se determinó como la longitud total (en *cm*) de los contactos perspectivos (ver presupuesto No 5) en el área de la ventana. Esta variable es cuantitativa, real y toma solamente valores mayores o iguales a cero. Como criterio de comparación se utilizó el de *error admisible*. Como magnitud del error se tomó 0.1 *cm*.

Variable No 12.- *DENANOM*: *Densidad anómala*. Se determinó a partir del conteo del número de anomalías incluidas en cada ventana de 1 $km^2$. Esta es una variable cuantitativa que toma valores mayores o iguales a cero. Como criterio de comparación se utilizó el de *error admisible* con valor del error igual a cero, o sea, desde el punto de vista de la perspectividad para fosforitas, cada valor de la densidad anómala es una situación diferente.

Variable No 13.- *SERGE*: *Serie genética de suelos* en la que se localiza el epicentro de la anomalía. Toma valores en la escala nominal siguiente:

1: Suelos ferralíticos.
2: Suelos escabrosos sobre rocas básicas.
3: Suelos escabrosos sobre rocas básicas y medias.
4: Suelos escabrosos sobre calizas duras.
5: Suelos escabrosos sobre areniscas calcáreas.
6: Suelos escabrosos sobre cocó y areniscas calcáreas.
7: Suelos fersialíticos.
8: Suelos pardos.
9: Suelos húmico-calcifórmicos.
10: Suelos aluviales.

Desde el punto de vista de la perspectividad para fosforitas se consideran vínculos entre algunos de estos suelos en cuanto a un cierto comportamiento análogo. Sobre la base de estos vínculos, como criterio de comparación se utilizó el denominado *conjuntos*. Como subconjuntos se consideraron:

{1}; {2,3}; {4}; {5,6}; {7}; {8}; {9}; {10}.

Variable No 14.- *DESUEL*: Desarrollo del suelo en el que se localizó el centro de la anomalía. Esta variable toma valores en la escala nominal siguiente:

1: Suelos escabrosos (no hay desarrollo de suelo).
2: Suelos muy poco desarrollados (profundidad 0-0.25 m).
3: Suelos poco desarrollados (0.25-0.50 m).
4: Suelos medianamente desarrollados (0.50-0.90m).
5: Suelos desarrollados (0.90-1.50 m).
6: Suelos muy desarrollados (más de 1.50 m).

Desde el punto de vista de la perspectividad para la fosforita se consideran análogos algunos valores anteriores. Por esta causa se escogió como criterio de comparación el de *conjuntos*. Los subconjuntos considerados fueron:

{1}; {2,3}; {4,5,6}

2.4) Utilizando estas variables y las ventanas seleccionadas, se construyó una matriz de datos que representa a los objetos de estudio.

2.5) Claramente en este caso, los resultados que se obtendrán utilizando el problema matemático formulado, a través de las variables definidas en la etapa 2.3, concuerdan con los objetivos planteados.

2.6) **Criterios para la evaluación**.- Se hace necesario la determinación de un procedimiento para evaluar el comportamiento del algoritmo que vayamos a utilizar en definitivas para la solución del problema y para ello es imprescindible precisar lo que se van a considerar errores del mismo y cómo vamos a "penar" la comisión de tales errores. En este caso esa idea se recogió en la siguiente expresión

$$E(\mathcal{A}) = 3M + 2L + 0.5N$$

donde

*M*: es la cantidad de unidades informacionales no perspectivas consideradas como perspectivas;
*L*: cantidad de perspectivas consideradas no perspectivas y
*N*: cantidad de unidades en las que el algoritmo se abstiene.

La expresión del criterio de evaluación $E(\mathcal{A})$, fue elaborada con el experto del área de aplicación sobre la base de un análisis cualitativo de las implicaciones de estos errores.

## 3) Selección de la forma de solución del problema

Inicialmente debe realizarse un análisis de los tuplos que conforman la matriz inicial con vistas a determinar el grado de informatividad de cada una de las variables consideradas y el grado de ajuste de las descripciones de estas variables y de los criterios de comparación utilizados al modelo geológico establecido.

3.1) La selección de la matriz de entrenamiento se realizó teniendo en cuenta el **peso informacional de los objetos y consideraciones geológicas**.

Como integrantes de la Matriz de Control se utilizaron los objetos (ventanas) no considerados para la Matriz de entrenamiento.

Los **pesos informacionales de las variables** consideradas se muestran en la Tabla 1.

El ordenamiento obtenido de las variables según los valores del peso informacional de los mismos, en términos generales responde al modelo geológico establecido. Solamente se apartan de este modelo las variables:

- *Valor asociado de la concentración de potasio en la anomalía AGE.*
- *Situación estratigráfica de las rocas asociadas a la ventana.*

En el caso de la primera variable el algoritmo reveló una regularidad no tenida en cuenta al conformar la Matriz de Entrenamiento. Las ventanas constituyentes de esta matriz se asocian espacialmente al yacimiento Meseta Roja donde se observa una asociación genética estrecha entre la mineralización fosfórica y glaucorítica lo que motiva que las zonas de desarrollo de cortezas de intemperismo ferro arcillosas, provocadas por la erosión de las menas a las que se asocian las anomalías AGE presentan enriquecimiento en potasio en relación con las anomalías consideradas no perspectivas.

**Tabla 1.**

| VARIABLE | PESO INFORMACIONAL |
|---|---|
| *SERGE* | 0,453 |
| *K* | 0,371 |
| *DENANOM* | 0,364 |
| *U(Ra)* | 0,358 |
| *FOC* | 0,297 |
| *DESVEL* | 0,284 |
| *LONPER* | 0,284 |
| *FTH* | 0,277 |
| *FU(Ra)* | 0,263 |
| *FK* | 0,257 |
| *OC* | 0,250 |
| *Th* | 0,250 |
| *A* | 0,236 |
| *SITSET* | 0,155 |

El bajo peso informacional de la variable, analizado comparativamente con el mayor peso obtenido por el atributo *longitud de los contactos perspectivos* permite concluir que la situación estratigráfica en general no resulta significativa en relación con la perspectividad para fosforitas. Lo realmente significativo es la presencia en la ventana de contactos entre los paquetes estratigráficos que definen la perspectividad fosfórica (deposiciones del Mioceno Inferior y Medio en contacto con rocas del Oligoceno y del Eoceno Medio a Superior).

3.2) En el ejemplo que analizamos estamos en presencia de **un problema de clasificación supervisada con 2 clases disjuntas en el que se admite la abstención de clasificación, los errores de clasificación posibles no tienen el mismo valor, no se admite la multiclasificación por carecer de sentido y en el que se dispone de una muestra de cada una de las clases**.

Las matrices de entrenamiento y de control quedaron conformadas como se indica en la tabla 2:

**Tabla 2.**

| CLASE | *MI* | *M* | *MC* |
|---|---|---|---|
| 2 | 25 | 12 | 13 |
| 1 | 32 | 16 | 16 |

donde *MI*, *M* y *MC* representan la matriz inicial, la de entrenamiento y la de control respectivamente.
3.3) Para resolver el problema se utilizó el sistema PROGNOSIS [36], utilizando los mecanismos de entrada y salida propios de este sistema.

## Solución del problema matemático

4.1) A partir de la selección de los parámetros descritos, se obtuvieron varias variantes de clasificación:

VARIANTE No 1.

Sistema de Conjuntos de Apoyo: Cardinal Fijo. Cardinal = 11.
Función de Similaridad: Un umbral. Umbral = 0.
Umbral de Similaridad: Media máxima.

VARIANTE No 2.

Sistema de Conjuntos de Apoyo: Definido por usuario.
Función de Similaridad: Un umbral. Umbral = 0.
Umbral de Similaridad: Media máxima.

Teniendo en cuenta los conjuntos de apoyo definidos en la $2^{da}$ variante es evidente que la variante No 1 está incluida en ella. Por esta causa se determinó que esta última era la más significativa, la cual será diseña utilizando las facilidades del sistema PROGNOSIS.
4.2) En este caso no fue necesario implementar el esquema de solución pues se utilizó el sistema PROGNOSIS.
4.3) Del total de ventanas consideradas para el entrenamiento (346), 81 clasificaron como perspectivas (23.4% del total).

Del total de ventanas clasificadas como perspectivas 49 (60.5% del total de ventanas clasificadas de perspectivas) se localizan en áreas donde ha sido establecido, por investigaciones anteriores, la mineralización fosforítica. En otras 25 ventanas (30.9% del total de perspectivas) representan indicios favorables para la mineralización fosfórica. Solamente para el caso de 7 ventanas (8,6% del total de perspectivas) no existen evidencias que permitan justificar la clasificación de las mismas.

Aplicando el criterio de evaluación definido en la etapa 2.6, los resultados obtenidos permiten considerar que el procedimiento utilizado conjuntamente con la aplicación del sistema PROGNOSIS permite resolver la tarea de clasificación de las anomalías AGE en relación con su perspectividad para fosforitas.

Como resultado de la clasificación de la matriz de trabajo (integrada por 2015 cuadrículas) se obtuvo que 195 (9.67% del total) cuadrículas presentaban analogías.

Analizando la distribución espacial de estas cuadrículas no se consideraron de interés aquellas que se presentaban de forma aislada. Este análisis permitió definir un total de 14 sectores de interés. El área de estos sectores representa el 13.65% del área total de estudio.

## 5) Análisis e interpretación de los resultados, respecto al problema

5.1) Los 14 sectores fueron sometidos a trabajos de comprobación que incluyeron tanto consulta y generalización de la información geológica de archivo (a distintas escalas) disponible para los sectores de interés, como la aplicación de un complejo de métodos geólogo - geofísicos de campo.
5.2) Como resultado de estas investigaciones se determinó la presencia de 4 sectores que se consideraron como de interés prioritario para la ocurrencia de mineralización en las condiciones de la zona de estudio. El área de estos sectores representa el 6.65% del área total de la Unidad Geológica estudiada.

# 5 Conclusiones

Teniendo en cuenta la potencialidad que para resolver problemas del mundo real tienen las herramientas del Reconocimiento de Patrones, basada en la inmensa cantidad de problemas prácticos relacionados con los principales problemas que aborda esta disciplina: selección de variables, clasificación supervisada, no supervisada y parcialmente supervisada entre otros, era de esperar que el volumen de problemas en los que satisfactoriamente estas herramientas se aplicasen fuese muy grande. Sin embargo esto no es así realmente.

Por otro lado, muchas de las supuestas soluciones a esos problemas prácticos no han pasado de ser interesantes y prometedoras publicaciones en revistas y congresos internacionales o sistemas computacionales que no han salido de los laboratorios de sus creadores, siendo sensible la ausencia de los mismos en los lugares donde especialistas del área de las aplicaciones deberían estar explotándolas en la práctica profesional.

Probablemente una de las causas que explican este hecho está relacionada por la ausencia de una metodología que garantice la introducción real y efectiva de estas herramientas en la práctica profesional de los especialistas de las áreas de las aplicaciones.

El objetivo de este trabajo, basado en la experiencia práctica de los autores y del grupo de investigadores relacionados con ellos, es el de brindarle una herramienta a los especialistas del área del Reconocimiento de Patrones y la Minería de Datos, y quizás a otros especialistas de otras áreas a las que estos principios metodológicos pudieran ser aplicables, para garantizar no sólo la solución del problema que pretenden resolver en la práctica sino también y para nosotros lo más importante, su real introducción en esa área profesional a la que quieren apoyar con sus herramientas. En otras palabras, ofrecer un procedimiento que garanticen que sus soluciones sean realmente **herramientas de los especialistas del área de las aplicaciones**, con la misma familiaridad que un estetoscopio lo es para un médico.

Ejemplos como el que se muestra en el epígrafe 4 y problemas relacionados con éstos también fueron desarrollados sobre la base de estos principios y el lector puede consultarlos en [20,23-34].

No consideramos que esta metodología es un procedimiento acabado y que no pueda ser mejorado, por ello, para los autores sería de mucha utilidad la retroalimentación de las experiencias en la aplicación de la misma, de aquellos que consideren procedente la metodología aquí expuesta.

# Referencias bibliográficas

1. García-Borroto, M., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Medina-Pérez, M.A., Ruiz-Shulcloper, J. (2010). LCMine: An Efficient Algorithm for Mining Discriminative Regularities and its Application in Supervised Classification. Pattern Recognition, Volume 43, Issue 9, pp. 3025–3034.
2. Rodríguez-González, A.Y., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Ruiz-Shulcloper, J. (2011). RP-Miner: A Relaxed Prune Algorithm for Frequent Similar Pattern Mining. Knowledge and Information Systems, Vol. 27, no. 3, pp. 451-471.
3. Rodríguez-González, A.Y., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Ruiz-Shulcloper, J. (2013). Mining frequent patterns and association rules using similarities. Expert Systems with Applications, Volume 40, Issue 17, pp. 6823–6836.
4. Carbajal-Hernández, J.J., Sánchez-Fernández, L.P., Villa-Vargas, L.A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F. (2013). Water quality assessment in shrimp culture using an analytical hierarchical process. Ecological Indicators 29, pp. 148-158.
5. Santos-Gordillo, J.A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F. (2004). Feature Selection using Typical Testors Applied to Estimation of Stellar Parameters. Computación y Sistemas 8/1 pp. 15-23.

6. Martínez-Trinidad, J.F., Velasco-Sánchez M., Contreras-Arevalo, E. (2000). Discovering Differences in Patients with Uveitis through Typical Testors by Class. Lecture Notes in Artificial Intelligence series 1910 pp. 524-529.
7. Tou, J.T., González, R. (1974). Pattern Recognition Principles. Addison Wesley, Reading, Mass.
8. Duda, R.O., Hart, P.E., Stork, D.G. (2001). Pattern Classification. Second ed., John Wiley & Sons.
9. Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition. 2da. edición. Academic Press.
10. Fu, K. (1974). Syntactic Methods in Pattern Recognition. Academic Press.
11. Schalkoff, Robert J. (1992). Pattern Recognition: Statistical, Structural and Neural Approaches, John Wiley & Sons, Inc.
12. Zhuravlev, Yu.I. (1978). On the algebraic approach to the solution of recognition and classification problems. Journal Problemi Kibernetiki, vol. 33, pp. 5-68. (In Russian).
13. Zhuravlev, Yu.I. (1998). An Algebraic Approach to Recognition or Classifications Problems. Pattern Recognition and Image Analysis. Vol. 8, No. 1, pp. 59-100.
14. Kandel, A. (1982). Fuzzy Techniques in Pattern Recognition, John Whiley and Sons.
15. Cheremesina, E.N., Ruiz-Shulcloper, J. (1992). Cuestiones Metodológicas de la Aplicación de Modelos Matemáticos de Reconocimiento de Patrones en Zonas del Conocimiento Poco Formalizadas. Revista Ciencias Matemáticas, vol 13; No.2; pp. 93-108, Cuba.
16. Ruiz-Shulcloper, J., Abidi (2002). Logical Combinatorial Pattern Recognition: A Review. In: Editor S.G. Pandalai. Recent Research Developments in Pattern Recognition, Pub. Transword Research Networks, USA, 3 pp 133-176.
17. Ruiz-Shulcloper, J. (2008). Pattern Recognition with Mixed and Incomplete Data. Proceedings of the 8th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-8-2007), pp. 1-4.
18. Ruiz-Shulcloper, J. (2008). Pattern Recognition with Mixed and Incomplete Data. Journal Pattern Recognition and Image Analysis, vol. 18, No. 4, pp. 563-576.
19. Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition. 2da. edición. Academic Press.
20. Ruiz-Shulcloper, J., Fuentes-Rodríguez, A. (1981). Un Modelo Cibernético para el Análisis de la Delincuencia Juvenil. Revista Ciencias Matemáticas Vol. II(1) pp.141-153. Cuba.
21. Voronin, Yu.A. (1982). Introduction to the Classification Theory. Novosibirsk. (In Russian).
22. Ruiz-Shulcloper, J., Pico-Peña, R., Alamitos-Ibarra, C., Valdés-Hernández, G., Manchado-Martín, A. (1992). "Modelación Matemática del Problema de Discriminación de Anomalías AGE Perspectivas para Rocas Fosfóricas de Génesis Sedimentaria". Revista Ciencias Matemáticas, vol 13; No.2; pp 159-171. Cuba y en "Reconocimiento de estructuras espaciales" pp. 65-80. Editorial Academia.
23. Douglas-De la Peña, M., Ruiz-Shulcloper, J. (1983). Un Algoritmo para el Pronóstico de Enfermedades Laborales Crónicas. Revista Ciencias Matemáticas Vol. IV(1) pp.133-155. Cuba.
24. López-Reyes, N., Ruiz-Shulcloper, J., Gil-Moreno, G., Viera, L. (1988). Un Sistema para el Pronóstico a Corto Plazo de Tormentas Ionosféricas. Reporte de Investigación ICIMAF, No.76, pp 1-25. Cuba.
25. Álvarez-Gómez, L., Ruiz-Shulcloper, J., Chuy-Rodríguez, T., Pico-Peña, R., Cotilla, M. (1992). Modelación Matemática del Pronóstico de Magnitudes Máximas de los Terremotos en la Región del Caribe. En: Editores J. Ruiz-Shulcloper, L. Álvarez-Gómez, V. Guitis. Reconocimiento de Estructuras Espaciales. pp 81-101, Editorial Academia. Cuba.
26. Gómez-Herrera, J.E., Rodríguez-Morán, O., Valladares-Amaro, S., Ruiz-Shulcloper, J., Pico-Peña, R., Echevarría-Rodríguez, G., Tenreyro-Pérez, R., Otero-Marrero, R., Cheremisina, E.N., Cruz-Toledo, R., Barceló-Carol, G., Álvarez-Castro, J., Barea-Centeno, M., García-Sánchez, R. (1994). Pronóstico Gasopetrolífero en la Asociación Ofeolítica Aplicando la Modelación Matemática. Revista Geofísica Internacional, Volumen 33, No. 3, July-Sept., pp 447-467. México.
27. Martínez-Trinidad, J. Fco., Velasco-Sánchez, M, Contreras-Arevalo, E. (2000). Discovering differences in patients with uveitis through typical testors by class. Lecture Notes in Artificial Intelligence 1910 pp. 524-529.
28. Ortíz-Posadas, M.R. (1997). Prognosis and evaluation of cleft palate patients' rehabilitation using pattern recognition techniques. Proceedings of World Congress on Medical Physics and Biomedical Engineering 35, 1, pp. 500-. Niza, France.
29. Ortíz-Posadas, M.R., Martínez-Trinidad, J.F., Ruiz-Shulcloper, J. (1996). A new approach to differential diagnosis of diseases. International Journal of Biomedical Computing, 40, 3, pp. 179-185.
30. Ortíz-Posadas, M.R., Maya-Behart, J., Lazo-Cortés, M. (1998). Evaluation of lips and cleft palate chirurgery using logical combinatorial approach to pattern recognition theory. Journal Revista Brasileira de Bioengenharia. Caderno de Engenharia Biomedica, 14, No. 1, pp. 7-21.

31. Ortíz-Posadas, M.R., Vega-Alvarado, L., Jiménez-Jacinto, V., Lazo-Cortés, M. (1998). The concept of analogy in medicine. A similarity function for cleft palate patients. Proceedings of III Taller Iberoamericano de Reconocimiento de Patrones. México, pp 247-256.
32. Ortíz-Posadas, M.R., Vega-Alvarado, L., Jiménez-Jacinto, V., Lazo-Cortés, M. (1998). A tool for quality service evaluation of the multidisciplinary clinic of lips-clef palate. Proceedings of I Congreso Latinoamericano de Ingeniería Biomédica. Mazatlán, México. pp. 796-799.
33. Ortíz-Posadas, M.R., Vega-Alvarado, L., Jiménez-Jacinto, V., Lazo-Cortés, M., Maya-Behart, J. (1999). Prognosis of cleft palate patients' rehabilitation using a partial precedence algorithm. Proceedings of IV Simposio Iberoamericano de Reconocimiento de Patrones. Conferencia Internacional CIMAF'99. La Habana, pp. 411-418.
34. Ortíz-Posadas, M.R., Vega-Alvarado, L., Maya-Behart, J. (2001). A New Approach to Classify Cleft Lip and Palate. The Cleft Palate-Craniofacial Journal, vol. 38, no. 6, pp. 545–550.
35. Dobrinin, V.N., Cheremesina, E.N. (1988). Métodos Matemáticos y Dispositivos Computacionales en las Investigaciones de Pronósticos Geológicos. Ed. Nauka, Moscú. (In Russian).
36. Ruiz-Shulcloper, J., Pico Peña, R., Alaminos Ibarra, C., Lazo Cortés, M., Boggiano Castillo, M.B., Barreto Fiú, E., Santana Machado, A., Álvarez-Gómez, L., Chuy-Rodríguez, T. (1993). "PROGNOSIS y sus Aplicaciones a las Geociencias". Revista Ciencias Matemáticas, Vol. 14, (2-3), pp. 124-144, Cuba.

RT_054, octubre 2013
Aprobado por el Consejo Científico CENATAV

**Editor:** Lic. Lucía González Bayona
**Diseño de Portada:** Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
**Indicaciones para los Autores:**
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
*Impreso en Cuba*