UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA INDUSTRIAL Y DE SISTEMAS
SECCION DE POST GRADO

"IDENTIFICACIÓN DE PATRONES DE EVASIÓN EN EL SISTEMA DE ADMINISTRACIÓN TRIBUTARIA USANDO TECNOLOGÍA DATA MINING"

TESIS

Para Optar el Grado Académico de:

MAESTRO EN CIENCIAS

MENCION EN INGENIERÍA DE SISTEMAS

EDGAR SÓCRATES VILCAPOMA ESCURRA

Lima – Perú
2003

DESCRIPTORES TEMÁTICOS

- Redes Neuronales
- Data Mining
- Retropropagación
- Predicción
- Evasión
- Auditoria
- Fiscalización

RESUMEN EJECUTIVO

Un problema permanente que entorpece la consecución de las metas de recaudación en la administración tributaria peruana es la evasión. El presente proyecto ha examinado la posibilidad de incorporar nuevas tecnologías y proponer un nuevo esquema para facilitar las tareas orientadas a combatir la evasión tributaria.

En la actualidad, la exploración de contribuyentes con indicios de evasión se realiza fundamentalmente basado en una estrategia de segmentación del universo de contribuyentes, para luego determinar los grupos de contribuyentes con indicios de evasión, haciendo uso intensivo de procedimientos manuales apoyados en consultas a sistemas informáticos no integrados. Este esquema, tiene una gran dependencia del analista tributario para la exploración de los datos en la selección de contribuyentes de probable inclusión en la lista de "fiscalizables", para programas de auditoria. Por lo tanto, la tarea de combatir la evasión no cubre todo el universo de contribuyentes, principalmente por la limitada disponibilidad de recursos para la implementación masiva del esquema anteriormente explicado.

Como alternativa al esquema actual, se propone un uso intensivo de la tecnología de la información, principalmente aquellas tecnologías orientadas al mejor aprovechamiento de la información almacenada en las bases de datos vinculados a los diversos sistemas tributarios existentes. Es en este contexto, donde se constata la necesidad de emplear una nueva tecnología para la explotación de las bases de datos: Data Mining con Redes Neuronales.

A través de la aplicación de Redes Neuronales, se cambia el paradigma de explotación de los datos históricos. Son rasgos fundamentales de esta nueva tecnología, la posibilidad de exploración de la totalidad de los datos a examinar para encontrar patrones ocultos que coadyuven a la predicción de escenarios futuros, con márgenes de error menores a los obtenidos por técnicas convencionales.

Para plasmar lo anterior, el proyecto empieza identificando del escenario de aplicación del proyecto, presentando en el primer capítulo una semblanza de la SUNAT. Luego en el segundo capítulo se define el problema de evasión tributaria, a través de la recolección de las distintas percepciones de los principales protagonistas al interior de la administración tributaria, obteniendo un modelo conceptual consensuado e integrado a partir de los modelos individuales, el mismo que permitió identificar dos componentes básicos del problema.

En el tercer capítulo se examina el componente tributario, mostrando el esquema actual de selección de contribuyentes y el nuevo esquema propuesto, presentando las variables relevantes asociadas al IGV. Luego en el cuarto capítulo se examina el componente tecnológico a través de la exploración de tecnologías emergentes para usufructuar la información tributaria, eligiendo el Data Mining como alternativa tecnológica, previa evaluación de las técnicas que subyacen a la misma, seleccionando las redes neuronales.

En el quinto capítulo se ingresa a profundidad a examinar las redes neuronales, poniendo énfasis en el modelo perceptrón multicapa y el algoritmo de retropropagación. Habiendo elegido el modelo y el algoritmo, en el capítulo sexto se describe la implementación de un sistema predictor multipropósito, detallando los objetos red, capa y nodo. En el séptimo capítulo se usa dicho sistema para el propósito específico de identificación de patrones de evasión, describiendo secuencias de pantallas de ejecución.

Finalmente, en el octavo capítulo se analiza los costos del proyecto de investigación y los beneficios potenciales de la implantación del nuevo esquema de identificación de potenciales evasores, esencialmente por el ahorro de horas-hombre de auditoria. No podían faltar las conclusiones y recomendaciones, plasmados en el noveno capítulo.

INTRODUCCIÓN

Objetivo

Identificar patrones de evasión en el sistema de administración tributaria, aplicando técnicas que subyacen a nuevas tecnologías de manejo de la información (Data Mining). Asimismo proveer herramientas que coadyuven a la predicción y proyección de escenarios futuros de fiscalización, utilizando para ello información histórica almacenada en los repositorios de datos de los sistemas tributarios.

Por lo tanto, se intenta cubrir la ausencia de herramientas de tratamiento de la información de índole predictiva, reduciendo costos de los programas de fiscalización, aportando alternativas de alto poder predictivo, orientadas a minimizar la ocurrencia de casos de auditoria no exitosos.

Antecedentes

La identificación de posibles casos de evasión en la administración tributaria peruana siempre ha sido una labor tediosa, esencialmente manual y con alta dependencia del analista tributario. Durante mucho tiempo, la labor de selección de contribuyentes con indicios de evasión se realizaba a través de un uso intensivo de procedimientos manuales apoyados en consultas a sistemas informáticos no integrados.

En la actualidad la exploración de contribuyentes con indicios de evasión se realiza fundamentalmente basándose en una estrategia de segmentación del universo de contribuyentes. Para cada segmento, se forman grupos según el tipo

de tributo que se desea examinar, habiéndose denominado a estos "clases". Asimismo se han calculado un conjunto de variables definidas a partir de elementos asociados principalmente a la declaración y pago que ha realizado el contribuyente.

La identificación del universo de potenciales fiscalizables, se realiza basándose, fundamentalmente, en la experiencia acumulada de casos anteriores de evasión, plasmados en informes escritos o en archivos de computador. Utilizando como referencia la información almacenada en el sistema, el analista tributario utiliza una diversidad de criterios para seleccionar aquellos contribuyentes que serán incluidos en los programas de auditoria.

Sin embargo, el esquema de selección de fiscalizables explicado, al ser altamente dependiente de la "experiencia" del analista que realiza la labor de selección de contribuyentes, en muchos casos ha involucrado márgenes de error significativos en los programas de auditoria a contribuyentes que luego demostraban no tener motivos para ser considerados como evasores. Esto definitivamente involucra un alto costo para la administración tributaria, ya que en tales casos se ha tenido que desplegar recursos para dicha labor.

La idea central de este nuevo proyecto es la de proveer una herramienta para minimizar la ocurrencia de casos fallidos en el proceso de selección de contribuyentes.

De las averiguaciones realizadas, no existe aplicación tributaria similar, empleando las técnicas elegidas de la tecnología propuesta, en ninguna administración tributaria a nivel de los países latinoamericanos. Asimismo, no existe información suficiente que confirme o descarte la existencia de aplicaciones tributarias con tecnología similar en las instituciones vinculadas a la recaudación tributaria en las países del primer mundo, tales como USA o Europa.

El insumo que ha servido de referencia para plantear la viabilidad del proyecto, está conformado por una diversidad de aplicaciones de índole financiera, tales como por ejemplo:

- Análisis de la situación financiera de una empresa: Basándose en información histórica de empresas del mismo rubro, algunas de las cuales han quebrado y otras no, se intenta predecir para otras empresas, la ocurrencia de una quiebra o descartar la misma.

- Detección de fraudes: Encontrar los patrones y tendencias de compra para detectar comportamientos fraudulentos en el momento de compras con tarjeta de crédito

- Riesgo crediticio: Se utiliza información histórica de clientes para evaluar la factibilidad de otorgar un préstamo, clasificándolos en intervalos, según su comportamiento a través del tiempo.

Fundamento

En la actualidad, el alto costo que involucra auditar a los contribuyentes, y concluir que en algunos casos no existen indicios de evasión, representa una preocupación notoria entre los responsables de los programas de fiscalización.

Como alternativa al esquema actual, se propone un uso intensivo de la tecnología de la información, principalmente aquellas orientadas al mejor aprovechamiento de la información almacenada en las bases de datos que subyacen a los diversos sistemas tributarios existentes. Es en este contexto, donde se constata la necesidad de emplear una nueva tecnología para la explotación de datos: DATA MINING.

El proceso de selección de contribuyentes tiene alta dependencia de la experiencia del analista tributario. Esta experiencia se traduce en la consideración

de una serie de criterios para el aprovechamiento de las variables calculadas a partir de los datos disponibles en los sistemas tributarios. Si bien en algunos casos, algunos analistas han tenido resultados relativamente satisfactorios (pocos casos no exitosos), para otro gran sector de analistas se ha constatado un alto índice de error al momento de seleccionar los contribuyentes fiscalizables.

Es evidente la complejidad inherente al proceso de selección de fiscalizables dada la diversidad de variables que se deben combinar para decidir si un contribuyente es fiscalizable o no. Se trata de proceso de toma de decisión no estructurada o débilmente estructurada. Esto representa un terreno fértil para la aplicación de técnicas innovadoras para el aprovechamiento de los datos y que ayuden al analista tributario en la toma de decisiones.

Alcance

El ámbito del proyecto a nivel geográfico lo constituye la institución encargada de la administración tributaria peruana, es decir la Superintendencia Nacional de Administración Tributaria (SUNAT). Al tener ésta un carácter monopólico (es la única a quien se le ha encargado administrar los tributos del gobierno central), la implantación del proyecto en el corto plazo se limitaría solamente a ésta institución. Sin embargo, existen otras instituciones hacia donde se puede extrapolar la idea, tales como ADUANAS, SAT (encargada de administrar los tributos municipales) o cualquier Municipalidad.

Por otro lado el alcance a nivel tecnológico es incorporar tecnología de explotación de datos de última generación (Data Mining), soportados por conceptos relativos a tópicos avanzados de base de datos e inteligencia artificial.

En materia tributaria, dada la complejidad y diversidad de tributos existentes se ha elegido aquel cuya evasión tiene alta incidencia en materia tributaria: el Impuesto General a las Ventas (IGV).

Contribución sistémica

La propuesta del proyecto de tesis es esencialmente sistémico, porque a través de la aplicación de nuevas tecnologías, intenta superar las limitaciones de los métodos fragmentados, inconexos y para nada sinérgicos que se usan en la actualidad.

La posibilidad de exploración y especialmente la utilización de la totalidad de la información almacenada en las bases de datos históricas a través de Data Mining, es holista, y desecha la visión tubular de explotación de los datos que subyacen a los métodos convencionales de predicción.

En la actualidad el esfuerzo desplegado en la lucha contra la evasión por los funcionarios de la administración tributaria está orientada a combatir la complejidad de detalles. La carencia de alternativas tecnológicas que posibiliten aprovechar eficazmente la cuantiosa información existente, ha hecho que la mayoría de tareas se basen en la elección de muestras de datos y el análisis estático de los mismos.

En contraste, la propuesta de aplicación del Data Mining intenta manejar la complejidad dinámica a través de la identificación de patrones de evasión tributaria, a partir del aprovechamiento de la ingente cantidad de datos históricos que se dispone. Es de particular importancia para el proyecto de tesis, proveer alternativas de análisis con alto poder predictivo.

Por lo tanto, lo anteriormente descrito se puede sintetizar como una propuesta de **"aplicación de moderna tecnología de explotación de datos con intención sistémica".**

CAPITULO I

ESCENARIO DE APLICACIÓN DEL PROYECTO

1.1 Presentación de la Empresa

La SUNAT es una institución pública descentralizada del Sector Economía y Finanzas, creada por Ley No. 24829 y conforme a su Ley General aprobada por Decreto Legislativo N° 501. Está dotada de personería jurídica de Derecho Público, patrimonio propio y autonomía administrativa, funcional, técnica y financiera.

La SUNAT tiene por finalidad administrar y aplicar los procesos de recaudación y fiscalización de los tributos internos, así como proponer y participar en la reglamentación de las normas tributarias. Para ello cuenta con la Intendencia Nacional de Principales Contribuyentes, la Intendencia Nacional de Servicios al Contribuyente, la Intendencia Nacional de Cumplimiento Tributario, la Intendencia Nacional de Sistemas de Información, la Intendencia Nacional de Planeamiento, la Intendencia Nacional de Administración, la Intendencia Nacional Jurídica; 10 Intendencias Regionales (Lima, Arequipa, La Libertad, Lambayeque, Piura, Cusco, Ica, Tacna, Loreto y Junín) y Oficinas Zonales (Huacho, Juliaca, Chimbote, Cajamarca, Cañete, Ucayali, San Martín y Huánuco), así como 21 Oficinas Remotas (Huaraz, Mollendo, Camaná, Puno, Madre de Dios, Abancay, Andahuaylas, Sicuani, Quillabamba, Ayacucho, Tarma, Huancavelica, Pasco, Pacasmayo, Chachapoyas, Jaén, Moyobamba, Tumbes, Talara, Moquegua e Ilo).

1.1.1 Conceptos Básicos

1.1.1.1 Tributación

Se refiere al conjunto de obligaciones que deben realizar los ciudadanos sobre sus rentas, sus propiedades, mercancias, o servicios que prestan, en beneficio del Estado, para su sostenimiento y el suministro de servicios tales como defensa, transportes, comunicaciones, educación, sanidad, vivienda, etc.

1.1.1.2 Tributo

Prestación generalmente pecuniaria que el Estado exige en virtud de una ley, para cubrir gastos que le demanda el cumplimiento de sus fines. El Código Tributario rige las relaciones jurídicas originadas por los tributos. Para estos efectos, el termino genérico tributo comprende impuestos, contribuciones y tasas.

1.1.1.3 Impuesto

Tributo cuyo cumplimiento no origina una contraprestación directa en favor del contribuyente por parte del Estado.

1.1.1.4 Contribución

Aporte voluntario de una cantidad para un fin determinado

1.1.1.5 Tasa

Es el tributo cuya obligación tiene como hecho generador la prestación efectiva por el Estado de un servicio público individualizado en el contribuyente. No es tasa el pago que se recibe por un servicio de origen contractual. Son especies de este género los arbitrios, los derechos y las licencias.

1.1.1.6 Contribuyente

Es aquel deudor tributario que realiza o respecto del cual se produce el hecho generador de la obligación tributaria. Se define también como la persona Natural o Jurídica que tenga patrimonio, ejerza actividades económicas o haga uso de un derecho que conforme a ley genere la obligación tributaria.

1.1.1.7 Obligación Tributaria

Es el vinculo entre el acreedor y el deudor tributario, establecido por ley y de derecho público. Tiene por objeto el cumplimiento de la prestación tributaria y es exigible coactivamente.

1.1.1.8 Declaración Jurada

Es la manifestación de hechos comunicados a la Administración Tributaria en la forma establecida por leyes o reglamentos, la cual puede constituir la base para la determinación de la obligación tributaria.

Los deudores tributarios deberán consignar en su Declaración, en forma correcta y sustentada, los datos solicitados por la Administración Tributaria.

Se presume, sin admitir prueba en contrario, que toda Declaración Tributaria es jurada.

1.1.1.9 Acreedor Tributario

Es la persona que tiene el derecho para exigir el pago de una deuda resultante del incumplimiento de una obligación tributaria.

1.1.1.10 Deudor Tributario

Es la persona obligada al cumplimiento de la prestación tributaria como contribuyente o responsable.

1.1.1.11 Deuda Tributaria

Suma adeudada al acreedor tributario por concepto de tributos, recargos, multas, intereses moratorios y de ser el caso los intereses que se generan por el acogimiento al beneficio de fraccionamiento o aplazamiento previsto en el Código Tributario.
Se entiende por deuda tributaria la que procede de un hecho imponible y todas las sanciones producidas en el desarrollo de la relación tributaria.

1.1.2 Tributos administrados

Con el fin de lograr un sistema tributario eficiente, permanente y simple se dictó la Ley Marco del Sistema Tributario Nacional (Decreto Legislativo N° 771), vigente a partir del 1 de enero de 1994.

La ley señala los tributos vigentes e indica quiénes son los acreedores tributarios: el Gobierno Central, los Gobiernos Locales y algunas entidades con fines específicos. Tratándose de los tributos correspondientes al Gobierno Central, los entes administradores son la SUNAT (tributos internos) y ADUANAS (derechos arancelarios).

Los principales tributos que administra la SUNAT son los siguientes:

Impuesto General a las Ventas: Es el impuesto que se aplica en las operaciones de venta e importación de bienes, así como en la prestación de distintos servicios comerciales, en los contratos de construcción o en la primera venta de inmuebles.

Impuesto a la Renta: Es aquél que se aplica a las rentas que provienen del capital, del trabajo o de la aplicación conjunta de ambos.

Régimen Unico Simplificado: Es un régimen simple que establece un pago único por el Impuesto a la Renta y el Impuesto General a las Ventas (incluyendo al Impuesto de Promoción Municipal). A él pueden acogerse únicamente las personas naturales o sucesiones indivisas, siempre que desarrollen actividades generadoras de rentas de tercera categoría (bodegas, ferreterías, bazares, puestos de mercado, etc.) y cumplan los requisitos y condiciones establecidas.

Impuesto Selectivo al Consumo: Es el impuesto que se aplica sólo a la producción o importación de determinados productos como cigarrillos, licores, cervezas, gaseosas, combustibles, etc.

Impuesto Extraordinario de Solidaridad: A partir del 1 de setiembre de 1998, este impuesto sustituyó a la Contribución al Fondo Nacional de Vivienda (FONAVI). La tasa vigente es 5%, y se aplica sobre las remuneraciones que abonan los empleadores y sobre las rentas que perciben los trabajadores y profesionales independientes.

Impuesto de Solidaridad en favor de la Niñez Desamparada Son sujetos de este impuesto las personas que soliciten la expedición o revalidación de pasaportes.

Impuesto a las Acciones del Estado. A partir del 1 de enero del 2001, se gravó la propiedad de las acciones del Estado, con una tasa del 5%. Es aplicable a las empresas cuyo capital, de manera directa o indirecta, pertenece íntegramente al Estado.

Aportaciones al ESSALUD y a la ONP: Mediante la Ley N° 27334 se encarga a la SUNAT la administración de las citadas aportaciones, manteniéndose como acreedor tributario de las mismas el Seguro

Social de Salud (ESSALUD) y la Oficina de Normalización Previsional (ONP).

1.1.3 Los clientes

Podemos identificar dos grandes grupos de clientes: los clientes externos y los clientes internos.

Se considera clientes externos de la Administración Tributaria, a toda persona natural o jurídica que debe pagar tributos.

Clientes internos lo constituyen todos los trabajadores de la institución, cuyo protagonismo ha demostrado ser determinante en la puesta en marcha de los proyectos de la SUNAT.

Tanto los clientes internos como externos, tienen expectativas, es decir, esperan recibir algo de la SUNAT.

1.1.3.1 Expectativas de los Clientes Externos

- Reducción de los impuestos.
- Recibir orientación de las obligaciones tributarias.
- Un servicio ágil y oportuno.
- Ampliación de los horarios de atención (Por ejemplo: los días sábados).
- Facilidades para el cumplimiento de sus obligaciones tributarias.

1.1.3.2 Expectativas de los Clientes Internos

- Capacitación permanente
- Agradable ambiente de trabajo
- Comunicación efectiva

- Adecuada política de remuneraciones
- Participación constante

1.1.4 Producto o Servicio entregado

La SUNAT brinda servicios. Podríamos decir de otro modo, que la SUNAT "vende imagen" de institución integra.

La principal preocupación de la Administración Tributaria es la de ofrecer todas las facilidades para que los contribuyentes (clientes externos) cumplan con sus obligaciones tributarias y para que sus trabajadores (clientes internos) se encuentren motivados.

1.1.5 La competencia

El Estado peruano, por ley, ha encargado a la SUNAT las tareas vinculadas a la administración de los tributos internos, dándole exclusividad. El público percibe a esta como una institución monopólica en materia tributaria, y de hecho lo es.

Sin embargo, esto no significa que no tenga competencia. Haciendo referencia a Michael Porter, el manifiesta que la competencia se genera "entre las fuerzas en contienda". Por lo tanto, desde este punto de vista, la SUNAT tiene un gran contendor: *la evasión tributaria.* En este sentido, la Administración Tributaria orienta sus esfuerzos a detectar y combatir todas las modalidades de evasión tributaria.

Además, Porter hace una apreciación interesante cuando dice que "competidor es todo aquel que se puede oponer al logro de nuestros objetivos ahora *y en el futuro*". Por lo tanto, no resulta ilógico suponer que en el futuro, los gobiernos de turno, puedan intentar la creación de un ente paralelo para la administración de algunos o todos los tributos internos del país. Esta suposición puede materializarse si la actual Administración

Tributaria no demuestra ser eficiente y eficaz en el logro de los objetivos y mas aún, si no satisface las expectativos del Gobierno.

1.2 Análisis Estratégico

1.2.1 Plan Estratégico

1.2.1.1 Misión.

La Superintendencia Nacional de Administración Tributaria (SUNAT) tiene por Misión Institucional:

"Contribuir al financiamiento sostenible del proceso de desarrollo económico, social e institucional del país, a partir del establecimiento de una relación honesta y justa con los contribuyentes que, vía la provisión de servicios de calidad al contribuyente y la generación efectiva de riesgo, permita asegurar la ampliación de la base tributaria y un adecuado nivel de recaudación."

1.2.1.2 Visión

La Superintendencia Nacional de Administración Tributaria (SUNAT) espera que al cabo de cinco años:

a. Los contribuyentes, informados de sus derechos y obligaciones tributarias, perciban una preocupación permanente en el personal SUNAT por atenderlos oportunamente y a satisfacción con productos y servicios de calidad.
b. La SUNAT cuente con capacidad institucional para atender y asistir al contribuyente de manera efectiva, oportuna, precisa y con sistemas eficaces y procedimientos uniformes.
c. Los contribuyentes mejoren su cumplimiento tributario debido a la capacidad técnica de la Superintendencia Nacional de Administración

Tributaria (SUNAT) en la generación de riesgo, la detección del incumplimiento y la aplicación de sanciones efectivas, conforme a ley.

d. La SUNAT cuente con capacidad institucional para promover de manera permanente el desarrollo personal y profesional de sus trabajadores, brindándoles un ambiente de trabajo que contribuya de manera importante al cumplimiento eficaz de labores.
e. La SUNAT cuente con capacidad institucional para mantener un bajo nivel de costos operativos indirectos (no vinculados a sus responsabilidades de servicios a los contribuyentes y a la generación de riesgo).
f. La SUNAT sea percibida por la opinión pública y los contribuyentes como una institución honesta que contribuye, junto con otros organismos públicos, a formular una política tributaria equitativa, uniforme (mínimos regímenes especiales), de amplia base tributaria (mínimas exoneraciones), con pocos impuestos y con tasas moderadas y uniformes.
g. La SUNAT cuente con capacidad institucional para asegurar que los sistemas y la información tributaria de los contribuyentes sea reservada y utilizada sólo para fines técnico-tributarios.

1.2.1.3 Valores Corporativos.

- Honestidad a toda prueba.
- Justicia (Exigir al contribuyente que pague lo que le corresponde).
- Equidad
- Puntualidad
- Solidaridad
- Etica

1.2.1.4 Objetivos

1.2.1.4.1 Objetivo General

Lograr la óptima orientación de los medios que intervienen en la

administración de los procesos de recaudación y fiscalización de los tributos.

1.2.1.4.2 Objetivos Específicos

- Proveer al estado los fondos necesarios para el funcionamiento.
- Lograr un adecuado servicio de atención y orientación al contribuyente.
- Reducir la evasión tributaria.
- Crear un verdadero riesgo para el evasor.
- Maximizar el cumplimiento voluntario.
- Lograr una fuerte imagen de integridad y eficacia de la administración tributaria.

1.2.1.5 Opción institucional frente a dilemas estratégicos

A pesar que la Visión y la Misión Institucionales señalan claramente una dirección estratégica hacia donde dirigir los esfuerzos institucionales, en dicho proceso se han de suscitar sin duda dilemas de accionar institucional que obligan a su planteamiento explícito al momento inicial de definir una estrategia. En ese sentido, la Superintendencia Nacional de Administración Tributaria (SUNAT) ha resuelto enfrentar eventuales dilemas estratégicos en el siguiente sentido:

1.2.1.5.1 Estructura Organizacional

a. La estructura organizacional responde a un accionar basado en la descentralización geográfica, donde las dependencias controlan todas las funciones básicas con respecto a todos los contribuyentes de su área de acción; sólo en el caso de la Intendencia Nacionales de Principales Contribuyentes (INPC) se justifica una especialización en un solo tipo de contribuyente.

b. A nivel de toma de decisiones, la estructura organizacional de línea coexiste con equipos de proyectos, cuyos líderes tienen autonomía de gestión (especialmente cuando se trata de equipos multifuncionales).
c. Se requiere incrementar el número de unidades operativas, a fin de facilitar el control eficiente del actual directorio de los contribuyentes de la Intendencia Regional Lima (donde cabe pensar en la creación de nuevas Intendencias Regionales, oficinas zonales y/o centros de atención en Lima Metropolitana).

1.2.1.5.2 Control del cumplimiento tributario

a. La programación de la fiscalización es fundamentalmente centralizada, pero se complementa con las programaciones especiales que, de manera coordinada, realiza cada dependencia a nivel descentralizado. De este modo, se homogenizan criterios a nivel nacional, pero se tiene siempre presente la retroalimentación proveniente de las regiones.
b. La auditoria es realizada en el campo fundamentalmente por contadores, pero es previamente planeada y asistida por equipos multidisciplinarios.
c. El auditor determina la deuda del contribuyente, pero no ejerce la función de cobranza (en consideración tanto al perfil y especialización que requiere la cobranza como al riesgo de centralizar todo el proceso en el auditor).
d. Si bien la auditoria constituye una función vital para la SUNAT, existen ámbitos en los que se podría delegar esta función a terceros, a fin de ampliar la cobertura y especialización del control.

1.2.1.5.3 Rol del orientador

a. El rol del orientador tributario es informar y asistir al contribuyente, ya sea transmitiéndole conocimiento sobre sus obligaciones y derechos, o

proporcionándole servicios y medios que faciliten su cumplimiento tributario.

1.2.1.5.4 Desarrollo de personal

a. La línea de carrera contempla una línea profesional alternativa para los profesionales que no siguen la línea directiva.
b. La capacitación integral es manejada por una sola entidad, si bien el desarrollo efectivo de los programas puede ser delegado parcial o totalmente a terceros.
c. El programa de becas es, básicamente, una inversión de la institución en capital humano y, consecuentemente, constituye un beneficio para el trabajador.
d. La política remunerativa considera el desempeño y los resultados obtenidos por el trabajador, como el nivel jerárquico y el puesto. La política de rotación de personal es una práctica continua, tanto entre diferentes unidades organizacionales como entre áreas geográficas.
e. Los destaques del personal de SUNAT a otras instituciones públicas y privadas constituyen una inversión institucional, aunque no constituyen una prioridad en el corto plazo.

1.2.1.5.5 Herramientas de apoyo:

a. Los sistemas facilitan y no determinan la aplicación de los procedimientos y normas legales.
b. La administración de la base de datos es centralizada porque ello asegura la calidad de los datos, pero debe evaluarse la administración y el acceso a los sistemas de manera descentralizada.
c. Los sistemas para declarar información y controlar la deuda son desarrollados internamente en SUNAT; sin embargo, debe evaluarse el encargar a terceros el desarrollo de los sistemas administrativos y legales estándares.

d. El proceso de notificaciones se realiza por terceros, aunque bajo ciertos controles por parte de SUNAT.
e. El manejo de la documentación interna debe tender a realizarse totalmente por vía electrónica.

1.2.1.5.6 Imagen institucional

a. Los lineamientos estratégicos de la Institución son difundidos a la opinión pública, para mayor conocimiento de los objetivos que guían el trabajo de SUNAT.
b. La SUNAT proyecta una única imagen a nivel nacional, la cual se retroalimenta con la proyección de su imagen internacional.
c. La estrategia de difusión de información contempla la suscripción de convenios bilaterales, así como la disponibilidad pública de información no vinculada a la reserva tributaria.

1.2.2 Diagnóstico Estratégico

1.2.2.1 Análisis Externo

1.2.2.1.1 Oportunidades

Establecer convenios de cooperación técnica con organismos internacionales.
Participar en la definición de la política tributaria.
Acrecentar el respaldo de la política económica del gobierno.
Consolidar rol protagónico en el destino de la economía del país.
Extender servicios de recaudación y cobranza a mas instituciones del Estado.

1.2.2.1.2 Amenazas

La variabilidad de la política tributaria de cada gobierno.

La inestabilidad económica del país.

Pauperización progresiva de la población.

La acentuada recesión en los sectores económicos.

Creciente disconformidad con la política tributaria en sectores de alta contribución.

1.2.2.2 Análisis Interno

1.2.2.2.1 Fortalezas

Contar con el respaldo del gobierno.

Poseer cuadros profesionales competentes en permanente actualización.

Procedimiento transparente para la selección de personal.

Frecuente incorporación de tecnología de punta para potenciar el cumplimiento de las funciones encomendadas.

Institución pública con autonomía funcional, económica, técnica, financiera y administrativa.

1.2.2.2.2 Debilidades

Los servicios que se otorgan al contribuyente son insuficientes en cobertura. A nivel nacional, existirían alrededor de 14,000 contribuyentes por orientador. Como consecuencia de ello, la información que se brinda a través de la orientación personalizada no es uniforme ni completa, las llamadas efectivamente atendidas lo son con tiempos de espera aún considerables, y por cada nueva campaña que entra en vigencia (Ejm.

Fraccionamientos) se generan largas colas que deben realizar los contribuyentes para ser atendidos.

La fuerza efectiva de SUNAT para generar riesgo ha disminuido, lo cual se expresa en los siguientes hechos: (a) disminución en un 50% del número de auditorias entre 1998 – 2000, (b) controles insuficientes, pues a manera de ejemplo, en el año 2000, el porcentaje promedio de auditorias realizadas respecto al número de inscritos no sobrepasaba el 8%, (c) los Programas de Fiscalización no se encuentran debidamente actualizados los estudios contables y/o contadores terminan conociéndolos de antemano y contrarrestan la labor de auditoria.

El personal de SUNAT tiene poco conocimiento de los lineamientos y objetivos institucionales. De otro lado, las líneas de comunicación son poco desarrolladas por lo que se constata que existe una poca integración del personal y que los ambientes físicos no son plenamente apropiados para el desempeño laboral. Finalmente, la ausencia de un plan de línea de carrera formal, así como algunas limitaciones de los sistemas de evaluación han generado cierta desmotivación en el personal.

La imagen de la Institución se ha deteriorado ante la opinión pública, en parte, debido a un accionar institucional no necesariamente técnico, con una Alta Dirección con frecuente rotación que ha mostrado en el tiempo diversas orientaciones y priorizaciones institucionales, que a nivel nacional fueron seguidas con distinta intensidad.

1.3 Análisis Funcional

1.3.1 Unidades Organizacionales

La SUNAT cuenta con la siguiente estructura organizacional

Alta Dirección

- Superintendente Nacional de Administración Tributaria
- Superintendente Nacional Adjunto de Tributos Internos
- Superintendente Nacional Adjunto de Tributos Aduaneros

Organos de Apoyo

- Secretaría General
- Comité Alta Dirección
- Instituto de Administración Tributaria y Aduanera

Organos de Control

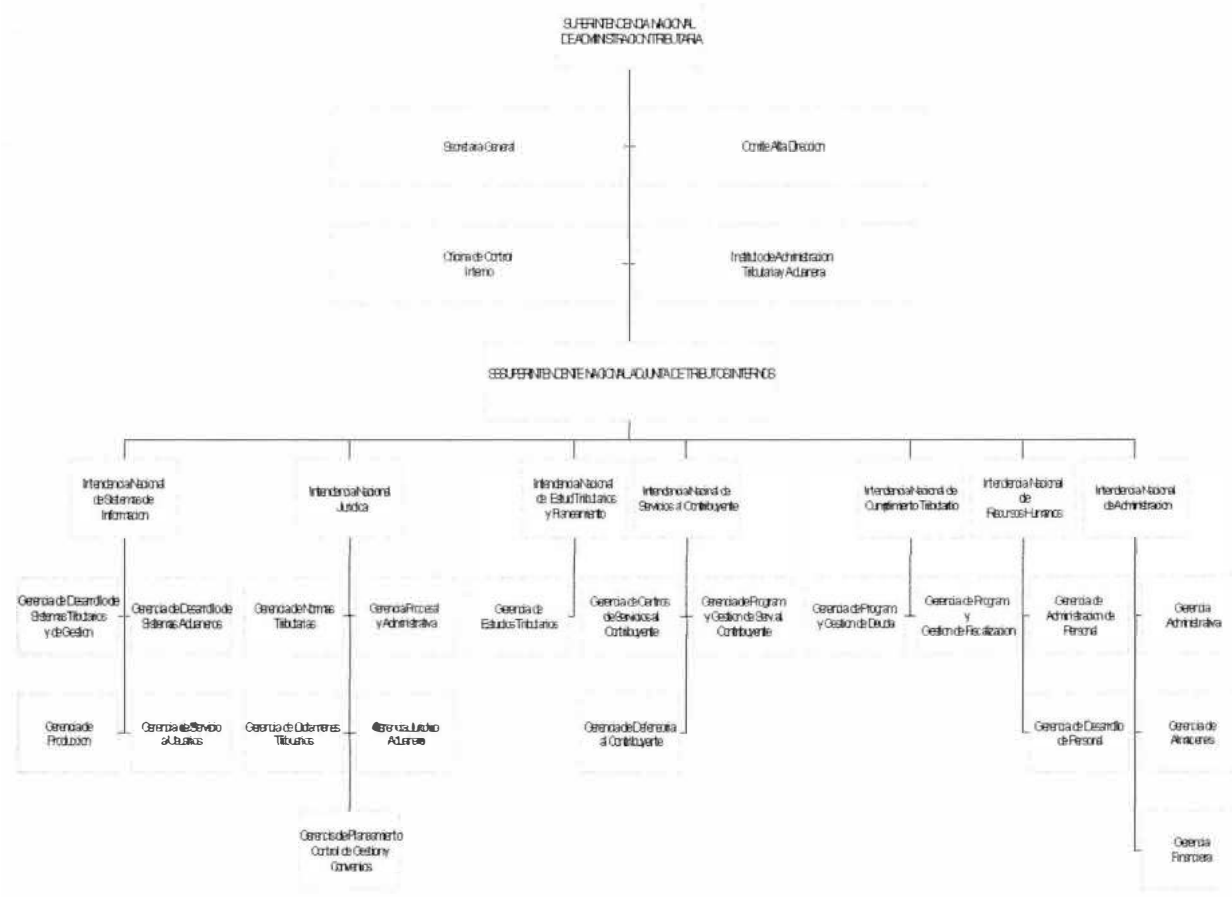
- Oficina de Control Interno

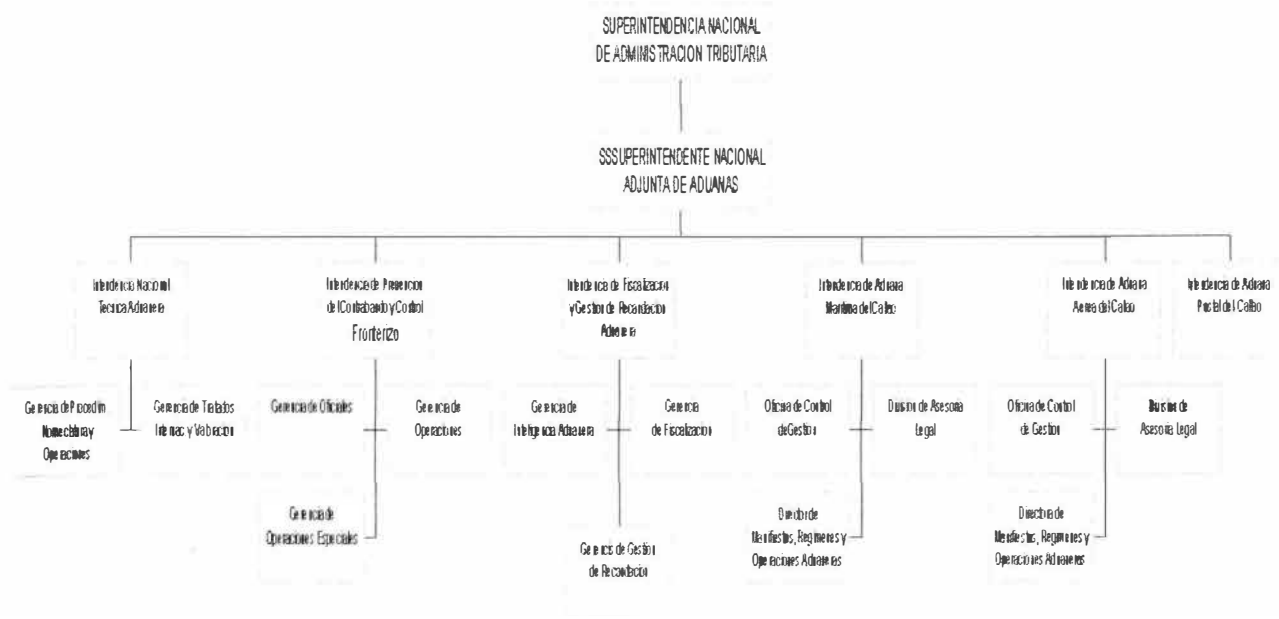
Organos de Línea

- Intendencia Nacional de Administración
- Intendencia Nacional de Recursos Humanos
- Intendencia Nacional de Cumplimiento Tributario
- Intendencia Nacional de Servicios al Contribuyente
- Intendencia Nacional de Estudios Tributarios y Planeamiento
- Intendencia Nacional Jurídica
- Intendencia Nacional de Sistemas de Información
- Intendencia Nacional de Principales Contribuyentes

Organos Descentralizados

- Intendencias Regionales y Oficinas Zonales





1.3.2 Principales Funciones del Negocio

Las principales funciones de negocio de la Administración Tributaria son:

- Planificar
- Recaudar
- Fiscalizar
- Reglamentar

1.3.2.1 Planificar

- Establecer los objetivos estratégicos del negocio
- Definir las metas operativas anuales para el logro de los objetivos estratégicos del negocio

1.3.2.2 Recaudar

Empadronar

Permite identificar al sujeto pasvio de las obligaciones tributarias

- **Recaudar**

  Recolectar los fondos provenientes del pago de las obligaciones tributarias

- **Controlar la Recaudación**

  Llevar un control de las obligaciones tributarias de los contribuyentes

- **Cobrar Coactivamente**

  Realizar el cobro coactivo de la deuda morosa

1.3.2.3 Fiscalizar

- **Inspeccionar**

  Permite conocer si el contribuyente esta cumpliendo con sus obligaciones tributarias.

- **Investigar**

  Permite conocer el monto de las obligaciones tributarias, accionar sobre el incumplimiento de las obligaciones formales y las obligaciones de pago.

- **Determinar**

  Permite determinar la omisión o falsedad en lo declarado y sancionar el incumplimiento de las obligaciones tributarias

1.3.2.4 Reglamentar

- **Normar**

  Apoyar al diseño de las reglas que rigen el marco tributario
  Proponer al Ministerio de Economía y Finanzas la reglamentación de las normas tributarias y participar en su elaboración

- **Resolver**

  Definir y sustentar la posición de la Administración, cuando existan aspectos tributarios controvertidos.

1.3.3 Facultades de la SUNAT

Para el desempeño de las funciones de negocio, la normatividad vigente ha asignado facultades a la administración tributaria. Entre las facultades de la SUNAT se deben señalar:

1.3.3.1 Facultad de Recaudación

Para el ejercicio de esta facultad, en julio de 1993, la SUNAT puso en marcha el Sistema de Recaudación Bancaria. En la actualidad, la SUNAT tiene suscrito un Convenio de Recaudación con siete bancos, los que reciben a través de sus sucursales y agencias en todo el país las declaraciones de pago de los contribuyentes, facilitando así el cumplimiento voluntario de sus obligaciones.

Los contribuyentes presentan hoy sus declaraciones a través de medios magnéticos utilizando como base un software denominado Programa de Declaración Telemática (PDT). El disquete es presentado por los Principales Contribuyentes en las oficinas de la SUNAT, mientras que los medianos y pequeños contribuyentes los realizan a través de la red bancaria. El PDT se encuentra en la página Web de la SUNAT.

Los bancos que conforman la red autorizada son los siguientes:

- Banco de Crédito del Perú.
- Interbank
- Banco Wiese Sudameris.
- Banco Continental.
- Banco de la Nación

Asimismo, con la finalidad de facilitar y evitar errores en la presentación de sus declaraciones tributarias, la SUNAT ha desarrollado programas de declaración telemática tales como el PDT - Renta Anual, PDT IGV - Renta Mensual y el PDT – Remuneraciones, entre otros.

1.3.3.2 Facultad de Determinación

Con esta facultad, la SUNAT verifica el cumplimiento de las obligaciones tributarias, identifica al deudor tributario y señala la base imponible de la obligación tributaria. De allí la emisión de Resoluciones de Determinación y Ordenes de Pago que son valores a cargo del contribuyente, emitidos en aquellos casos en que éstos no calcularon correctamente sus tributos o no los pagaron totalmente.

1.3.3.3 Facultad de Cobranza Coactiva

Es la facultad que se ejerce a través del ejecutor coactivo como última fase del proceso de cobro de la deuda tributaria exigible al contribuyente o responsable de los tributos.

1.3.3.4 Facultad de Fiscalización

Esta facultad otorga a la SUNAT la potestad de investigar, inspeccionar y controlar el cumplimiento de las obligaciones tributarias. Para ello cuenta con un programa y tecnología moderna que le permite la detección automática del incumplimiento.

1.3.3.5 Facultad de Sanción

La SUNAT sanciona las infracciones derivadas del incumplimiento de las siguientes obligaciones:

a) Inscribirse en el RUC.
b) Emitir y exigir comprobantes de pago.
c) Llevar libros y registros contables.
d) Presentar declaraciones y comunicaciones.
e) Permitir el control de la Administración Tributaria.
f) Otras obligaciones tributarias.

CAPITULO II

DEFINICIÓN DEL PROBLEMA A PARTIR DE LA COSMOVISIÓN DE LOS USUARIOS DE LOS SISTEMAS TRIBUTARIOS.

2.1 Identificación de la Situación Problema: La Evasión Tributaria

La administración tributaria peruana tiene múltiples problemas por resolver; sin embargo dada la limitación de recursos, fundamentalmente humanos, solo se pueden afrontar parte de ellos. Es en este contexto que surge la preocupación por identificar el problema más álgido y plantear alternativas para resolverlo.

La primera labor realizada para la elección del problema que debe analizarse fue interactuar con personas representativas de la administración tributaria. La idea básica ha sido recoger las diferentes percepciones sobre los problemas que se enfrenta en la actualidad. Para esto, se ha aplicado parcialmente la secuencia metodológica de la Metodología de Sistemas Blandos (MSB).

El objetivo de esta etapa del proyecto es recopilar la diversidad de percepciones y luego integrarlas en un enunciado consensuado, referido al problema de mayor trascendencia, cuya solución tenga gran impacto para la institución en particular y el país en general.

Para la selección de las personas se ha identificado primeramente los elementos del sistema contenedor del problema y el sistema solucionador del problema. Luego se han elegido de ambos grupos a personas con las que se ha interactuado para recoger sus opiniones.

La totalidad de elegidos han coincidido en que la administración tributaria peruana tiene múltiples problemas, pero a la vez todos han incidido en mayor o menor

grado en el problema de la evasión tributaria como obstáculo para la consecución de las metas de recaudación en particular y los objetivos institucionales en general.

2.1.1 El Sistema Contenedor del Problema

Cabe precisar que el Sistema Contenedor del Problema (SCP) según la MSB, "es aquella porción de la realidad conformado por el sistema y su entorno que lo circunda, donde existen personas que conforman grupos culturales y que adoptan el papel de 'vivir los problemas' de esa realidad". Es decir, contiene aquellas personas que son directamente afectadas por la existencia del problema, pero que no son necesariamente protagonistas en la búsqueda de soluciones. Son aquellos que "se han acostumbrado" al escenario problemático vigente.

Podemos citar como elementos del SCP de la administración tributaria a los siguientes grupos de personas:

Auditores
Fedatarios
Usuarios finales de los sistemas de fiscalización

Estos grupos están cotidianamente inmersos en el diario quehacer para "combatir" la evasión, pero sus acciones obedecen a lineamientos establecidos por sus superiores u otras áreas de la administración tributaria; son simplemente ejecutores de planes de acción pre-establecidos, orientados a encontrar casos de evasión en un universo de contribuyentes que otros han seleccionado.

Los miembros del SCP desarrollan su labor con diligencia y consiguen en varios casos encontrar reparos que revelan casos de evasión, pero en otros casos, por mas revisión que realicen no encuentran indicio alguno que demuestre un accionar incorrecto del contribuyente. Esto último es desalentador para el auditor, y excesivamente oneroso para la administración tributaria, dado que se han perdido

muchas horas-hombre y otros recursos, para finalmente concluir en una auditoria no exitosa.

2.1.2 El Sistema Solucionador del Problema

Según la MSB, el Sistema Solucionador del Problema (SSP), "Está conformado por aquellas personas que tienen vocación de 'solucionadores' y que han tomado la decisión de "solucionar" los problemas existentes en el SCP. Es el sistema que recogiendo los requerimientos y aspiraciones del SCP, propone 'soluciones' a ser implantadas en el SCP".

En la administración tributaria peruana mucho predomina el "día a día". La gran mayoría de personas está inmersa en "resolver" problemas cotidianos y recurrentes. Son pocos los que se detienen sobre la marcha, analizan la situación e intentan resolverlo de una manera diferente y creativa.

Este proyecto, en el intento de encontrar una opción innovadora para enfrentar el problema de la evasión tributaria, indagó por aquellas personas que en teoría deberían ser protagonistas en la solución del problema. Aunque algunos no accedieron a dar opiniones, básicamente por restricción de tiempo, se consiguió en otros una amplia predisposición a opinar sobre el particular, y que en alguna medida "tomaron la decisión de solucionar el problema", requisito esencial según la MSB para pertenecer al SSP.

Aunque cualquier persona puede pertenecer al SSP, dado que es más cuestión de actitud que aptitud, las siguientes personas han sido seleccionadas como miembros del SSP en el contexto del proyecto, en el intento de abordar el problema de la evasión tributaria:

Intendente Nacional de Estudios Tributarios y Planeamiento
Analista Tributario
Analista de Sistemas

Jefe de Programas de Fiscalización

Auditor

La contribución de estos cinco elementos representativos será crucial para encontrar nuevas formas de enfrentar exitosamente el problema de la evasión.

Cabe mencionar que según la MSB, el SCP y el SSP no son conjuntos disjuntos, es decir no son excluyentes necesariamente. Por lo tanto, podemos encontrar miembros de la organización que pertenecen tanto al SCP como el SSP. Tal es el caso del quinto elemento del SSP (el auditor) que también fue indicado como parte del SCP. Lógicamente, no todos los auditores tendrán la aptitud o actitud de jugar este doble rol, sin embargo encontramos algunos que si estaban dispuestos a contribuir a resolver el problema de la evasión tributaria desde una perspectiva diferente a su labor habitual.

Identificar aquellos miembros del SCP que pueden pertenecer al SSP es de vital importancia para entender adecuadamente el problema, ya que son ellos los que viven el problema. El auditor, en este caso, nos ha permitido conocer en detalle la problemática de la tarea de auditoria en directa relación con la evasión tributaria.

2.1.3 Cuadro Pictográfico de la Situación Problema

El cuadro pictográfico, según la MSB, es la representación "a mano alzada" de la "primera impresión" que el analista percibe de la realidad que está abordando, basándose en la observación directa, en referencias y/o datos preliminares que haya podido recoger, sin entrar en mayor detalle.

Se muestra a continuación un cuadro pictográfico obtenido como resultado de una primera fase de exploración orientada a identificar el clima organizacional que predomina en la administración tributaria peruana en el contexto de la lucha contra la evasión.

2.2 Alternativas para afrontar la situación problema

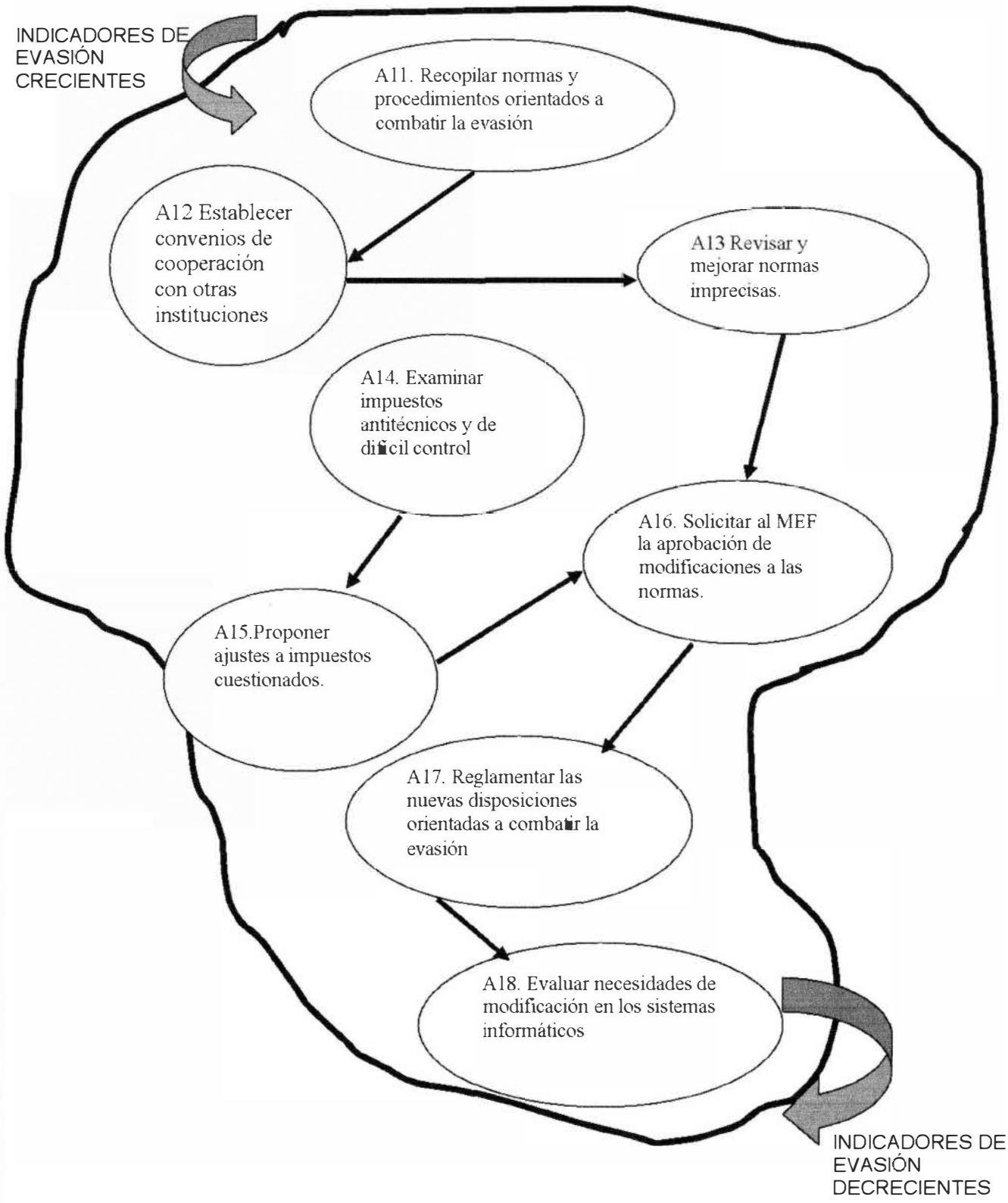
2.2.1 Recolección de percepciones

A cada una de los miembros del SSP señalado anteriormente, se les solicitó su percepción sobre el problema que se está examinando. La modalidad de recopilación, combinó en algunos casos medios escritos, electrónicos (e-mails) y comunicación directa (diálogo).

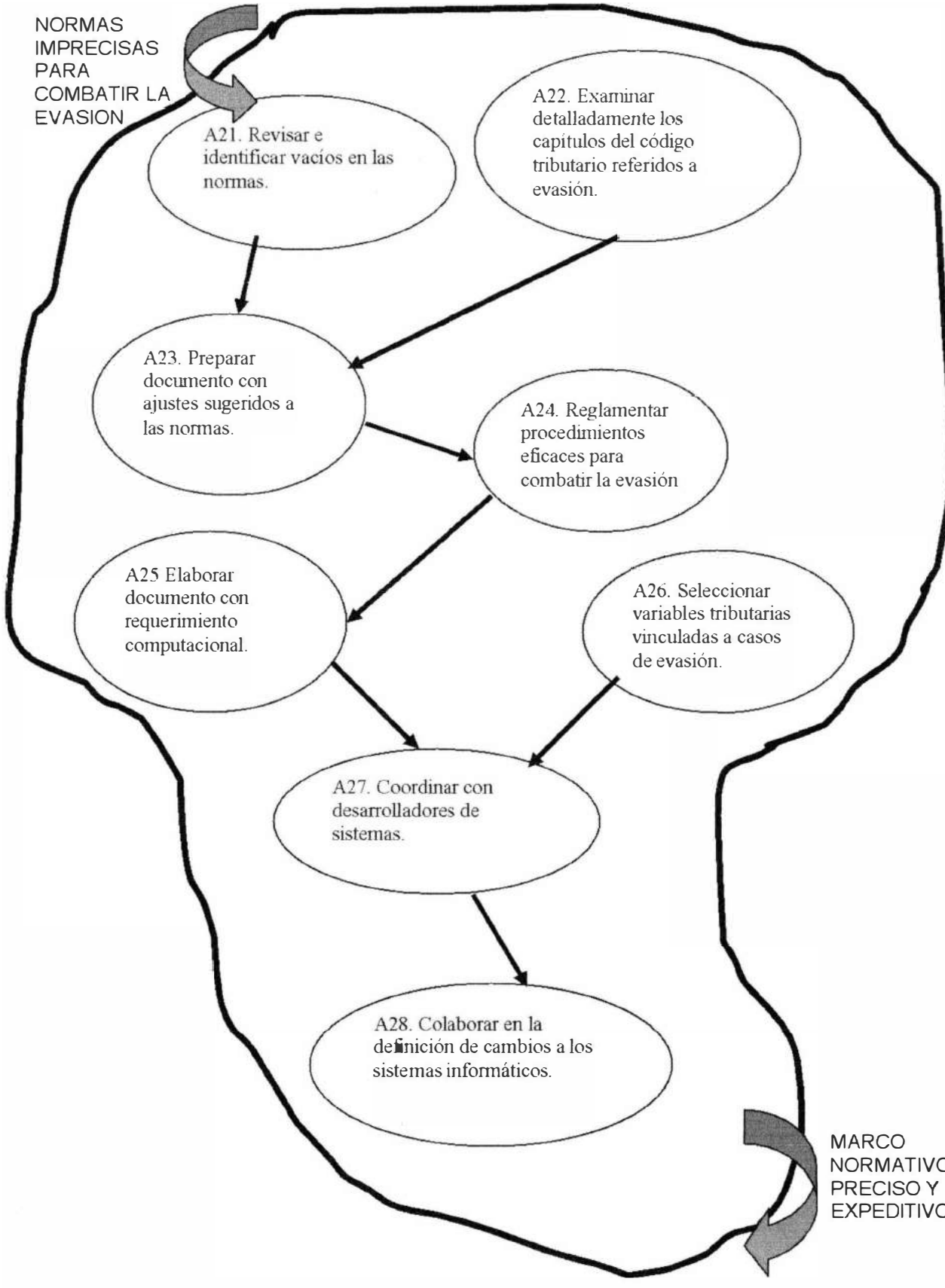
Utilizando la nomenclatura de la MSB, a cada una de las personas indicadas las denominaremos en adelante observantes. La percepción del problema de cada uno de estos observantes constituye el weltanschauung o cosmovisión de la realidad que se está abordando.

2.2.2 Obtención de modelos conceptuales

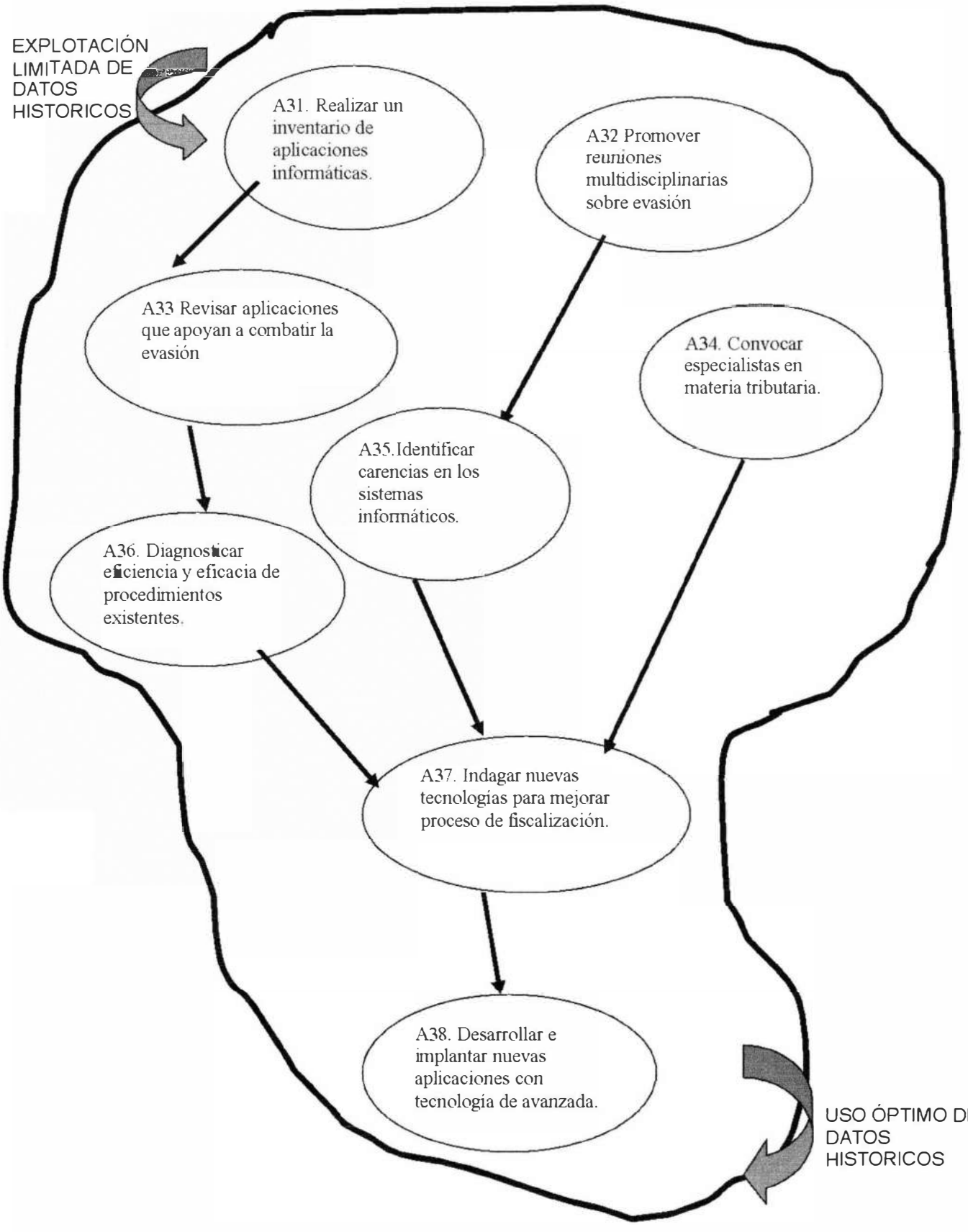
Modelo Conceptual del Intendente Nacional de Estudios Tributarios y Planeamiento



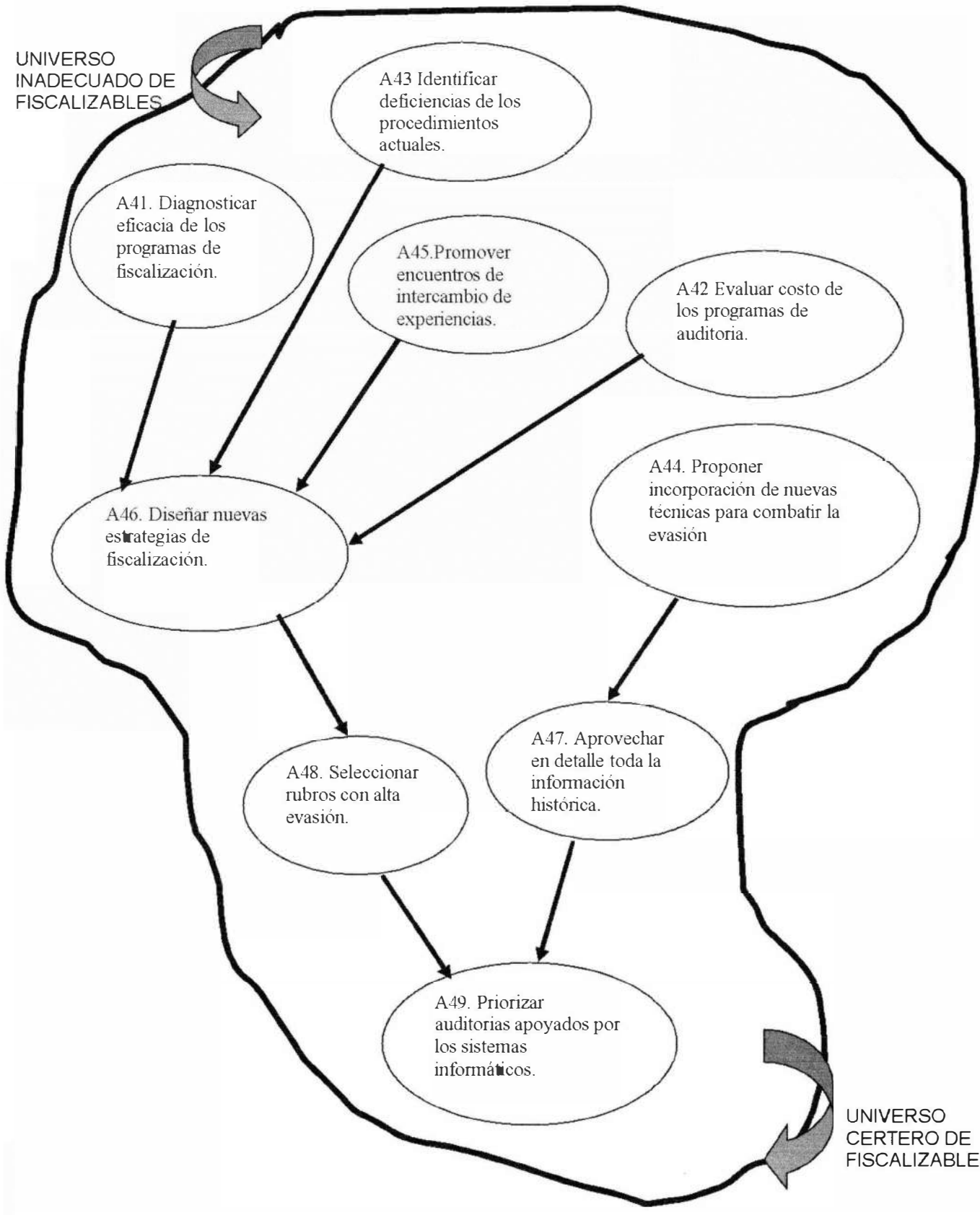
Modelo Conceptual del Analista Tributario



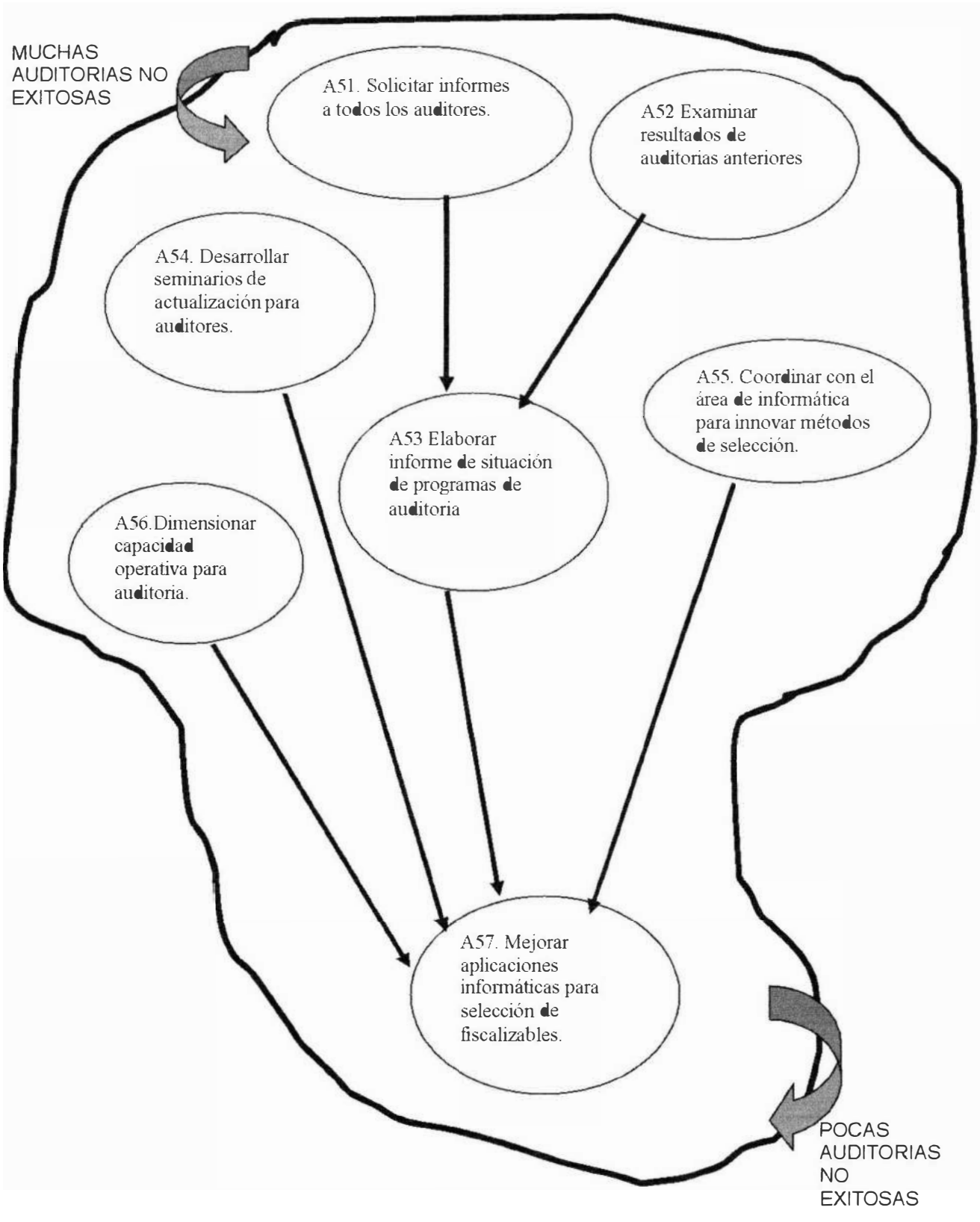
Modelo Conceptual del Analista de Sistemas



Modelo Conceptual del Jefe de Programas de Fiscalización



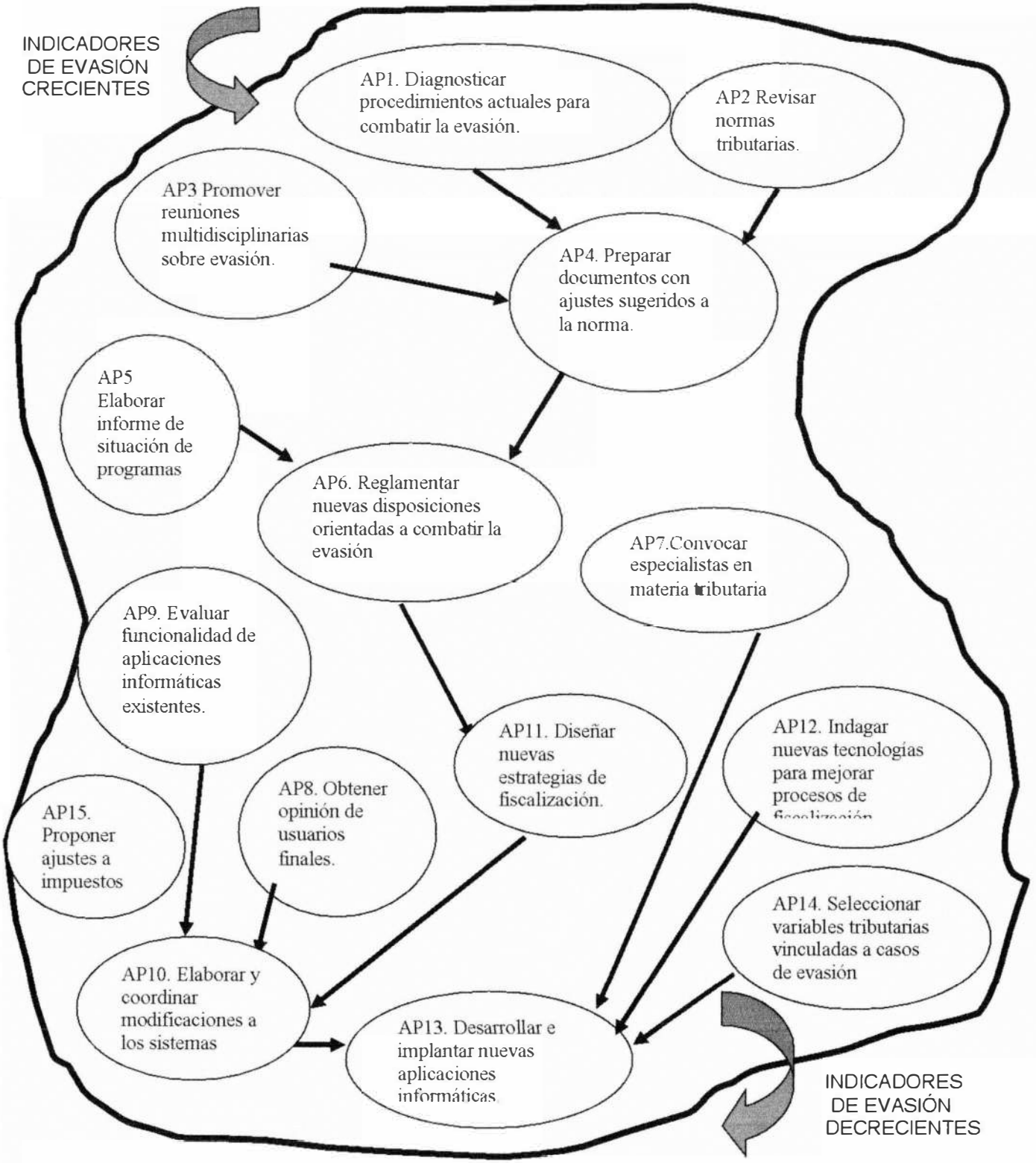
Modelo Conceptual del Auditor



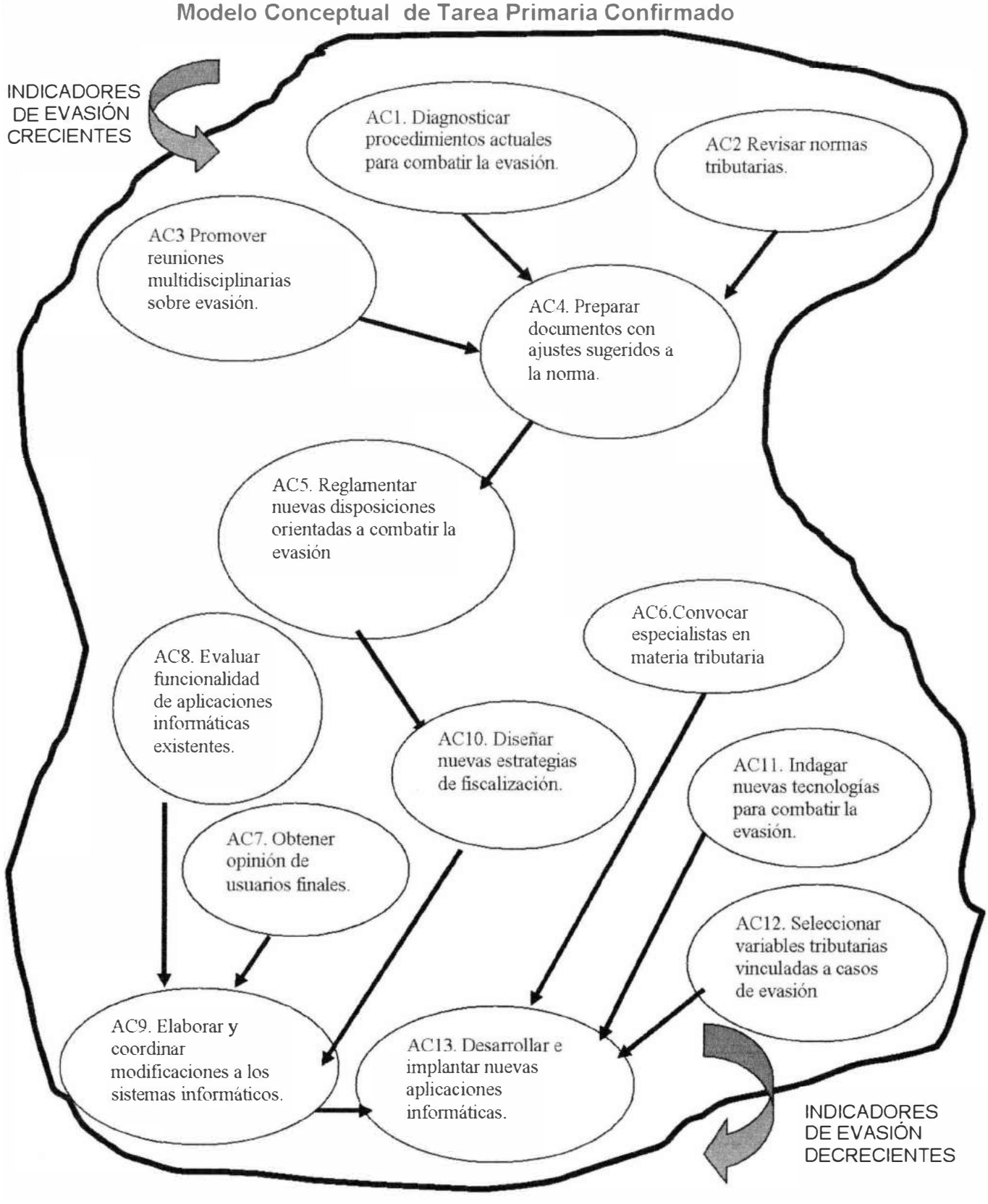
2.2.3 Integración de modelos conceptuales

2.2.3.1 Unificación de modelos conceptuales individuales

Modelo Conceptual de Tarea Primaria

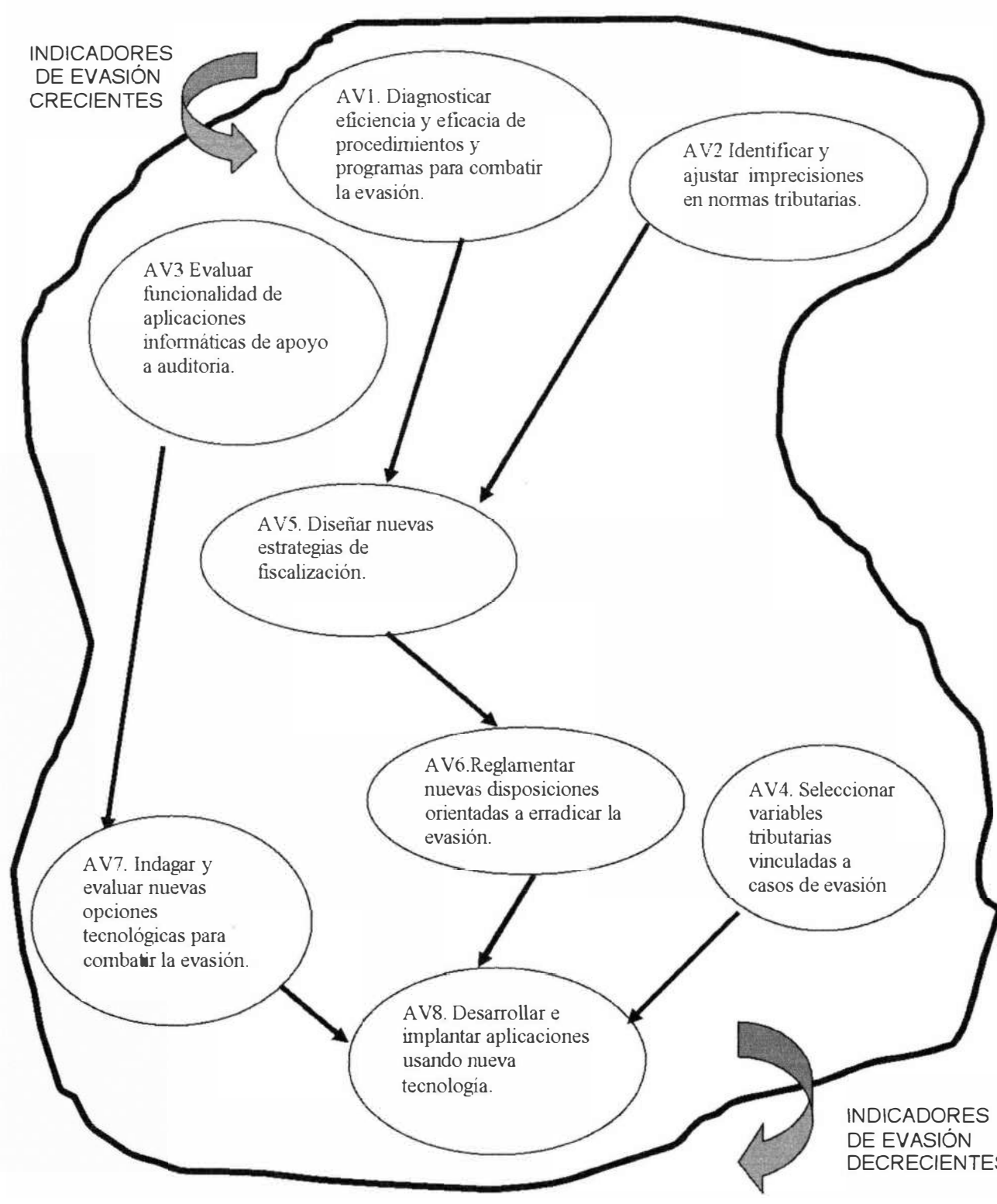


2.2.3.2 Depuración del modelo conceptual unificado



2.2.4 Modelo Conceptual Integrado y Consensuado

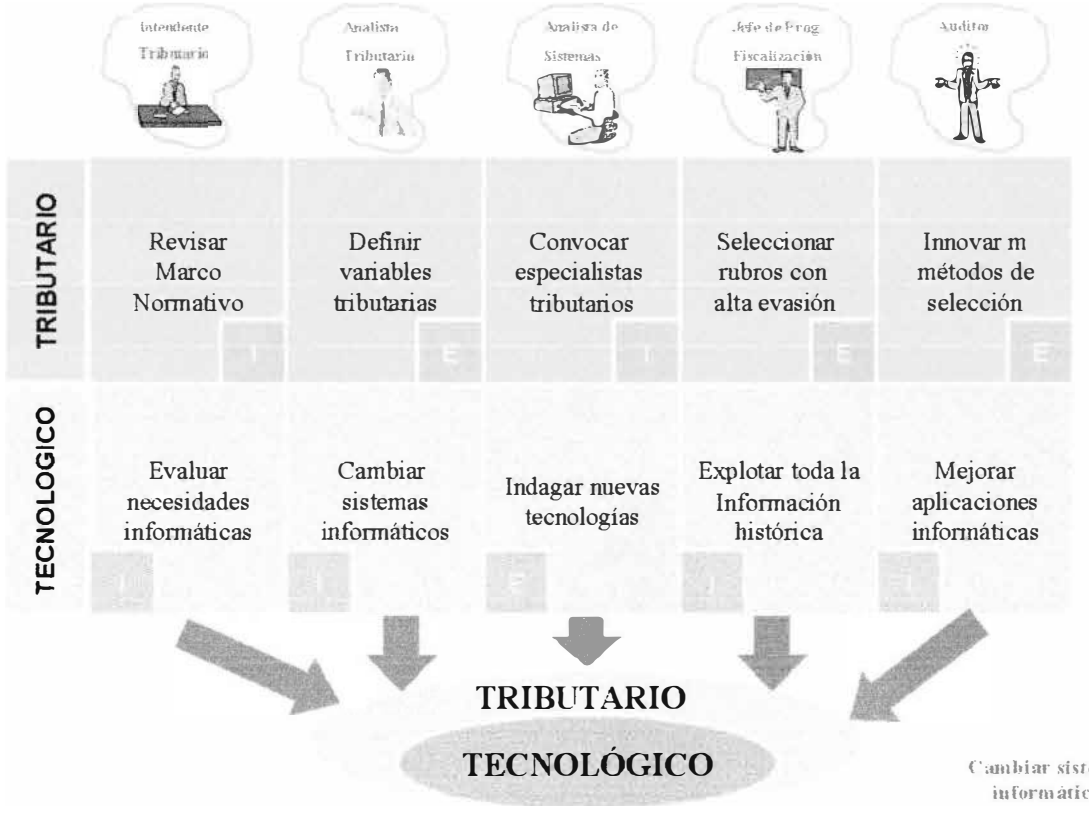
Modelo Conceptual de Tarea Primaria Confirmado y Validado



2.3 Identificación de los componentes del problema

Si bien es la evasión tributaria el problema más álgido, su reiterada referencia en la recopilación de percepciones no ha hecho sino corroborar la hipótesis o premisa de situación problema que como analista ya se tenía Sin embargo, hay otro valor agregado que se desprende de la aplicación parcial de la MSB a la situación problema: la posibilidad de identificar componentes del mismo.

La muestra de observantes (Intendente Tributario, Analista Tributario, Analista de Sistemas, Jefe de Programas de Fiscalización y el Auditor), han incluido en su percepción referencias a temas tributarios y de tecnología de información. Ciertamente con diferentes niveles de intensidad; algunos en forma explícita y otros implícitamente. Esto ha dado lugar a la identificación de dos componentes principales que subyacen a las propuestas para enfrentar la evasión: el componente tributario y el componente tecnológico. Como resultado de examinar sus percepciones se ha asignado una 'E' en el caso haya hecho referencia al componente en forma explícita o una 'I' si solamente lo mencionó implícitamente.



CAPITULO III

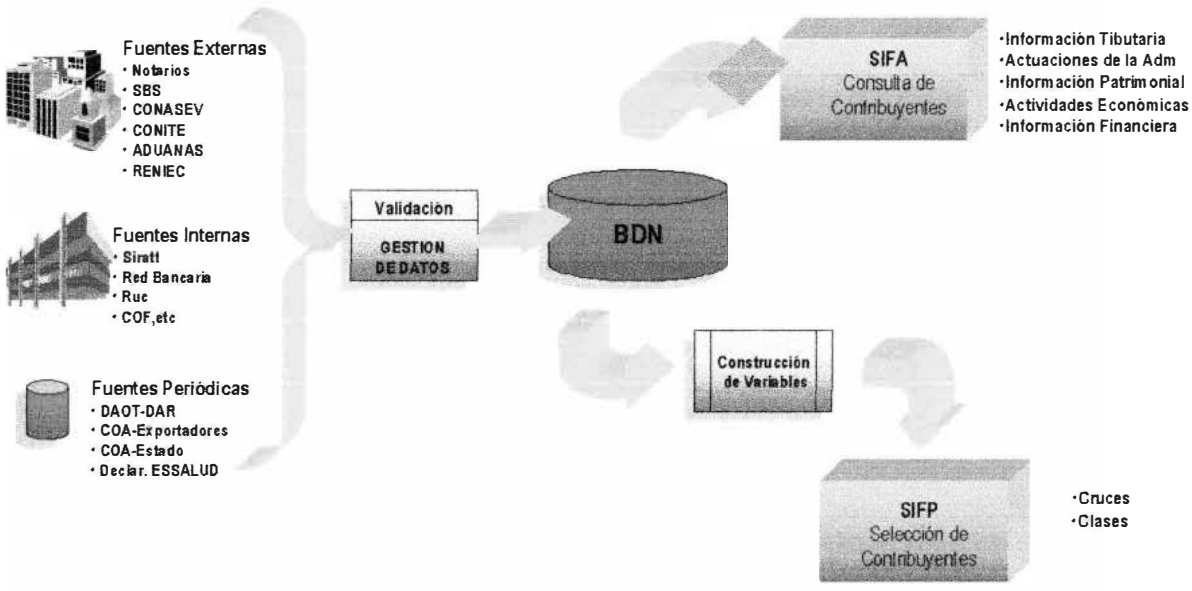
ANÁLISIS DEL MARCO TRIBUTARIO VINCULADO A LA BUSQUEDA DE ESCENARIOS DE EVASION

3.1 Escenario de Fiscalización

El escenario lo representa la interacción de los responsables de la programación de campañas de fiscalización (programadores o usuarios del Sistema Integrado de Fiscalización para Programadores – SIFP) a través de la emisión de Ordenes de Fiscalización (OF) y los auditores que reciben dicho insumo para sus tareas de auditoria a través del Sistema Integrado de Fiscalización para Auditores (SIFA).

3.2 Esquema actual de selección de contribuyentes fiscalizables

El esquema actual para la selección de contribuyentes fiscalizables se enmarca dentro de los programas de fiscalización, que se ejecutan con cierta periodicidad.



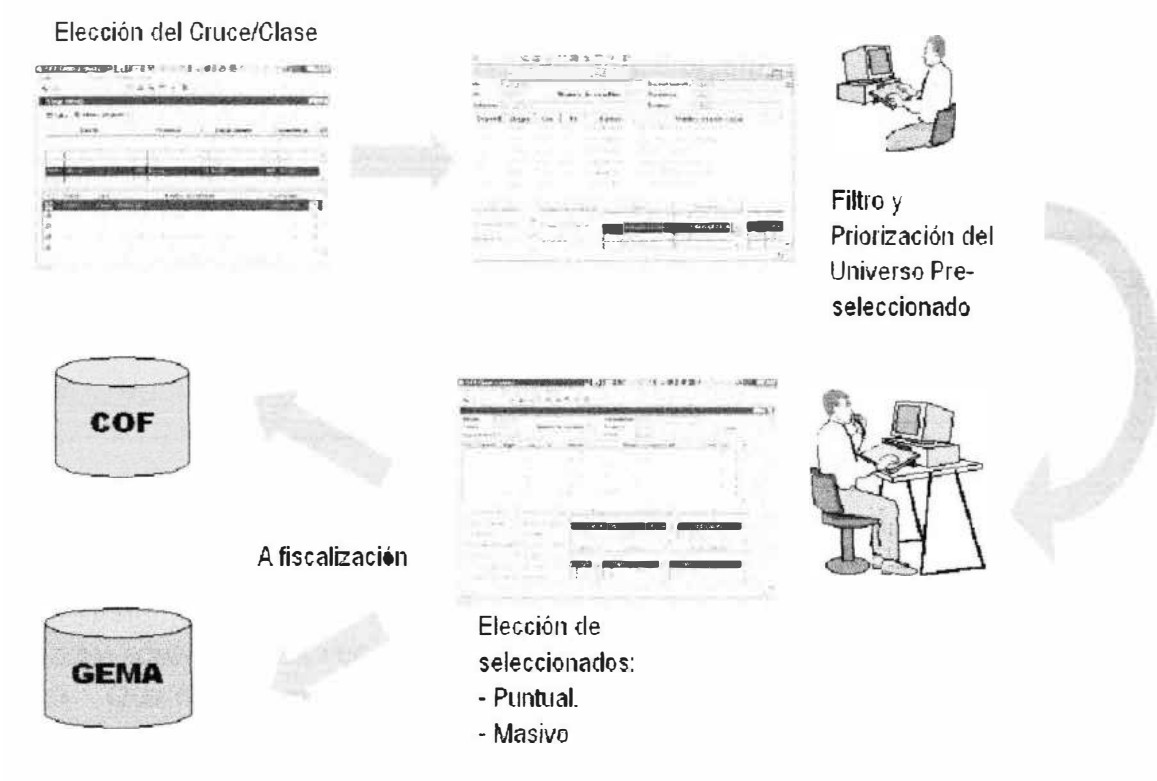
Un programa de fiscalización, es un conjunto de elementos que delimitan las acciones de fiscalización dirigidas a un universo.

Los elementos que configuran un programa de fiscalización son:

- Método de Selección
- Sector o actividad económica
- Dependencias
- Segmento
- Tipo de actuación
- Impuestos

3.2.1 Métodos de selección

Los métodos de selección son los Cruces y Clases que ofrece el SIFP.



CRUCE:

Método de selección que permite confrontar la información procedente de las diversas fuentes de información y presentar los resultados en una matriz de filas y columnas.

Descripción Cruce Régimen General Ventas

Objetivo:

Identificar aquellos contribuyentes que muestran diferencias entre las Ventas Imputadas por DAOT y las Ventas Internas declaradas.

Definición del Universo

- Contribuyentes que cumplen simultáneamente con las siguientes condiciones:

- Que son informados como proveedores en DAOT.
- Que presentaron alguna declaración mensual o anual del impuesto a la renta de 3ra categoría.

CRUCE REGIMEN GENERAL - VENTAS (R01)

	VTA NO DEC						
TOT VTA INTERNAS	]- inf , 0]	]0 , 5m]	]5m , 10m]	]10m - 50m]	]50m , 100m]	]100m, 500m]	]500m
nd	5,000	0	0	0	0	0	0
]0 , 1m]	30,000	40,000	20,000	1,000	500	100	80
]1m , 10,m]	40,000	70,000	60,000	50,000	2,000	300	10
]10m , 100m	35,000	30,000	25,000	35,000	1,500	750	60
]500m , 1M]	20,000	10,000	10,000	10,000	4,000	600	30
] 1M	15,000	8,000	6,000	4,000	3,500	500	10

NOTA IMPORTANTE
EL TAMAÑO DE LOS TRAMOS Y EL NUMERO DE CONTRIBUYENTES, SON DATOS SUPUESTOS

m = miles de soles
M = millones de soles

LA ZONA SOMBREADA ES LA DE MAYOR IMPORTANCIA

- Que no presentaron declaraciones correspondientes al RER o al RUS.

CLASE:

Permite seleccionar contribuyentes que cumplen con un perfil de evasión genérico, el cual se construye mediante la asociación de un determinado número de variables o características.

Ejemplo de un clase:

CONDICIONES	TRAMO
1. Ventas imputadas por terceros.	_ mayor a cero y menor a 5,000 _ mayor a 5,000 y menor a 10,000 _ mayor a 10,000 y menor a 20,000 _ mayor a 20,000 y menor a 50,000 _ mayor a 50,000
2. Indicador de no haber presentado Declaración Anual del Impuesto a la Renta de Tercera	
3. Indicador de Presentar la razón Débito / Crédito < 1	_ mayor a cero y menor a 0.1 _ mayor a 0.1 y menor a 0.2 _ mayor a 0.2 y menor a 0.5 _ mayor a 0.5 y menor a 0.8 _ mayor a 0.8 y menor a 1
4. Indicador de declaraciones falsas.	
5. Indicador de no haber solicitado autorización de impresión de comprobantes de pago.	

Conectores Lógicos:

Y : Indica que la variable a la cual se aplica está encendida.

O : Indica que al menos una de las variables a la cual se aplica está encendida.

N Indica que la variable a la cual se aplica " no está encendida".

E Indica que exclusivamente las variables que tiene este operador "están encendidas".

Tipos de Variables

Variables Tipo 1

Construidas a partir de información de terceros

Variables Tipo 2

Construidas utilizando información del propio contribuyente

Variables Tipo 3

Creadas a partir de antecedentes de la fiscalización

Descripción de la Clase IGV:

Definición del Universo

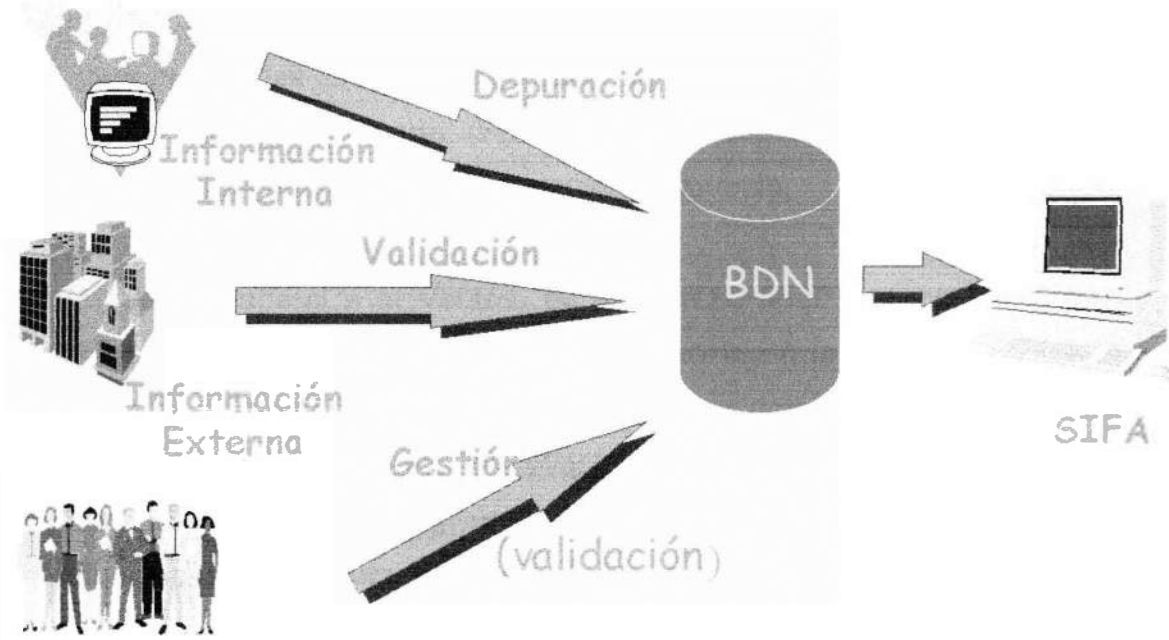
- Contribuyentes que cumplen simultáneamente las siguientes condiciones:
- Que hayan presentado declaraciones mensuales de IGV en el período.
- Que NO hayan presentado declaraciones correspondientes al RUS

Filtrar o excluir del universo, a las entidades cuyas operaciones de IGV no son de trascendencia para los objetivos de este clase.

3.2.2 La Base de Datos Nacional

Representa la gran base de datos centralizada, que concentra información proveniente de los sistemas tributarios transaccionales, así como de información proveniente de fuentes externas.

La información interna y externa es sometida a un proceso de validación y depuración antes de ingresar a la base de datos nacional (BDN). Es de este repositorio que se alimenta el Sistema Integrado de Fiscalización para Auditores (SIFA) para generar el universo de contribuyentes fiscalizables.



3.3 Delimitación del espectro de tributos para proponer un nuevo esquema

3.3.1 Identificación de impuestos con alta índice de evasión

Las manifestaciones de evasión en el sistema de administración tributaria se distribuyen entre los diversos tributos que esta administra. Sin embargo, hay tres que merecen especial atención por el protagonismo que representan en la recaudación: el Impuesto Selectivo al Consumo, el Impuesto a la Renta y el Indice General a las Ventas.

El Impuesto Selectivo al Consumo, viendo desde la perspectiva del porcentaje que representa en el total de la recaudación, no amerita considerarlo como prioridad en el análisis del proyecto del nuevo esquema para su fiscalización. Su control podrá continuar con el sistema convencional, que tampoco es intensivo por los motivos expuestos.

El Impuesto a la Renta, con sus categorías que van desde el Impuesto a la Renta de Primera Categoría hasta el Impuesto a la Renta de Quinta Categoría, tiene un

alto protagonismo en la recaudación tributaria. Sin embargo, las diversas categorías hacen complejo su análisis. Las manifestaciones de evasión alrededor de cada una de las categorías son diversas. Merecería tal vez, dar prioridad a la Renta de Cuarta Categoría (Profesionales Independientes), pero su análisis detallado se ha pospuesto para un siguiente proyecto.

El Impuesto General a la Ventas, es un terreno fértil para buscar patrones de evasión. Su carácter masivo y su alta incidencia en la recaudación tributaria lo hacen especialmente atractivo para examinarlo a detalle. Además, es uno de los impuestos con mayor índice de evasión, y además existe un mayor análisis documentado sobre los programas de fiscalización y auditorias realizadas, que conducen a darle un protagonismo en la lucha contra la evasión.

Lo anteriormente mencionado sobre niveles de evasión ha sido recogido de la interacción con especialistas tributarios. No se ha pretendido reformular modelos tributarios, dado que no está contemplado en el alcance de este proyecto. Se seleccionará por lo tanto el IGV como el impuesto que se analizará con mas detalle en adelante.

Se muestra a continuación una breve semblanza del IGV, para un mejor entendimiento en la identificación de variables relacionadas.

3.3.2 El Impuesto General a las Ventas (IGV)

El Impuesto General a las Ventas es a la fecha, quizá, el tributo más importante en el ámbito nacional porque:

- Grava la transferencia de bienes y la prestación de servicios en el país, con lo que incide directamente en los precios relativos de los mismos, así como en la economía nacional y en el comercio internacional.

- El Impuesto General a las Ventas representa para el Estado un extraordinario recurso financiero, pues es el impuesto con mayor recaudación en el país.

Por ello el marco teórico y, las normas de este impuesto, constituye parte importante del estudio y análisis de la tributación nacional.

3.3.2.1 Operaciones Gravadas

3.3.2.1.1 Venta en el país de bienes muebles.

Está gravada la venta de bienes muebles ubicados en el territorio nacional y realizada en cualquiera de las etapas del ciclo de producción y distribución.

3.3.2.1.2 Prestación o utilización de servicios en el país.

Con relación a la prestación de servicios en el país, debe entenderse que se encuentra gravada independientemente del lugar en que se pague o se perciba la contraprestación, del lugar donde se celebre el contrato, y siempre que el sujeto que lo presta se encuentre domiciliado en él.

3.3.2.1.3 Contratos de construcción.

Están gravados los contratos de construcción que se ejecuten en el territorio nacional, cualquiera sea su denominación, sujeto que lo realice, lugar de celebración del contrato o de percepción de los ingresos.

3.3.2.1.4 Primera venta de inmuebles ubicados en el país

Está gravada la primera venta de inmuebles ubicados en el territorio nacional, que realicen los constructores de los mismos.

Sobre el particular, se considera constructor a cualquier persona que se dedique en forma habitual a la venta de inmuebles, construidos totalmente por ella, o en parte o toda, por un tercero para dicha persona.

3.3.2.1.5 Importación de bienes

Se encuentra gravada la importación o internación de bienes, cualquiera sea el sujeto que la realice.

3.3.2.2 Sujetos del impuesto

Las empresas, entidades o personas que realicen una operación gravada y trasladen el impuesto en cualquiera de las etapas de la cadena de producción o comercialización, aplicándolo y pagándolo sobre el valor agregado en cada etapa, son consideradas sujetos del impuesto, sea en calidad de contribuyentes o de responsables solidarios. En todos los casos que la Ley señala, tienen la obligación de declarar y pagar el impuesto en los plazos establecidos por la SUNAT. No obstante, cabe indicar que el Impuesto General a la Ventas es asumido por el consumidor o usuario final, al momento de cancelar el precio de venta de un bien o servicio.

3.3.2.2.1 Contribuyentes

Son contribuyentes del IGV:

a) Las personas naturales, jurídicas, sociedades conyugales que ejerzan la opción prevista en la legislación del Impuesto a la Renta, sucesiones indivisas,

sociedades irregulares, patrimonios fideicometidos de sociedades titulizadoras, los fondos mutuos de inversión en valores y los fondos de inversión, que realicen actividad empresarial.

b) También son contribuyentes, la comunidad de bienes, los consorcios, joint ventures u otras formas de contratos de colaboración empresarial, que lleven contabilidad independiente

c) En el caso de comisionistas, es sujeto del impuesto la persona por cuya cuenta se realiza la venta (Comitente).

d) Tratándose de entrega de bienes en consignación y otras formas similares en las que la venta se realiza por cuenta propia, son sujetos del IGV tanto el que entrega el bien como el consignatario.

3.3.3 Selección de variables relevantes vinculadas al IGV

Para la búsqueda de patrones de evasión en el IGV, es vital identificar que variables son relevantes para su determinación. Mostraremos a continuación un subconjunto de variables con su descripción, definidas por especialistas y utilizadas en el esquema actual de selección de contribuyentes.

Debe precisarse que la elección de estas variables respecto a otras que también se han definido, ha sido resultado de la aplicación de dos criterios: por un lado la opinión del especialista tributario (lo cual tomamos como un legado) y por otro lado la disponibilidad de datos en la Base de Datos Nacional para esas variables, las cuales aparecen como campos en tablas.

En términos generales, las variables seleccionadas giran alrededor de montos de compras (nacionales e importadas), ventas (mercado interno y exportación) y cantidad de presentaciones de formularios. El tramo de medición se considera anual.

Código	Descripción
CIC	Código de Identificación del Contribuyente
mto_comimp_anu	Compras nacionales imputadas por terceros (DAOT, COA-estado o COA-exportadores). Para cada contribuyente informado como cliente (DAOT, COA-estado o COA-exportadores)
mto_comimpu_tot	Suma del monto de las Compras Imputadas mediante el DAOT más las Compras Importadas según ADUANAS.
mto_comp_totales	Monto anual del total de la base imponible de las Compras Gravadas y Compras No Gravadas, declaradas en el IGV (internas + importaciones)
mto_dif_pos_prov	Acumula para cada informante del DAOT las diferencias positivas (Compras según contribuyente en proceso por DAOT - Compras según proveedor por DAOT), siempre que sus proveedores sean informantes del DAOT en el ejercicio en proceso
mto_ing_nd	Para cada contribuyente informado como proveedor DAOT , contiene el monto de las ventas no declaradas (diferencia de ventas imputadas anuales por la DAOT- (total anual de ventas internas declaradas + total de rentas brutas de 1ra categoría declaradas + total de renta bruta de 2da categoría declarada + total de ingresos atribuidos)
mto_vtaimp_anu	Para todo contribuyente informado como proveedor DAOT, contiene la suma de las ventas que le han sido imputadas en el período de cruce.
mto_vtas_net_gra	Monto de Ventas Netas Internas Gravadas anualizadas. Para cada contribuyente contiene el valor de la suma del valor de las "Ventas Netas Internas Gravadas mensuales" de los meses comprendidos dentro del período de cruce, de los formularios 118, 119, 219, 620 y 621.
mto_vtas_totales	Ventas Totales anuales declaradas en el IGV. Por contribuyente sumar los Montos Totales de Ventas Internas

	mas el Monto de Exportación, del periodo de proceso.
ctd_decla_igv	Número de declaraciones mensuales de IGV que un contribuyente presenta en el año y que no son rectificatorias.
ctd_nocon_cl	Cuenta todas las veces que al contribuyente en proceso se le ha encontrado infracción que puede generar cierre de local, para todas las infracciones cuyas actas se emitieron en el período de cruce.
ctd_rect_igv	El número de veces que un contribuyente rectifica su declaración de IGV CTA PROPIA en un año.

CAPITULO IV

EXPLORACIÓN DE TECNOLOGÍAS EMERGENTES PARA USUFRUCTUAR LA INFORMACION TRIBUTARIA.

4.1 Métodos convencionales para el aprovechamiento de la información de los sistemas tributarios.

En la administración tributaria peruana, la gran mayoría de usuarios usa intensivamente y casi exclusivamente los sistemas transaccionales, principalmente por la recargada labor de nivel operativo que desempeñan. Las decisiones se toman basados en conclusiones aproximadas obtenidas de examinar reportes inmensos que se obtienen a partir de los sistemas operacionales. Solo un reducido número de usuarios usa intensivamente sistemas que acceden a base de datos con información resumida o consolidada.

4.2 Opciones tecnológicas de vanguardia para el tratamiento de datos

En la actualidad las grandes organizaciones tienen almacenados en sus bases de datos una cantidad inmensa de información. En muchos casos, dicha información histórica, almacenada a través de los años, no es aprovechada mas que en un pequeño porcentaje, por no disponer los usuarios de los medios adecuados para su explotación.

Hace buen tiempo irrumpieron en el mercado de tecnologías de información varias propuestas para el aprovechamiento de la ingente cantidad de datos almacenados en los repositorios de las organizaciones. Sin embargo, a mi entender,

particularmente dos se han ido consolidando en el mercado de las bases de datos: el DataWarehouse y el Data Mining.

La primera irrumpió comercialmente antes que la segunda, y existe ya muchos resultados en varias organizaciones, tanto favorables como desfavorables. La segunda, es mas reciente en términos comerciales, y las aplicaciones existentes son por ahora muy pocas, sin embargo promete bastante, teniendo en cuenta que el insumo básico que requiere (data histórica almacenada en detalle) existe en la mayoría de organizaciones.

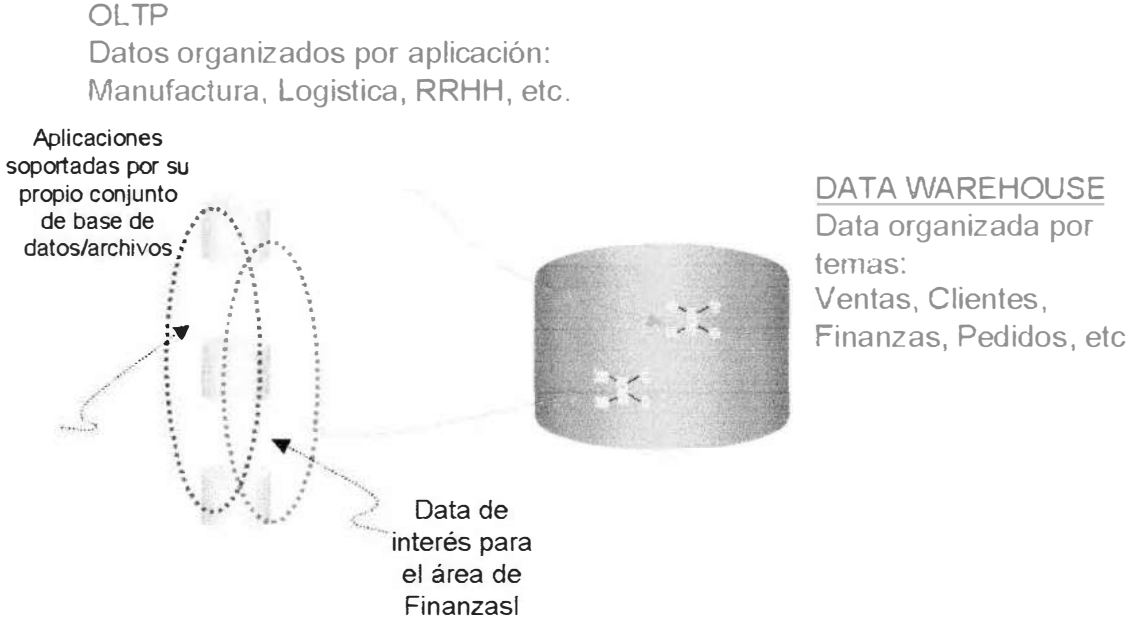
4.2.1 Data Warehouse

4.2.1.1 Definición de Data Warehouse

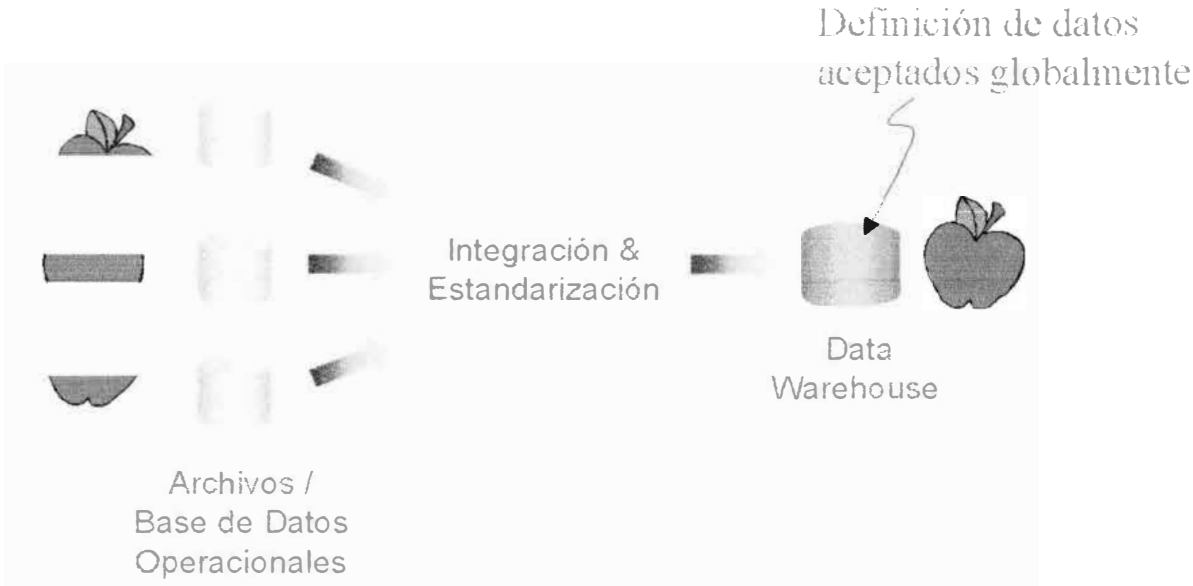
"Una colección de datos, Orientado a un tema, Integrado, Variante en el Tiempo, No volatil, para servir de apoyo en la toma de decisiones." - W. H. Inmon.

4.2.1.2 Características de un Data Warehouse

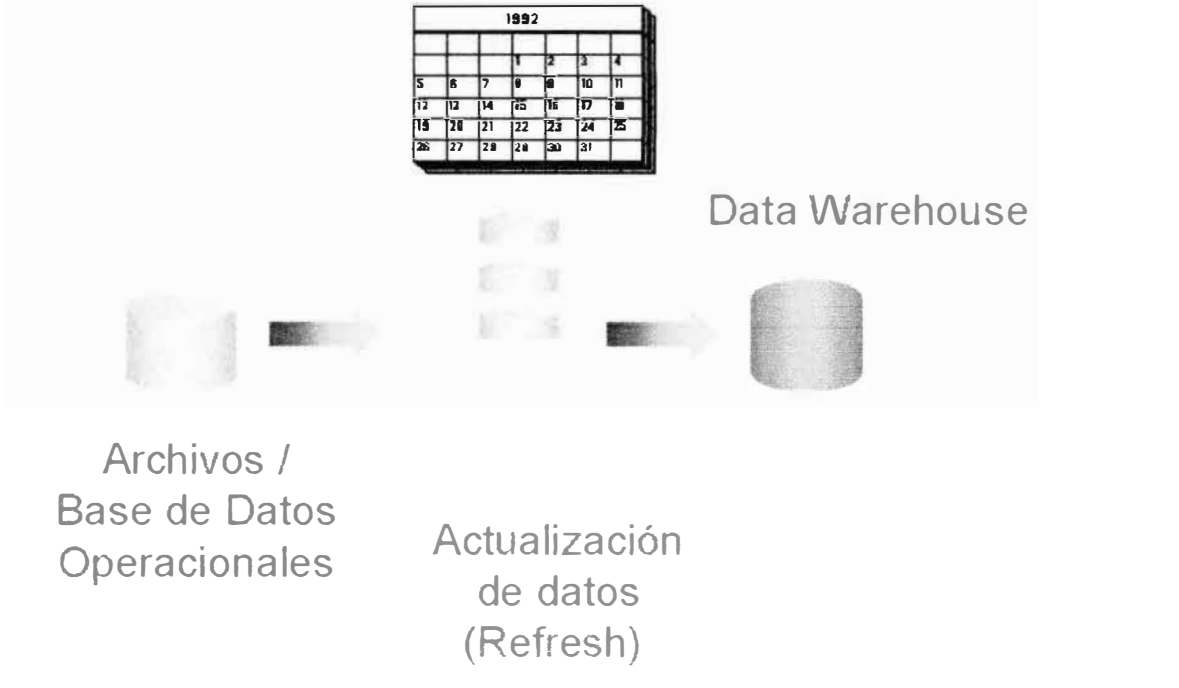
Es orientado a un tema:



Es integrado:



Es variante en el tiempo:



Es no volátil:

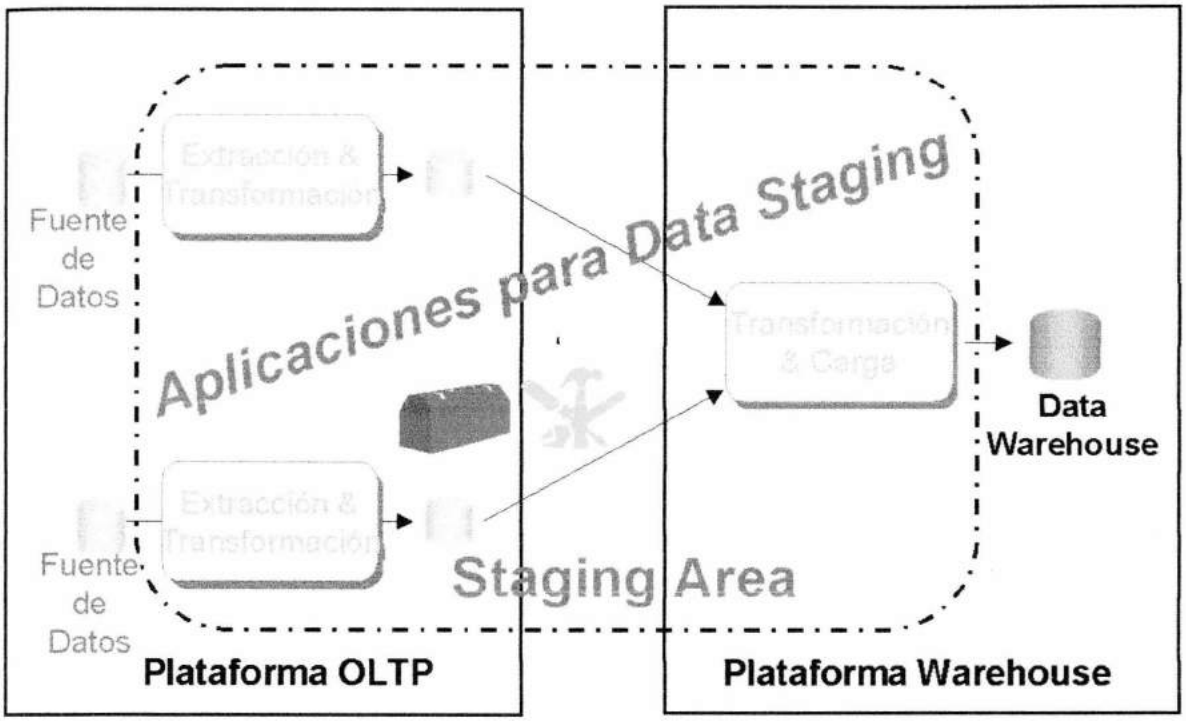
4.2.1.3 Construcción de un Data Warehouse (extracción, transformación)

Para construir un Data Warehouse se extrae información de los sistemas operacionales (sistemas transaccionales) y previa a su incorporación al Warehouse se realiza la transformación necesaria. Muchas veces esta etapa involucra cambiar de formato a algunos datos, depurar inconsistencias en los datos extraídos y estandarizar los datos. Toda esta labor no se realiza en el warehouse, sino en un área intermedia de trabajo temporal, denominado "staging area". Este lugar intermedio, sirve de escenario temporal para las tareas de transformación.

El Staging Area contiene tablas normalizadas y archivos usados para:

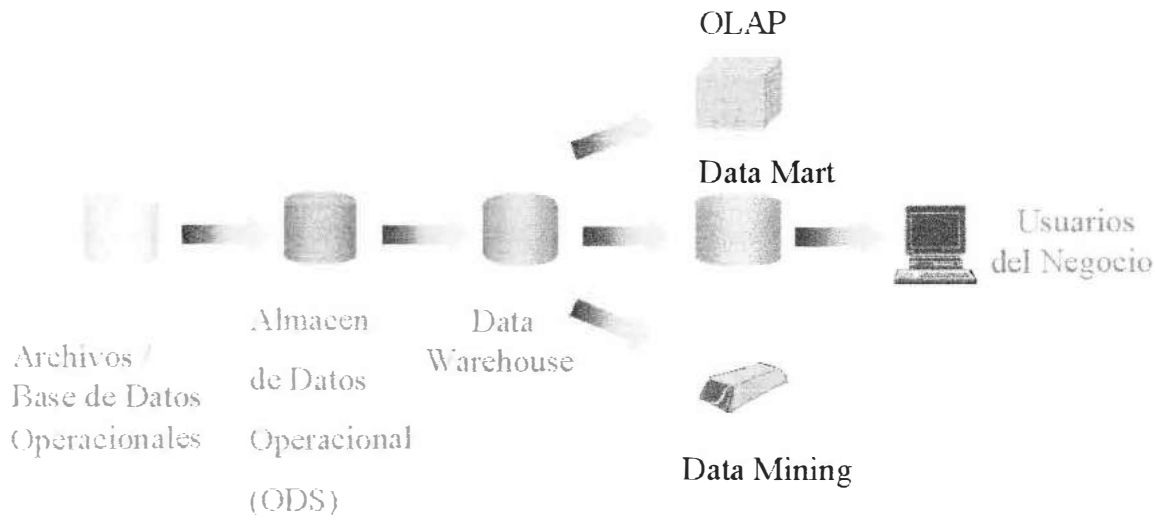
- Integración de datos transaccionales (OLTP data).
- Validación.
- Búsqueda y Referencias cruzadas.

Los Usuarios de Negocio siempre acceden directamente la base de datos dimensional.



4.2.1.4 Opciones de explotación del DataWarehouse (OLAP, DATA MART, DATA MINING)

La existencia de un Data Warehouse en una organización abre interesantes opciones para la explotación de los datos que tiene almacenado. Una posibilidad es hacer consultas en linea en múltiples dimensiones (OLAP). Otra alternativa es generar Data Marts (almacenes de datos para un área específica de negocio) y la tercera es realizar Data Mining para la búsqueda de patrones de información ocultos.



4.2.2 Data Mining

Data Mining, ***la extracción de información oculta y predecible de grandes bases de datos***, es una poderosa tecnología nueva con gran potencial para ayudar a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse). Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información (knowledge-driven). Los ***análisis prospectivos*** automatizados ofrecidos por un producto así van más allá de los eventos pasados provistos por herramientas retrospectivas típicas de sistemas de soporte de decisión. Las herramientas de Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas y a los cuales los usuarios de esta información casi no están dispuestos a aceptar. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas.

Muchas compañías ya colectan y refinan cantidades masivas de datos. Las técnicas de Data Mining pueden ser implementadas rápidamente en plataformas ya existentes de software y hardware para acrecentar el valor de las fuentes de información existentes y pueden ser integradas con nuevos productos y sistemas

pues son traídas en línea (on-line). Una vez que las herramientas de Data Mining fueron implementadas en computadoras cliente servidor de alta performance o de procesamiento paralelo, pueden analizar bases de datos masivas para brindar respuesta a preguntas tales como, "¿Cuáles clientes tienen más probabilidad de responder al próximo mailing promocional, y por qué? y presentar los resultados en formas de tablas, con gráficos, reportes, texto, hipertexto, etc.

¿Cómo Trabaja el Data Mining?

¿Cuán exactamente es capaz Data Mining de decirle cosas importantes que usted desconoce o que van a pasar? La técnica usada para realizar estas hazañas en Data Mining se llama *Modelado*. Modelado es simplemente el acto de construir un modelo en una situación donde usted conoce la respuesta y luego la aplica en otra situación de la cual desconoce la respuesta. Por ejemplo, si busca un galeón español hundido en los mares lo primero que podría hacer es investigar otros tesoros españoles que ya fueron encontrados en el pasado. Notaría que esos barcos frecuentemente fueron encontrados fuera de las costas de Bermuda y que hay ciertas características respecto de las corrientes oceánicas y ciertas rutas que probablemente tomara el capitán del barco en esa época. Usted nota esas similitudes y arma un modelo que incluye las características comunes a todos los sitios de estos tesoros hundidos. Con estos modelos en mano sale a buscar el tesoro donde el modelo indica que en el pasado hubo más probabilidad de darse una situación similar. Con un poco de esperanza, si tiene un buen modelo, probablemente encontrará el tesoro.

Este acto de construcción de un modelo es algo que la gente ha estado haciendo desde hace mucho tiempo, seguramente desde antes del auge de las computadoras y de la tecnología de Data Mining. Lo que ocurre en las computadoras, no es muy diferente de la manera en que la gente construye modelos. Las computadoras son cargadas con mucha información acerca de una variedad de situaciones donde una respuesta es conocida y luego el software de Data Mining en la computadora debe correr a través de los datos y distinguir las

características de los datos que llevarán al modelo. Una vez que el modelo se construyó, puede ser usado en situaciones similares donde usted no conoce la respuesta.

Si alguien le dice que tiene un modelo que puede predecir el uso de los clientes, ¿Cómo puede saber si es realmente un buen modelo? La primera cosa que puede probar es pedirle que aplique el modelo a su base de clientes - donde usted ya conoce la respuesta. Con Data Mining, la mejor manera para realizar esto es dejando de lado ciertos datos para aislarlos del proceso de Data Mining. Una vez que el proceso está completo, los resultados pueden ser comparados contra los datos excluidos para confirmar la validez del modelo. Si el modelo funciona, las observaciones deben mantenerse para los datos excluidos.

4.2.2.1 Fundamentos de Data Mining

Las técnicas de Data Mining son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. Data Mining toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva. Data Mining está listo para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:

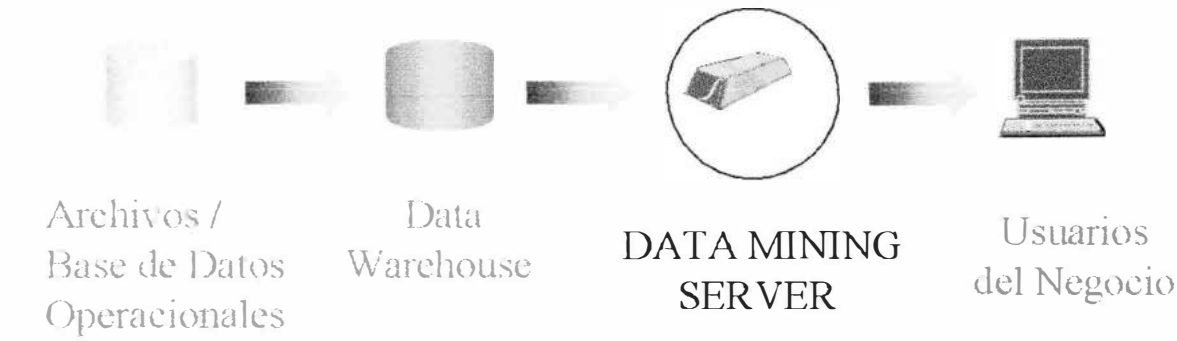
- Recolección masiva de datos
- Potentes computadoras con multiprocesadores
- Algoritmos de Data Mining

Las bases de datos comerciales están creciendo a un ritmo sin precedentes. Un estudio del META GROUP sobre los proyectos de Data Warehouse encontró que

el 19% de los que contestaron están por encima del nivel de los 50 Gigabytes, mientras que el 59% espera alcanzarlo en el segundo trimestre de 1999. En algunas industrias, tales como ventas al por menor (retail), estos números pueden ser aún mayores. MCI Telecommunications Corp. cuenta con una base de datos de 3 terabytes + 1 terabyte de índices y overhead corriendo en MVS sobre IBM SP2. La necesidad paralela de motores computacionales mejorados puede ahora alcanzarse de forma más costo - efectiva con tecnología de computadoras con multiprocesamiento paralelo. Los algoritmos de Data Mining utilizan técnicas que han existido por lo menos desde hace 10 años, pero que sólo han sido implementadas recientemente como herramientas maduras, confiables, entendibles que consistentemente son más performantes que métodos estadísticos clásicos.

En la evolución desde los datos de negocios a información de negocios, cada nuevo paso se basa en el previo. Por ejemplo, el acceso a datos dinámicos es crítico para las aplicaciones de navegación de datos (drill through applications), y la habilidad para almacenar grandes bases de datos es crítica para Data Mining.

Los componentes esenciales de la tecnología de Data Mining han estado bajo desarrollo por décadas, en áreas de investigación como estadísticas, inteligencia artificial y aprendizaje de máquinas. Hoy, la madurez de estas técnicas, junto con los motores de bases de datos relacionales de alta performance, hicieron que estas tecnologías fueran prácticas para los entornos de data warehouse actuales.



4.2.2.2 La Evolución hacia el Data Mining

PASO DE EVOLUCION	PREGUNTAS	TECNOLOGÍAS	CARACTERISTICAS
Recolección de data (1960s)	"Cual fue mi total de ventas en los últimos 5 años?"	Computadoras, cintas, discos.	Entrega de datos estática.
Acceso a la Data (1980s)	"Cuales fueron las ventas unitarias por distrito el pasado mes de Marzo?"	Base de Datos Relacionales (RDBMS), SQL, ODBC.	Entrega de datos dinámica a distintos niveles simples.
Data Warehousing y Soporte de Decisiones (1990s)	"Cuáles fueron las ventas unitarias en el distrito de Miraflores el pasado mes de Marzo?. Drill Down para obtener Miraflores.	Tecnología OLAP, Base de Datos Multidimensionales, Data Warehouses.	Entrega de datos dinámica en múltiples niveles de jerarquía.
Data Mining (aún en desarrollo)	"Que podrá pasar con las ventas unitarias en Miraflores el próximo mes? Por qué?"	Algoritmos avanzados, multiprocesadores, base de datos masivas.	Entrega de la información proactiva y predictivamente.

4.2.2.3 Alcance del Data Mining

El nombre de Data Mining deriva de las similitudes entre buscar valiosa información de negocios en grandes bases de datos - por ej.: encontrar información de la venta de un producto entre grandes montos de Gigabytes almacenados - y minar una montaña para encontrar una veta de metales valiosos. Ambos procesos requieren examinar una inmensa cantidad de material, o investigar inteligentemente hasta encontrar exactamente donde residen los valores. Dadas bases de datos de suficiente tamaño y calidad, la tecnología de Data Mining puede generar nuevas oportunidades de negocios al proveer estas capacidades:

Predicción automatizada de tendencias y comportamientos. Data Mining automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos. Un típico ejemplo de problema predecible es el marketing apuntado a objetivos (targeted marketing). Data Mining usa datos en mailing promocionales anteriores para identificar posibles objetivos para maximizar los resultados de la inversión en futuros mailing. Otros problemas predecibles incluyen pronósticos de problemas financieros futuros y otras formas de incumplimiento, e identificar segmentos de población que probablemente respondan similarmente a eventos dados.

Descubrimiento automatizado de modelos previamente desconocidos. Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso. Otros problemas de descubrimiento de modelos incluye detectar transacciones fraudulentas de tarjetas de créditos e identificar *datos anormales* que pueden representar errores de tipeado en la carga de datos.

Las técnicas de Data Mining pueden redituar los beneficios de automatización en las plataformas de hardware y software existentes y puede ser implementadas en

sistemas nuevos a medida que las plataformas existentes se actualicen y nuevos productos sean desarrollados. Cuando las herramientas de Data Mining son implementadas en sistemas de procesamiento paralelo de alta performance, pueden analizar bases de datos masivas en minutos. Procesamiento más rápido significa que los usuarios pueden automáticamente experimentar con más *modelos* para entender datos complejos. Alta velocidad hace que sea práctico para los usuarios analizar inmensas cantidades de datos. Grandes bases de datos, a su vez, producen mejores predicciones.

Las bases de datos pueden ser grandes tanto en profundidad como en ancho:

Más columnas. Los analistas muchas veces deben limitar el número de variables a examinar cuando realizan análisis manuales debido a limitaciones de tiempo. Sin embargo, variables que son descartadas porque parecen sin importancia pueden proveer información acerca de modelos desconocidos. Un Data Mining de alto rendimiento permite a los usuarios explorar toda la base de datos, sin preseleccionar un subconjunto de variables.

Más filas. Muestras mayores producen menos errores de estimación y desvíos, y permite a los usuarios hacer inferencias acerca de pequeños pero importantes segmentos de población.

4.2.2.4 Una arquitectura para Data Mining

Para aplicar mejor estas técnicas avanzadas, éstas deben estar totalmente integradas con el data warehouse así como con herramientas flexibles e interactivas para el análisis de negocios. Varias herramientas de Data Mining actualmente operan fuera del warehouse, requiriendo pasos extra para extraer, importar y analizar los datos. Además, cuando nuevos conceptos requieren implementación operacional, la integración con el warehouse simplifica la aplicación de los resultados desde Data Mining. El Data warehouse analítico resultante puede ser aplicado para mejorar procesos de negocios en toda la

organización, en áreas tales como manejo de campañas promocionales, detección de fraudes, lanzamiento de nuevos productos, etc.

El punto de inicio ideal es un data warehouse que contenga una combinación de datos de seguimiento interno de todos los clientes junto con datos externos de mercado acerca de la actividad de los competidores. Información histórica sobre potenciales clientes también provee una excelente base para proyecciones. Este warehouse puede ser implementado en una variedad de sistemas de bases relacionales y debe ser optimizado para un acceso a los datos flexible y rápido.

Un servidor multidimensional OLAP permite que un modelo de negocios más sofisticado pueda ser aplicado cuando se navega por el data warehouse. Las estructuras multidimensionales permiten que el usuario analice los datos de acuerdo a como quiera mirar el negocio - resumido por línea de producto, u otras perspectivas claves para su negocio. El servidor de Data Mining debe estar integrado con el data warehouse y el servidor OLAP para insertar el análisis de negocios directamente en esta infraestructura. Un avanzado, metadata centrado en procesos define los objetivos del Data Mining para resultados específicos tales como manejos de campaña, prospecting, y optimización de promociones. La integración con el data warehouse permite que decisiones operacionales sean implementadas directamente y monitoreadas. A medida que el data warehouse crece con nuevas decisiones y resultados, la organización puede "minar" las mejores prácticas y aplicarlas en futuras decisiones.

Este diseño representa una transferencia fundamental desde los sistemas de soporte de decisión convencionales. Más que simplemente proveer datos a los usuarios finales a través de software de consultas y reportes, el servidor de Análisis Avanzado aplica los modelos de negocios del usuario directamente al warehouse y devuelve un análisis proactivo de la información más relevante. Estos resultados mejoran los metadatos en el servidor OLAP proveyendo una estrato de metadatos que representa una vista fraccionada de los datos. Generadores de reportes, visualizadores y otras herramientas de análisis pueden

ser aplicadas para planificar futuras acciones y confirmar el impacto de esos planes.

4.2.2.5 Técnicas que subyacen al Data Mining

a) Arboles de decisión: estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. Métodos específicos de árboles de decisión incluyen Arboles de Clasificación y Regresión (CART: Classification And Regression Tree) y Detección de Interacción Automática de Chi Cuadrado (CHAI: Chi Square Automatic Interaction Detection)

b) Regla de inducción: la extracción de reglas if-then de datos basados en significado estadístico.

c) Análisis Estadístico: Regresión, Ji Cuadrado, etc.

d) Análisis Cluster: Permite clasificar una población en un número determinado de grupos, en base a semejanzas y desemejanzas de perfiles existentes entre los diferentes componentes de dicha población.

e) Análisis Discriminante: Método de clasificación de individuos en grupos que previamente se han establecido, y que permite encontrar la regla de clasificación de los elementos de estos grupos, y por tanto identificar cuáles son las variables que mejor definen la pertenencia al grupo.

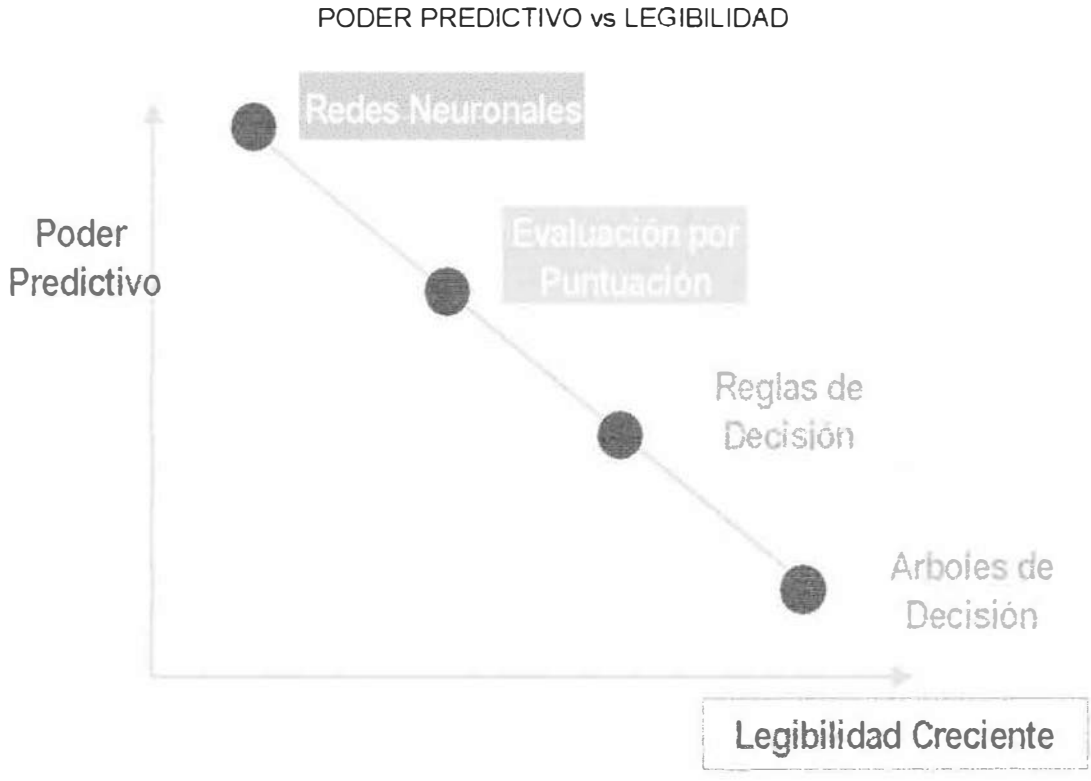
f) Método del vecino más cercano: una técnica que clasifica cada registro en un conjunto de datos basado en una combinación de las clases del/de los *k* registro (s) más similar/es a él en un conjunto de datos históricos (donde *k* ³ 1). Algunas veces se llama la técnica del vecino *k*-más cercano.

g) Series temporales: Es el conocimiento de una variable a través del tiempo y bajo el supuesto de no ocurrencia de cambios estructurales, poder realizar predicciones. Suelen basarse en el estudio de ciclos, tendencias, estacionalidades, que se diferencian por el ámbito del tiempo abarcado, para por composición obtener la serie original.
Un enfoque hibrido, en los que la serie se puede explicar no sólo en función del tiempo sino como combinación de otras variables de entorno más estables, y por tanto más facilmente predecibles.

h) Lógica Difusa (fuzzy logic): Es una generalización del concepto de estadística. La estadística clásica se basa en la teoría de probabilidades, a su vez ésta en la técnica conjuntista, en la que la relación de pertenencia a un conjunto es dicotómica (el 2 es par o no es par).
En la noción de conjunto borroso, como aquel en que la pertenencia tiene cierta graduación (¿Un día a 20°C es caluroso?), dispondremos de una estadística más amplia y con resultados más cercanos al modo de razonamiento humano.

i) Algoritmos genéticos: técnicas de optimización que usan procesos tales como combinaciones genéticas, mutaciones y selección natural en un diseño basado en los conceptos de evolución.

j) Redes neuronales artificiales: modelos predecible no-lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.



4.2.3 Elección de la tecnología para afrontar la evasión tributaria

Data Mining como tecnología para la explotación de grandes bases de datos con propósitos de predicción es una gran promesa siempre que se seleccione adecuadamente la técnica que mejor se adapte a las necesidades particulares del escenario que se quiere predecir.

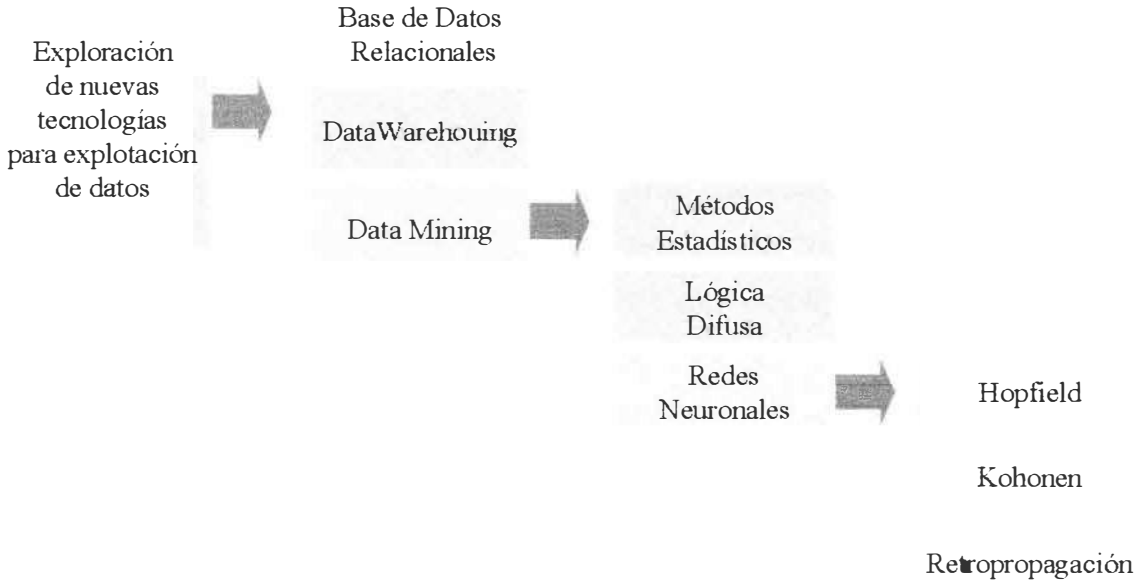
Luego de examinar las virtudes y debilidades de cada una de las técnicas, y teniendo en cuenta la exigencia de tener un alto nivel de certeza en la tarea de predicción de los potenciales evasores, se ha seleccionado la técnica redes neuronales.

Cabe precisar que otro de los argumentos a favor de las redes neuronales, es que su implementación no requiere la definición de reglas que involucren conocimiento tributario dentro del núcleo de procesamiento. Todas las otras técnicas, en mayor

o menor grado, requieren que un experto en el negocio construye un modelo estadístico o una sofisticada ecuación que relacione una variable dependiente (predecida) con variables independientes, o en otros casos que el experto define un conjunto de reglas a través de los cuales se pueda inferir un posible resultado con un margen de certeza.

La implementación de un Data Mining basado en Redes Neuronales requiere mas bien un gran despliegue de esfuerzo para comprender los algoritmos neuronales con fuerte componente matemático previa a su implementación usando un lenguaje de programación.

El esquema siguiente, trata de resumir la fase de exploración tecnológica hasta la elección de la alternativa que se usará.



Precisamente en el capítulo siguiente vamos a explorar a detalle las Redes Neuronales, poniendo énfasis en el algoritmo de retropropagación que luego se utilizará para la construcción de un sistema predictor.

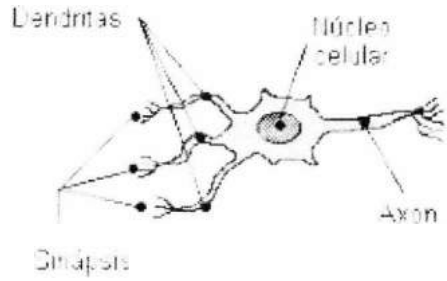
CAPITULO V

DATA MINING CON REDES NEURONALES PARA PREDICCIONES CON ALTA CERTEZA

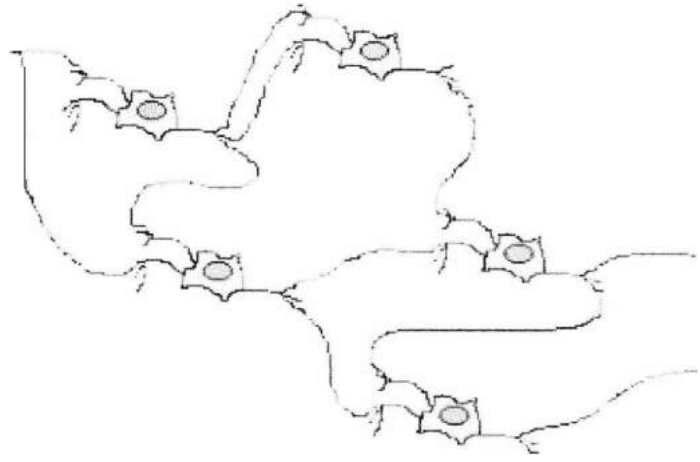
5.1 Generalidades

Las Redes Neuronales surgieron del movimiento conexionista, que nació junto con la Inteligencia Artificial (IA) simbólica o tradicional. Esto fue hacia los años 50, con algunos de los primeros ordenadores de la época y las posibilidades que ofrecían. La IA simbólica se basa en que todo conocimiento se puede representar mediante combinaciones de símbolos, derivadas de otras combinaciones que representan verdades incuestionables o axiomas. Así pues, la IA tradicional asume que el conocimiento es independiente de la estructura que maneje los símbolos, siempre y cuando la 'máquina' realice algunas operaciones básicas entre ellos. En contraposición, los 'conexionistas' intentan representar el conocimiento desde el estrato más básico de la inteligencia: el estrato físico. Creen que el secreto para el aprendizaje y el conocimiento se halla directamente relacionado con la estructura del cerebro: concretamente con las neuronas y la interconexión entre ellas.

Así pues, trabajan con grupos de neuronas artificiales, llamadas Redes Neuronales. La estructura básica de una neurona natural es:



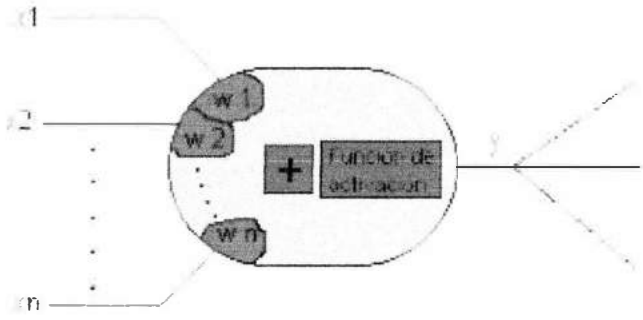
Éstas funcionan como sigue: Cada neurona puede tener infinitas entradas llamadas Dendritas que condicionan el estado de su única salida, el Axón. Este Axón puede ir conectado a una Dendrita de otra neurona mediante la Sinápsis correspondiente, de la siguiente manera:



El Axón da un nivel eléctrico correspondiente a sus entradas y a la importancia que les da a cada una de ellas. Así pues, una neurona puede no reaccionar ante un nivel muy alto de una de sus entradas, o dar una salida muy favorable cuando otra de ellas está mínimamente activa.

En las primeras etapas de nuestra vida, cuando realizamos el aprendizaje de nuestros cerebros, entrenamos nuestras neuronas mediante el éxito o fracaso de una acción a unos estímulos sensoriales. Cuando cierta acción realizada en respuesta a alguna entrada sensorial es exitosa (por ejemplo, al beber agua calmamos la sed), las conexiones sinápticas entre un grupo de neuronas se

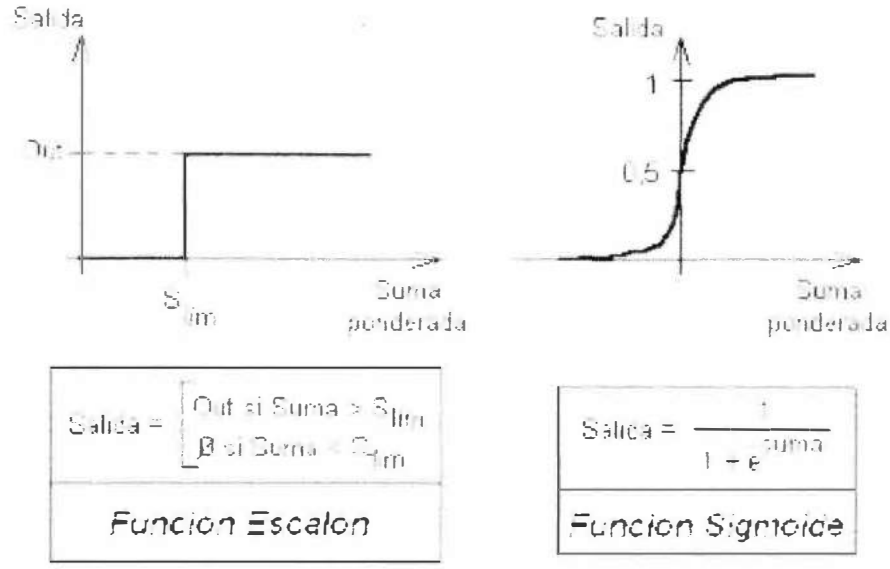
fortalecen, de manera que cuando tengamos una sensación sensorial parecida, la salida será la correcta. De esta forma se forman fuertes conexiones entre grupos de neuronas, que pueden servir para realizar otras acciones complejas. El esquema de una neurona artificial es:



Esta neurona funciona de la siguiente manera: cada entrada x tiene su peso asociado w, que le dará más o menos importancia en la activación de la neurona. Internamente se calcula la suma de cada entrada multiplicada por su peso:

$$\text{Suma Ponderada} = \sum x_i \cdot W_i$$

Con este valor de suma ponderada calculamos una función de activación, que será la salida que nos dará la neurona. Las dos funciones de activación más usada son el Escalón y la Sigmoide:



Más adelante ya se verá para que sirve cada una, pero principalmente se diferencian en que la Sigmoide (llamada así por su forma de S) es diferenciable en todos sus puntos y la Escalón no.

5.2 Algunos usos de las redes neuronales

Un nino pequeño puede mirar una foto familiar y reconocer inmediatamente al padre, a la mascota familiar y a los árboles del jardín con un 100% de efectividad.

Una computadora, sin embargo, requiere mucho trabajo de programación para poder hacer lo mismo, y aun así, el resultado dista mucho de tener la precisión humana.

Si la computadora posee un sistema de reconocimiento de patrones, puede obtener imágenes de una cámara de vídeo y actuar por sí misma en tareas tales como el reconocimiento de entradas y salidas del Personal de una empresa identificando al sujeto mediante la captura de una imagen de vídeo.

Un sistema alternativo consiste en los lectores de huellas digitales. Un scanner diminuto digitaliza la huella digital y forma un patrón con ella, que puede buscarse en la base de datos de personal. Actualmente también se encuentra disponible la

tecnología de lectura de iris por rayos infrarrojos. La persona se para frente a un lector y este digitaliza una imagen del ojo. El iris forma un patrón único para cada persona que puede ser usado como método de identificación similar a las huellas digitales

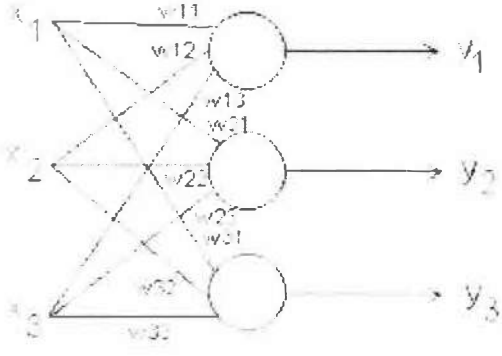
De la misma manera, para el ser humano es relativamente sencillo reconocer los patrones de la escritura manuscrita siendo que esta tarea es sumamente difícil para la computadora. Existen programas que tratan de lograr este objetivo llamados genéricamente OCR (Optical Character Recognition programs) que reconocen la escritura manuscrita tratándola como si fueran patrones

Los operadores de sonares no son otra cosa que personas que reconocen los patrones auditivos registrados por el sonar. Si se desea que la computadora haga este trabajo, necesita un sistema de reconocimiento de patrones para realizar la tarea.

5.3 Modelos neuronales

5.3.1 El perceptrón unicapa

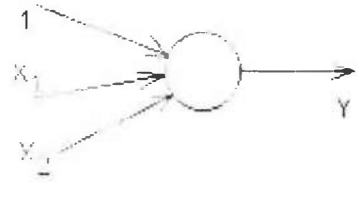
Un Perceptrón unicapa no es más que un conjunto de neuronas no unidas entre sí, de manera que cada una de las entradas del sistema se conectan a cada neurona, produciendo cada una de ellas su salida individual:



Como ya hemos dicho, un conjunto de neuronas no sirve para nada si previamente no le enseñamos qué debe hacer. Existen tres métodos de aprendizaje para un Perceptrón: Supervisado, Por Refuerzo y No Supervisado.

- En el **Aprendizaje Supervisado** se presentan al Perceptrón unas entradas con las correspondientes salidas que queremos que aprenda. De esta manera la red primeramente calcula la salida que da ella para esas entradas y luego, conociendo el error que está cometiendo, ajusta sus pesos proporcionalmente al error que ha cometido (si la diferencia entre salida calculada y salida deseada es nula, no se varían los pesos).
- En el **Aprendizaje No Supervisado**, solo se presentan al Perceptrón las entradas y, para esas entradas, la red debe dar una salida parecida.
- En el **Aprendizaje Por Refuerzo** se combinan los dos anteriores, y de cuando en cuando se presenta a la red una valoración global de como lo está haciendo.

Nosotros, de momento, solo nos centraremos en el Aprendizaje Supervisado. Existen dos maneras de realizarlo. Para entenderlo mejor lo haremos mediante un ejemplo. Imaginemos que queremos hacer aprender a un Perceptrón Unicapa la función AND, de manera que la salida del Perceptrón solo nos dará un 1 si todas sus entradas son 1. Como solo vamos a tener una salida y, suponiendo dos entradas, el Perceptrón estará compuesto simplemente por una neurona, siendo también la tabla a aprender:



X_1	X_2	X_3
0	0	0
0	1	0
1	0	0
1	1	1

Vemos que existe una tercera entrada con un "1". Ésta se incluye siempre para disponer de un "lindero entrenable", o sea, para que el Perceptrón pueda empezar a aprender algo. No tiene más importancia que esa, pero no debemos olvidarla. Para entrenar a este Perceptrón realizaremos los siguientes pasos por orden:

- Fijar unos pesos iniciales para todas las neuronas del Perceptrón
- Presentar una combinación de entradas al Perceptrón
- Hacer que calcule su salida para esas entradas con los pesos que actualmente tiene.
- Conociendo la salida correcta que nos debería dar el Perceptrón, modificar sus pesos de cualquiera de las siguientes maneras:

$$w_i(t+1) = w_i(t) + (C - P) \cdot x_i$$

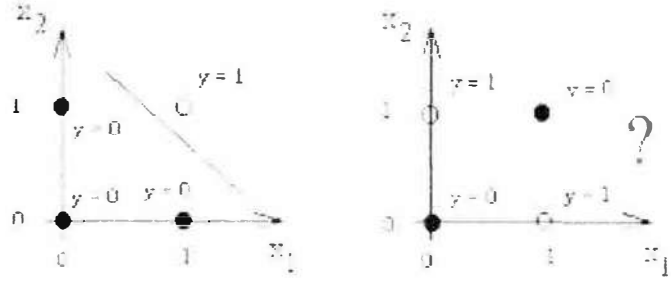
$$w_i(t+1) = w_i(t) + (C - P) \cdot \alpha_i$$

  siendo C la salida esperada y P la salida calculada por el Perceptrón. Alfa es una variable de razón de aprendizaje cualquiera, preferiblemente entre 0 y 1.

- Presentar la siguiente combinación al Perceptrón y repetir el proceso hasta obtener unos errores aceptables.

En principio, podría parecer que el Perceptrón tiene una potencia ilimitada para aprender, pero si en el ejemplo anterior intentamos hacerle aprender una función EXOR, por muchas iteraciones que se haga, ¡ el Perceptrón es incapaz de aprenderla !. Cuando a principios de 1969, Minsky y Paper (en aquella época, fervientes seguidores de la IA simbólica y padres de la moderna IA) pusieron al descubierto estas graves deficiéncias del Perceptrón en su libro *Perceptrons*,

supuso la aletargación del movimiento conexionista durante casi 20 años. Según Minsky y Paper el Perceptrón unicapa era incapaz de aprender las funciones que no fuesen linealmente separables, esto es, que para n variables de entrada se debe poder hacer pasar un 'plano' de (n-1) dimensiones que divida las dos clases totalmente. Si no es así, el Perceptrón no sirve para nada. Un ejemplo claro: observar como la función AND es linealmente separable y la EXOR no lo es:



Función AND Función EXOR

No es posible hacer pasar una línea que divida los puntos negros y blancos en dos partes separadas en la función EXOR Todo esto nos lleva al Perceptrón Multicapa.

5.3.2 El Perceptrón Multicapa

5.3.2.1 Arquitectura

Un Perceptrón multicapa está compuesto por una capa de entrada, una capa de salida y una o más capas ocultas; aunque se ha demostrado que para la mayoría de problemas bastará con una sola capa oculta (Funahashi, 1989; Hornik, Stinchcombe y White, 1989). En la figura 1 podemos observar un perceptrón típico formado por una capa de entrada, una capa oculta y una de salida.

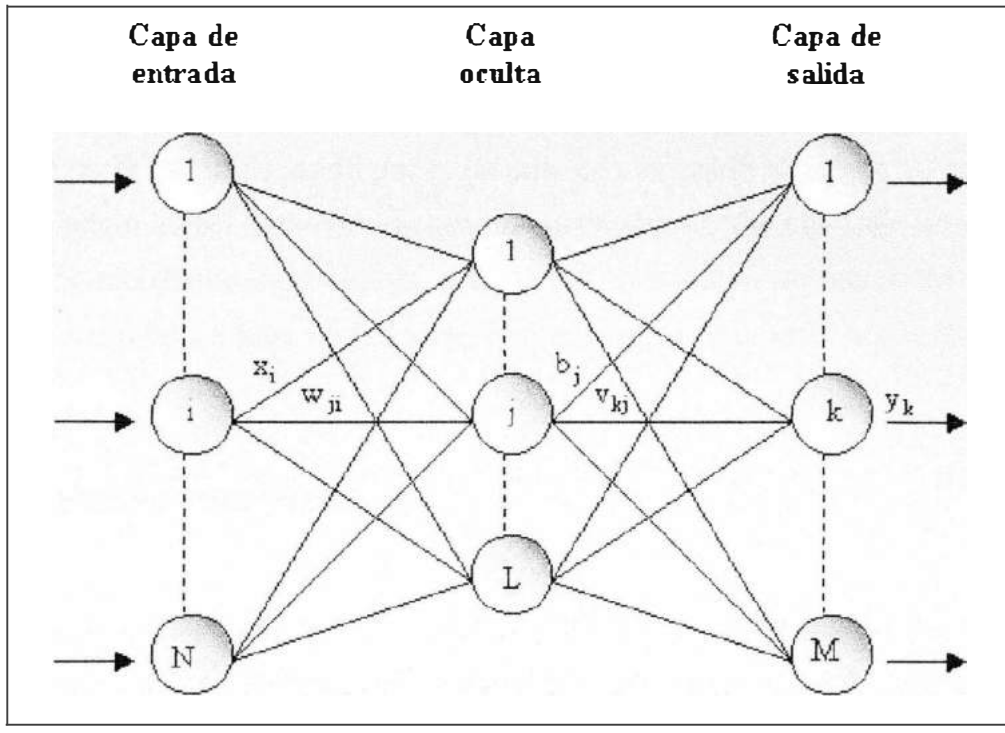


Figura 1: Perceptrón multicapa

En este tipo de arquitectura, las conexiones entre neuronas son siempre *hacia* delante, es decir, las conexiones van desde las neuronas de una determinada capa hacia las neuronas de la siguiente capa; no hay conexiones laterales --esto es, conexiones entre neuronas pertenecientes a una misma capa--, ni conexiones hacia atrás --esto es, conexiones que van desde una capa hacia la capa anterior. Por tanto, la información siempre se transmite desde la capa de entrada hacia la capa de salida.

En el presente documento, hemos considerado w_{ji} como el peso de conexión entre la neurona de entrada i y la neurona oculta j, y v_{kj} como el peso de conexión entre la neurona oculta j y la neurona de salida k.

5.3.2.2 Algoritmo retropropagacion

En el algoritmo *retropropagacion* podemos considerar, por un lado, una etapa de entrenamiento o aprendizaje donde se modifican los pesos de la red de manera que coincida la salida deseada por el usuario con la salida obtenida por la red ante la presentación de un determinado patrón de entrada y, por otro lado una etapa de funcionamiento donde se presenta, ante la red entrenada, un patrón de entrada y éste se transmite a través de las sucesivas capas de neuronas hasta obtener una salida.

5.3.2.2.1 Etapa de Aprendizaje

En la etapa de aprendizaje, el objetivo que se persigue es hacer mínima la discrepancia o error entre la salida obtenida por la red y la salida deseada por el usuario ante la presentación de un conjunto de patrones denominado grupo de entrenamiento. Por este motivo, se dice que el aprendizaje en las redes *retropropagacion* es de tipo supervisado, debido a el usuario (o supervisor) determina la salida deseada ante la presentación de un determinado patrón de entrada.

La función de error que se pretende minimizar para cada patrón p viene dada por:

$$E_p = \frac{1}{2}\sum_{k=1}^{M}(d_{pk} - y_{pk})^2$$

donde d_{pk} es la salida deseada para la neurona de salida k ante la presentación del patrón p. A partir de esta expresión se puede obtener una medida general de error mediante:

$$E = \sum_{p=1}^{P} E_p$$

La base matemática del algoritmo *retropropagacion* para la modificación de los pesos es la técnica conocida como gradiente decreciente (Rumelhart, Hinton y

Williams, 1986). Teniendo en cuenta que E_p es función de todos los pesos de la red, el gradiente de E_p es un vector igual a la derivada parcial de E_p respecto a cada uno de los pesos. El gradiente toma la dirección que determina el incremento más rápido en el error, mientras que la dirección opuesta --es decir, la dirección negativa--, determina el decremento más rápido en el error. Por tanto, el error puede reducirse ajustando cada peso en la dirección:

$$-\sum_{p=1}^{P} \frac{\partial E_p}{\partial w_{ji}}$$

Vamos a ilustrar el proceso de aprendizaje de forma gráfica: El conjunto de pesos que forma una red neuronal puede ser representado por un espacio compuesto por tantas dimensiones como pesos tengamos. Supongamos para simplificar el problema que tenemos una red formada por dos pesos, el paisaje se puede visualizar como un espacio de dos dimensiones. Por otra parte, hemos comentado que el error cometido es función de los pesos de la red; de forma que en nuestro caso, a cualquier combinación de valores de los dos pesos, le corresponderá un valor de error para el conjunto de entrenamiento. Estos valores de error se pueden visualizar como una superficie, que denominaremos superficie del error.

Como se muestra en la figura 2(A), la superficie del error puede tener una topografía arbitrariamente compleja.

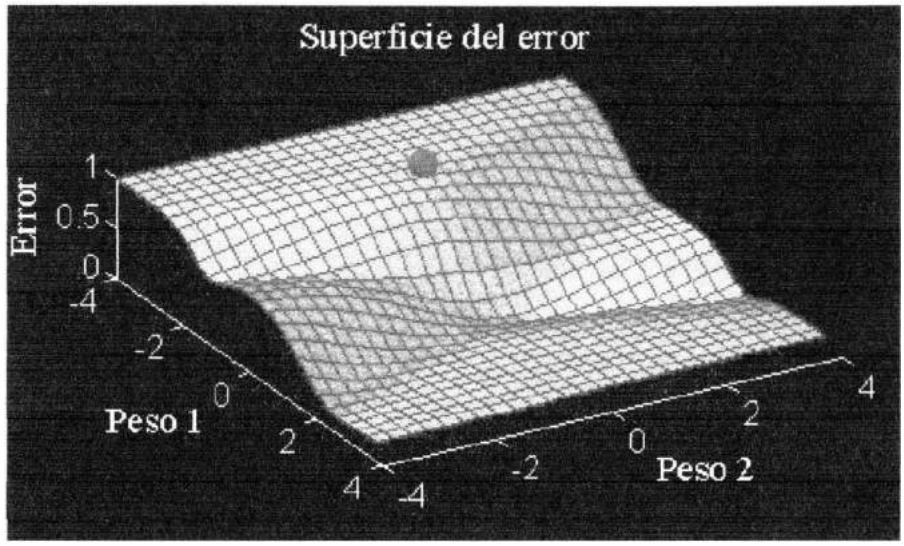


Figura 2 (A): Superficie del error

Con la imagen de la superficie del error en mente, el proceso de entrenamiento comienza en un determinado punto, representado por la bola roja, definido por los pesos iniciales de la red (figura 2(A)). El algoritmo de aprendizaje se basa en obtener información local de la pendiente de la superficie --esto es, del gradiente--, y a partir de esa información modificar iterativamente los pesos de forma proporcional a dicha pendiente, a fin de asegurar el descenso por la superficie del error hasta alcanzar el mínimo más cercano desde el punto de partida. La figura 2(B) muestra el proceso descrito mediante la representación del descenso de la bola roja hasta alcanzar una llanura.

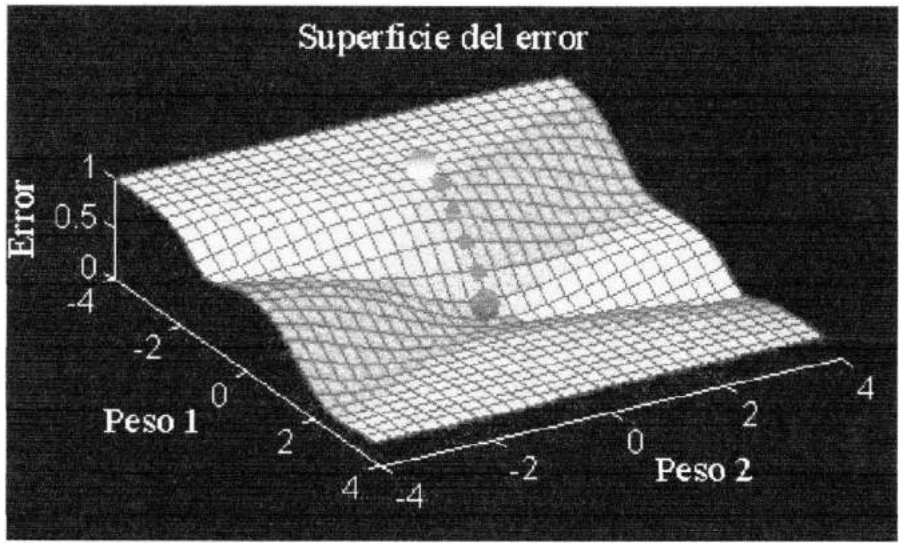


Figura 2 (B): Superficie del error

Con un número mayor de pesos el espacio se convierte en un plano multidimensional inimaginable, aunque se seguirían aplicando los mismos principios comentados en el ejemplo.

Un peligro que puede surgir al utilizar el método de gradiente decreciente es que el aprendizaje converja en un punto bajo, sin ser el punto más bajo de la superficie del error. Tales puntos se denominan mínimos locales para distinguirlos del punto más bajo de esta superficie, denominado mínimo global. Sin embargo, el problema potencial de los mínimos locales se dan en raras ocasiones en datos reales (Rzempoluck, 1998).

A nivel práctico, la forma de modificar los pesos de forma iterativa consiste en aplicar la regla de la cadena a la expresión del gradiente y añadir una tasa de aprendizaje. Así, cuando se trata del peso de una neurona de salida:

$$\Delta v_{kj}(n+1) = \eta \sum_{p=1}^{P} \delta_{pk} b_{pj}$$

donde

$$\delta_{pk} = (d_{pk} - y_{pk}) f'(net_{pk})$$

y n indica la iteración. Cuando se trata del peso de una neurona oculta:

$$\Delta w_{ji}(n+1) = \eta \sum_{p=1}^{P} \delta_{pj} x_{pi}$$

donde

$$\delta_{pj} = f'(net_{pj}) \sum_{k=1}^{M} \delta_{pk} v_{kj}$$

Se puede observar que el error o valor delta asociado a una neurona oculta j, viene determinado por la suma de los errores que se cometen en las k neuronas de salida que reciben como entrada la salida de esa neurona oculta j. De ahí que el algoritmo también se denomine propagación del error hacia atrás.

Para la modificación de los pesos, la actualización se realiza después de haber presentado todos los patrones de entrenamiento. Este es el modo habitual de proceder y se denomina aprendizaje por lotes o modo *batch*. Existe otra modalidad denominada aprendizaje en serie o modo *on line* consistente en actualizar los pesos tras la presentación de cada patrón de entrenamiento. En este modo, se debe tener presente que el orden en la presentación de los patrones debe ser aleatorio, puesto que si siempre se siguiese un mismo orden, el entrenamiento estaría viciado a favor del último patrón del conjunto de entrenamiento, cuya actualización, por ser la última, siempre predominaría sobre las anteriores (Martín del Brío y Sanz, 1997).

Con el fin de acelerar el proceso de convergencia de los pesos, Rumelhart, Hinton y Williams (1986) sugirieron añadir en la expresión del incremento de los pesos un factor momento, □, el cual tiene en cuenta la dirección del incremento tomada en la iteración anterior. Así, cuando se trata del peso de una neurona de salida:

$$\Delta v_{kj}(n+1) = \eta(\sum_{p=1}^{P} \delta_{pk} b_{pj}) + \alpha \Delta v_{kj}(n)$$

Cuando se trata del peso de una neurona oculta:

$$\Delta w_{ji}(n+1) = \eta(\sum_{p=1}^{P} \delta_{pj} x_{pi}) + \alpha \Delta w_{ji}(n)$$

5.3.2.2.2 Etapa de Funcionamiento

Cuando se presenta un patrón p de entrada X_p: $x_{p1},...,x_{pi},...,x_{pN}$, éste se transmite a través de los pesos w_{ji} desde la capa de entrada hacia la capa oculta. Las neuronas de esta capa intermedia transforman las señales recibidas mediante la aplicación de una función de activación proporcionando, de este modo, un valor de salida. Este se transmite a través de los pesos v_{kj} hacia la capa de salida, donde aplicando la misma operación que en el caso anterior, las neuronas de esta última capa proporcionan la salida de la red. Este proceso se puede explicar matemáticamente de la siguiente manera:

La entrada total o neta que recibe una neurona oculta j, net_{pj}, es:

$$net_{pj} = \sum_{i=1}^{N} w_{ji} x_{pi} + \theta_j$$

donde θ es el umbral de la neurona que se considera como un peso asociado a una neurona ficticia con valor de salida igual a 1.

El valor de salida de la neurona oculta j, b_{pj}, se obtiene aplicando una función f(.) sobre su entrada neta:

$$b_{pj} = f(net_{pj})$$

De igual forma, la entrada neta que recibe una neurona de salida k, net_{pk}, es:

$$net_{pk} = \sum_{j=1}^{L} v_{kj} b_{pj} + \theta_k$$

Por último, el valor de salida de la neurona de salida k, y_{pk}, es:

$$y_{pk} = f(net_{pk})$$

Por cierto, recordemos que en el Perceptrón Multicapa, cada neurona necesita un "1" para tener un lindero entrenable. Un ejemplo práctico de un Perceptrón multicapa podría ser su uso en visión artificial. Dada su capacidad para generalizar, las redes neuronales ya han demostrado su importancia en este campo. Se puede hacer que la red aprenda dos o tres formas básicas y que dada otra que no ha aprendido, la clasifique como la más cercana a las que conoce y además dé un porcentaje de lo segura que está que esa nueva forma se parezca a una que ha aprendido.

Las redes neuronales todavía se han de desarrollar mucho. Aún se debe estudiar para que sirven realmente, conocer en que tareas pueden resultar realmente útiles, ya que por ejemplo es difícil saber cuánto tiempo necesita una red para aprender cierta tarea, cuántas neuronas necesitamos como mínimo para realizar cierta tarea, etc... Las redes neuronales pueden llegar a ser algo realmente importante, pero todavía hace falta tiempo para estudiar como almacenan el conocimiento y para desarrollar el hardware paralelo específico que requieren.

En la robótica, las redes neuronales también parecen prometer mucho, sobre todo en su sensorización, para que el robot sea capaz de generalizar lo que siente como estímulos individuales a considerar.

La inteligencia Artificial es un tema demasiado complejo para querer ir deprisa. Solo tiene medio siglo de vida y está en pañales, y lo que ha demostrado hasta ahora (sistemas expertos, demostradores de teoremas) funcionan en un entorno muy restringido, siendo totalmente inútiles fuera de él. Además carecen de sentido común, que es quizás la parte más notable de la inteligencia humana. Se especula que este 'sentido común' solo se puede conseguir con el conocimiento masivo. Si

esto es así se necesitará inculcar a la máquina términos muy sencillos y hacer que aprenda a partir de ellos nuevas cosas que le faciliten la comprensión para aceptar nuevo conocimiento. Todo parece indicar que los resultados más interesantes no se conseguirán ni con la IA simbólica ni con el conexionismo, sino con una mezcla de los dos.

5.3.2.3 Variantes del algoritmo retropropagacion

Desde que en 1986 se presentara la regla *retropropagacion*, se han desarrollado diferentes variantes del algoritmo original. Estas variantes tienen por objeto acelerar el proceso de aprendizaje. A continuación, comentaremos brevemente los algoritmos más relevantes.

La regla *delta-bar-delta* (Jacobs, 1988) se basa en que cada peso tiene una tasa de aprendizaje propia, y ésta se puede ir modificando a lo largo del entrenamiento. Por su parte, el algoritmo QUICKPROP (Fahlman, 1988) modifica los pesos en función del valor del gradiente actual y del gradiente pasado. El algoritmo de gradiente conjugado (Battiti, 1992) se basa en el cálculo de la segunda derivada del error con respecto a cada peso, y en obtener el cambio a realizar a partir de este valor y el de la derivada primera. Por último, el algoritmo RPROP (*Resilient propagation*) (Riedmiller y Braun, 1993) es un método de aprendizaje adaptativo parecido a la regla *delta-bar-delta*, donde los pesos se modifican en función del signo del gradiente, no en función de su magnitud.

5.3.2.4 Fases en la aplicación de un perceptrón multicapa

En el presente apartado se van a exponer los pasos que suelen seguirse en el diseño de una aplicación neuronal (Palmer, Montaño y Calafat, 2000). En general, una red del tipo perceptrón multicapa intentará resolver dos tipos de problemas. Por un lado, los problemas de predicción consisten en la estimación de una variable continua de salida a partir de la presentación de un conjunto de variables

predictoras de entrada (discretas y/o continuas). Por otro lado, los problemas de clasificación consisten en la asignación de la categoría de pertenencia de un determinado patrón a partir de un conjunto de variables predictoras de entrada (discretas y/o continuas).

5.3.2.4.1 Selección de las variables relevantes y preprocesamiento de los datos

Para obtener una aproximación funcional óptima, se deben elegir cuidadosamente las variables a emplear. Más concretamente, de lo que se trata es de incluir en el modelo las variables predictoras que realmente predigan la variable dependiente o de salida, pero que a su vez no covaríen entre sí (Smith, 1993). La introducción de variables irrelevantes o que covaríen entre sí, puede provocar un sobreajuste innecesario en el modelo. Este fenómeno aparece cuando el número de parámetros o pesos de la red resulta excesivo en relación al problema a tratar y al número de patrones de entrenamiento disponibles. La consecuencia más directa del sobreajuste es una disminución sensible en la capacidad de generalización del modelo que como hemos mencionado, representa la capacidad de la red de proporcionar una respuesta correcta ante patrones que no han sido empleados en su entrenamiento.

Un procedimiento útil para la selección de las variables relevantes (Masters, 1993) consiste en entrenar la red con todas las variables de entrada y, a continuación, ir eliminando una variable de entrada cada vez y reentrenar la red. La variable cuya eliminación causa el menor decremento en la ejecución de la red es eliminada. Este procedimiento se repite sucesivamente hasta que llegados a un punto, la eliminación de más variables implica una disminución sensible en la ejecución del modelo.

Una vez seleccionadas las variables que van a formar parte del modelo, se procede al preprocesamiento de los datos para adecuarlos a su tratamiento por la

red neuronal. Cuando se trabaja con un perceptrón multicapa es muy aconsejable --aunque no imprescindible-- conseguir que los datos posean una serie de cualidades (Masters, 1993; Martín del Brío y Sanz, 1997; SPSS Inc., 1997; Sarle, 1998). Las variables deberían seguir una distribución normal o uniforme en tanto que el rango de posibles valores debería ser aproximadamente el mismo y acotado dentro del intervalo de trabajo de la función de activación empleada en las capas ocultas y de salida de la red neuronal.

Teniendo en cuenta lo comentado, las variables de entrada y salida suelen acotarse a valores comprendidos entre 0 y 1 ó entre -1 y 1. Si la variable es de naturaleza discreta, se utiliza la codificación dummy. Por ejemplo, la variable sexo podría codificarse como: 0 = hombre, 1 = mujer; estando representada por una única neurona. La variable nivel social podría codificarse como: 1 0 0 = bajo, 0 1 0 = medio, 0 0 1 = alto; estando representada por tres neuronas. Por su parte, si la variable es de naturaleza continua, ésta se representa mediante una sola neurona, como, por ejemplo, el CI de un sujeto.

5.3.2.4.2 Creación de los conjuntos de aprendizaje, validación y test

En la metodología de las RNA, a fin de encontrar la red que tiene la mejor ejecución con casos nuevos --es decir, que sea capaz de generalizar--, la muestra de datos es a menudo subdividida en tres grupos (Bishop, 1995; Ripley, 1996): entrenamiento, validación y test.

Durante la etapa de aprendizaje de la red, los pesos son modificados de forma iterativa de acuerdo con los valores del grupo de entrenamiento, con el objeto de minimizar el error cometido entre la salida obtenida por la red y la salida deseada por el usuario. Sin embargo, como ya se ha comentado, cuando el número de parámetros o pesos es excesivo en relación al problema --fenómeno del sobreajuste--, el modelo se ajusta demasiado a las particularidades irrelevantes

presentes en los patrones de entrenamiento en vez de ajustarse a la función subyacente que relaciona entradas y salidas, perdiendo su habilidad de generalizar su aprendizaje a casos nuevos.

Para evitar el problema del sobreajuste, es aconsejable utilizar un segundo grupo de datos diferentes a los de entrenamiento, el grupo de validación, que permita controlar el proceso de aprendizaje. Durante el aprendizaje la red va modificando los pesos en función de los datos de entrenamiento y de forma alternada se va obteniendo el error que comete la red ante los datos de validación. Este proceso se ve representado en la figura 3. Podemos observar cómo el error de entrenamiento y el error de validación van disminuyendo a medida que aumenta el número de iteraciones, hasta alcanzar un mínimo en la superficie del error, momento en el que podemos parar el aprendizaje de la red.

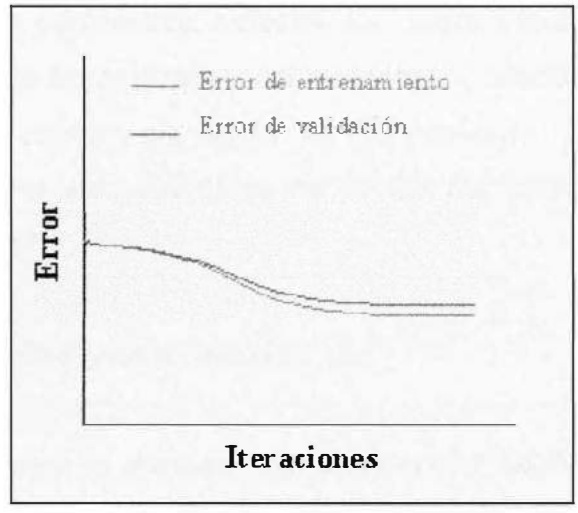


Figura 3: Evolución del error de entrenamiento y el error de validación

Con el grupo de validación podemos averiguar cuál es el número de pesos óptimo --y así evitar el problema del sobreajuste--, en función de la arquitectura que ha tenido la mejor ejecución con los datos de validación. Como se verá más adelante, mediante este grupo de validación también se puede determinar el valor de otros parámetros que intervienen en el aprendizaje de la red.

Por último, si se desea medir de una forma completamente objetiva la eficacia final

del sistema construido, no deberíamos basarnos en el error que se comete ante los datos de validación, ya que de alguna forma, estos datos han participado en el proceso de entrenamiento. Se debería contar con un tercer grupo de datos independientes, el grupo de test el cuál proporcionará una estimación insesgada del error de generalización.

5.3.2.4.3 Entrenamiento de la red neuronal

Una vez visto el funcionamiento del algoritmo *retropropagacion*, a continuación, se proporcionan una serie de consejos prácticos acerca de cuatro grupos de parámetros relacionados con el aprendizaje cuyo valor no se puede conocer *a priori* dado un problema, sino que deben ser determinados mediante ensayo y error. La utilización de un grupo de validación ayudará a conocer el valor óptimo de cada uno de estos parámetros: valor de los pesos iniciales, arquitectura de la red, valor de la tasa de aprendizaje y del momento, y función de activación de las neuronas de la capa oculta y de salida. Así, la configuración de parámetros que obtenga el menor error ante los datos de validación, será la seleccionada para pasar a la fase de test.

5.3.2.4.4 Elección de los pesos iniciales

Cuando una red neuronal es diseñada por primera vez, se deben asignar valores a los pesos a partir de los cuales comenzar la etapa de entrenamiento. Los pesos de umbral y de conexión se pueden inicializar de forma totalmente aleatoria, si bien es conveniente seguir algunas sencillas reglas que permitirán minimizar la duración del entrenamiento.

Es conveniente que la entrada neta a cada unidad sea cero, independientemente del valor que tomen los datos de entrada. En esta situación, el valor devuelto por la función de activación que se suele utilizar --la función sigmoidal--, es un valor intermedio, que proporciona el menor error si los valores a predecir se distribuyen

simétricamente alrededor de este valor intermedio (como habitualmente sucede). Además, al evitar los valores de salida extremos se escapa de las zonas saturadas de la función sigmoidal en que la pendiente es prácticamente nula y, por tanto el aprendizaje casi inexistente.

Para alcanzar este objetivo, la forma más sencilla y utilizada consiste en realizar una asignación de pesos pequeños generados de forma aleatoria, en un rango de valores entre –0.5 y 0.5 o algo similar (SPSS Inc. 1997).

5.3.2.4.5 Arquitectura de la red neuronal

Respecto a la arquitectura de la red, se sabe que para la mayoría de problemas prácticos bastará con utilizar una sola capa oculta (Funahashi, 1989; Hornik, Stinchcombe y White, 1989).

El número de neuronas de la capa de entrada está determinado por el número de variables predictoras. Así, siguiendo los ejemplos de variables comentados anteriormente, la variable sexo estaría representada por una neurona que recibiría los valores 0 ó 1. La variable estatus social estaría representada por tres neuronas que recibirían las codificaciones (1 0 0), (0 1 0) ó (0 0 1). Por último, la variable puntuación en CI estaría representada por una neurona que recibiría la puntuación previamente acotada, por ejemplo, a valores entre 0 y 1.

Por su parte, el número de neuronas de la capa de salida está determinado bajo el mismo esquema que en el caso anterior. Si estamos ante un problema de clasificación, cada neurona representará una categoría obteniendo un valor de activación máximo (por ejemplo, 1) la neurona que representa la categoría de pertenencia del patrón y un valor de activación mínimo (por ejemplo, 0) todas las demás neuronas de salida. Cuando intentamos discriminar entre dos categorías, bastará con utilizar una única neurona (por ejemplo, salida 1 para la categoría A, salida 0 para la categoría B). Si estamos ante un problema de estimación, tendremos una única neurona que dará como salida el valor de la variable a

estimar.

Por último, el número de neuronas ocultas determina la capacidad de aprendizaje de la red neuronal. No existe una receta que indique el número óptimo de neuronas ocultas para un problema dado. Recordando el problema del sobreajuste, se debe usar el mínimo número de neuronas ocultas con las cuales la red rinda de forma adecuada (Masters, 1993; Smith, 1993; Rzempoluck, 1998). Esto se consigue evaluando el rendimiento de diferentes arquitecturas en función de los resultados obtenidos con el grupo de validación.

5.3.2.4.6 Tasa de aprendizaje y factor momento

El valor de la tasa de aprendizaje (□) tiene un papel crucial en el proceso de entrenamiento de una red neuronal, ya que controla el tamaño del cambio de los pesos en cada iteración. Se deben evitar dos extremos: un ritmo de aprendizaje demasiado pequeño puede ocasionar una disminución importante en la velocidad de convergencia y la posibilidad de acabar atrapado en un mínimo local; en cambio, un ritmo de aprendizaje demasiado grande puede conducir a inestabilidades en la función de error, lo cual evitará que se produzca la convergencia debido a que se darán saltos en torno al mínimo sin alcanzarlo. Por tanto, se recomienda elegir un ritmo de aprendizaje lo más grande posible sin que provoque grandes oscilaciones. En general, el valor de la tasa de aprendizaje suele estar comprendida entre 0.05 y 0.5, (Rumelhart, Hinton y Williams, 1986).

El factor momento (□) permite filtrar las oscilaciones en la superficie del error provocadas por la tasa de aprendizaje y acelera considerablemente la convergencia de los pesos, ya que si en el momento n el incremento de un peso era positivo y en n + 1 también, entonces el descenso por la superficie de error en n + 1 será mayor. Sin embargo, si en n el incremento era positivo y en n + 1 es negativo, el paso que se da en n + 1 es más pequeño, lo cual es adecuado, ya que eso significa que se ha pasado por un mínimo y los pasos deben ser menores

para poder alcanzarlo. El factor momento suele tomar un valor próximo a 1 (por ejemplo, 0.9) (Rumelhart, Hinton y Williams, 1986).

5.3.2.4.7 Función de activación de las ocultas y de salida

Hemos visto que para obtener el valor de salida de las neuronas de la capa oculta y de salida, se aplica una función, denominada función de activación, sobre la entrada neta de la neurona. El algoritmo *retropropagacion* exige que la función de activación sea continua y, por tanto, derivable para poder obtener el error o valor delta de las neuronas ocultas y de salida. Se disponen de dos formas básicas que cumplen esta condición: la función lineal (o identidad) y la función sigmoidal (logística o tangente hiperbólica). En las figuras 4, 5 y 6 se presentan las expresiones matemáticas y la correspondiente representación gráfica de la función lineal, la sigmoidal logística (con límites entre 0 y 1) y la sigmoidal tangente hiperbólica (con límites entre -1 y 1):

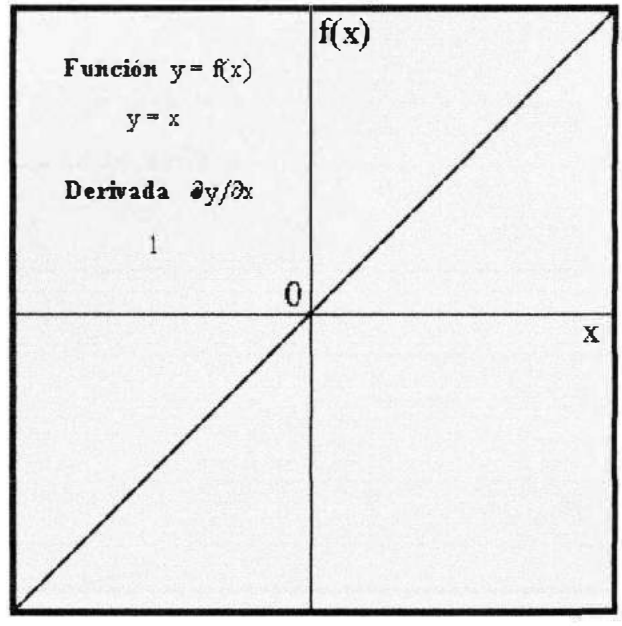


Figura 4: Función lineal

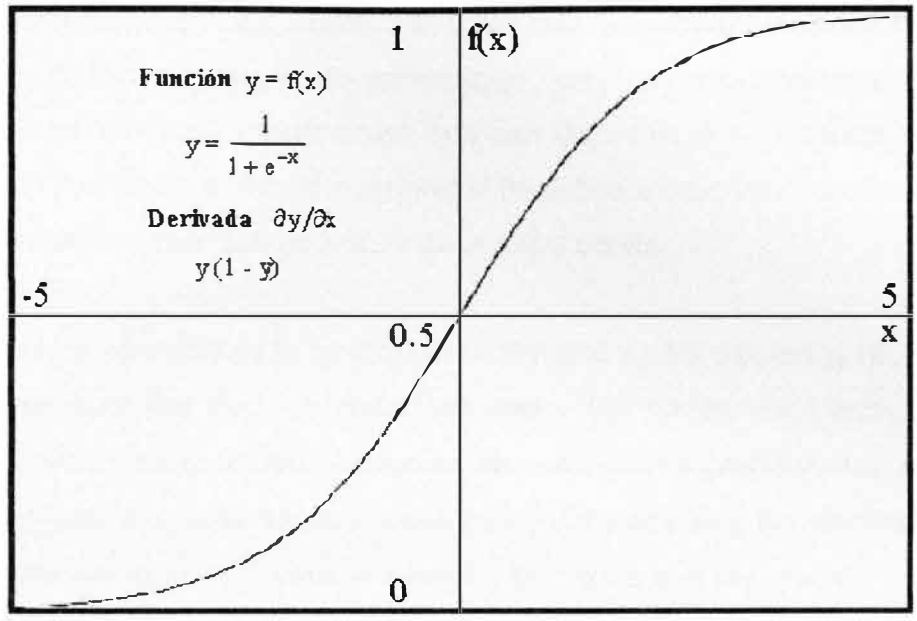


Figura 5: Función sigmoidal logística

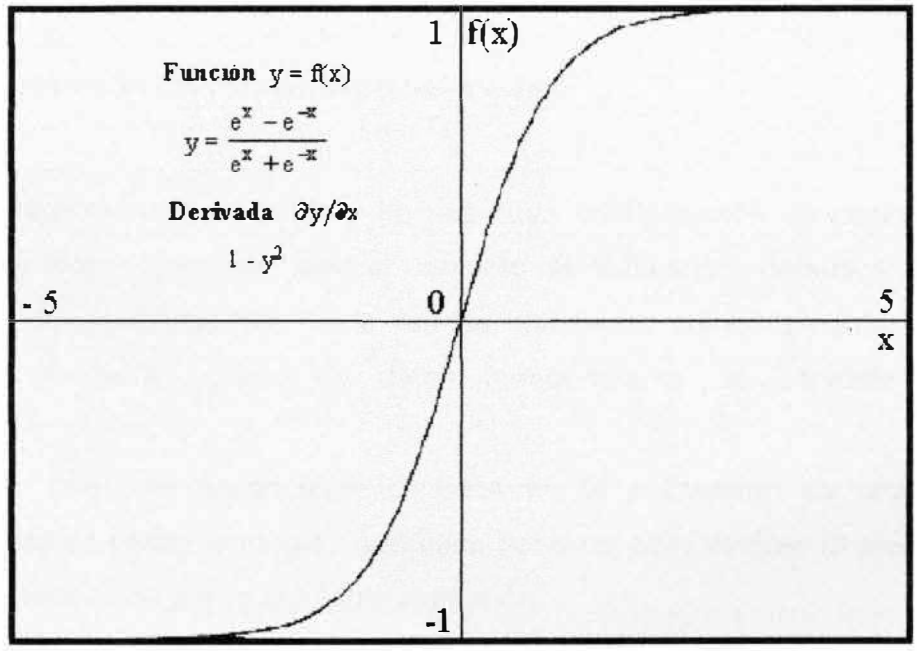


Figura 6: Función sigmoidal tangente hiperbólica

Debemos tener en cuenta que para aprovechar la capacidad de las RNA de aprender relaciones complejas o no lineales entre variables, es absolutamente imprescindible la utilización de funciones no lineales al menos en las neuronas de

la capa oculta (Rzempoluck, 1998). Las RNA que no utilizan funciones no lineales, se limitan a solucionar tareas de aprendizaje que implican únicamente funciones lineales o problemas de clasificación que son linealmente separables. Por tanto, en general se utilizará la función sigmoidal (logística o tangente hiperbólica) como función de activación en las neuronas de la capa oculta.

Por su parte, la elección de la función de activación en las neuronas de la capa de salida dependerá del tipo de tarea impuesto. En tareas de clasificación, las neuronas normalmente toman la función de activación sigmoidal. Así, cuando se presenta un patrón que pertenece a una categoría particular, los valores de salida tienden a dar como valor 1 para la neurona de salida que representa la categoría de pertenencia del patrón, y 0 ó -1 para las otras neuronas de salida. En cambio, en tareas de predicción o aproximación de una función, generalmente las neuronas toman la función de activación lineal.

5.3.2.5 Evaluación del rendimiento del modelo

Una vez seleccionado el modelo de red cuya configuración de parámetros ha obtenido la mejor ejecución ante el conjunto de validación, debemos evaluar la capacidad de generalización de la red de una forma completamente objetiva a partir de un tercer grupo de datos independiente, el conjunto de test.

Cuando la tarea de aprendizaje consiste en la estimación de una función, normalmente se utiliza la media cuadrática del error para evaluar la ejecución del modelo y viene dada por la siguiente expresión:

$$MC_{Error} = \frac{\sum_{p=1}^{P}\sum_{k=1}^{M}\left(d_{pk} - y_{pk}\right)^2}{P \cdot M}$$

Cuando se trata de un problema de clasificación de patrones es más cómodo basarnos en la frecuencia de clasificaciones correctas e incorrectas. A partir del valor de las frecuencias, podemos construir una tabla de confusión y calcular diferentes índices de asociación y acuerdo entre el criterio y la decisión tomada por la red neuronal. Por último, cuando estamos interesados en discriminar entre dos categorías, especialmente si utilizamos la red neuronal como instrumento diagnóstico (por ejemplo, salida = 0 -> sujeto sano; salida = 1 -> sujeto enfermo), es interesante hacer uso de los índices de sensibilidad, especificidad y eficacia, y del análisis de curvas ROC (*Receiver operating characteristic*) (Palmer, Montaño y Calafat, 2000).

CAPITULO VI

DESARROLLO DE UN SISTEMA PREDICTOR MULTIDIMENSIONAL Y MULTIPROPÓSITO USANDO REDES NEURONALES

6.1 Consideraciones

La tarea de predicción del sistema será soportado por una estructura tipo red neuronal, la cual consta de un conjunto de capas (capa de entrada, capas ocultas y capa de salida), conteniendo cada una cantidad de neuronas necesarias para realizar tareas de entrenamiento (aprendizaje) y posterior predicción. El análisis y diseño estará orientado directamente a la posterior implementación del algoritmo retropropagacion con un esquema de programación orientado a objetos.

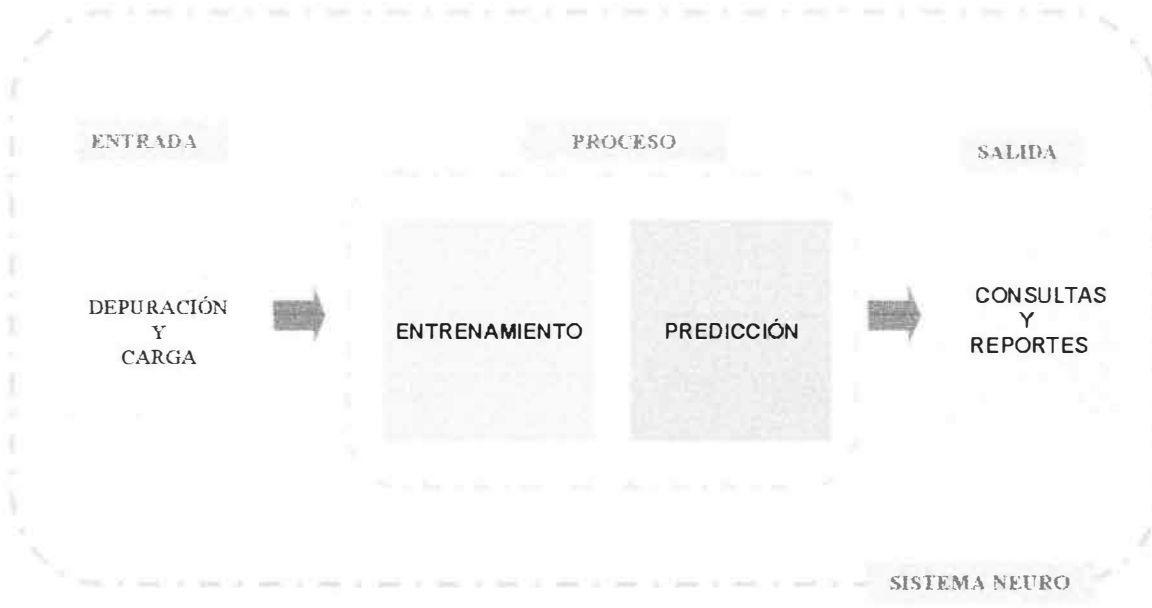
Como resultado de la tarea de modelamiento, se detallarán mas adelante los tres tipos de objetos que soportaran el sistema a desarrollar: el objeto red, el objeto capa y el objeto nodo. Para cada uno de ellos se ha definido un conjunto de atributos y un conjunto de métodos. La relación entre estos tres tipos de objetos se sintetiza en lo siguiente: “un objeto red contiene un conjunto de objetos capa, los cuales a su vez contienen un conjunto de objetos nodo”.

La labor de programación construye los tres objetos anteriormente indicados, implementando los métodos que usará el objeto para llevar a cabo la fase de entrenamiento y predicción. El lenguaje de programación que se está utilizando es Power Builder v8.0, con un manejador de base de datos SQL – Server v.7.0

En las páginas siguientes se podrá observar mayor detalle de la implementación de los objetos y las interfaces del sistema.

6.2 Esquema del sistema predictor

Como todo sistema, posee los tres elementos fundamentales: entrada, proceso y salida. En el diagrama de bloques siguiente se muestra en forma sucinta lo que involucra cada uno de ellos, los mismos que se detallaran en secciones posteriores.



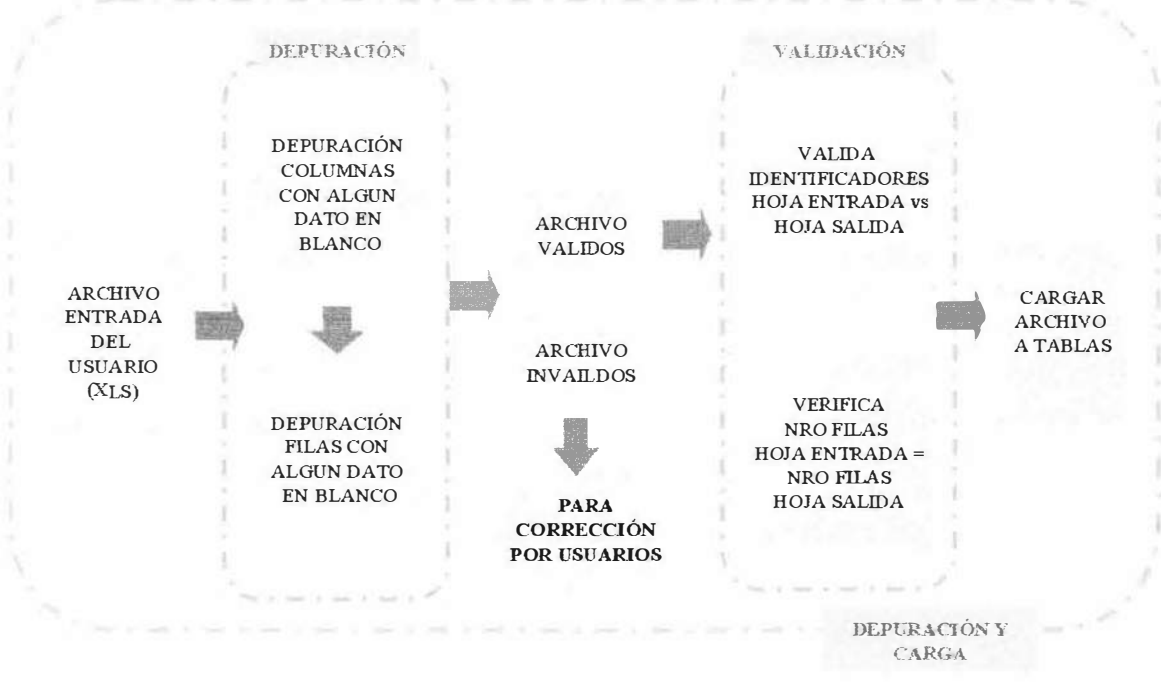
6.2.1 Entradas del sistema

Los datos que se suministran al sistema está conformado por un archivo con filas y columnas en formato MS-Excel. A cada fila de dicho archivo lo denominaremos un patrón. Cada patrón tiene un conjunto de columnas, las cuales representan las variables que van a ser utilizadas para propósitos de entrenamiento o predicción, según el caso.

6.2.1.1 Esquema de depuración y carga

La entrada del sistema tiene como objetivo primordial examinar el contenido del archivo de datos de entrada, el cual debe ser preparado en formato MS-Excel por convención del sistema. Esto consiste básicamente en una depuración y validación de los datos.

La depuración examina las filas y columnas del archivo para identificar los casos con datos en blanco; aquellos casos con algún dato en blanco, son filtrados y almacenados en un archivo MS-Excel de datos inválidos para la corrección por parte del usuario. Los datos validos, incluidos en otro archivo MS-Excel, pasan a un siguiente nivel de validación, en el que se verifica la simetría de la cantidad de identificadores de la hoja de entrada con la hoja de salida, y que la extensión de la hoja de entrada y la hoja de salida sea la misma, es decir que tengan la misma cantidad de filas.



6.2.1.2 Nomenclatura de los datos de entrada en MS-Excel

Cabe precisar, que para que el sistema pueda reconocer correctamente cada elemento del archivo, se debe prepara el mismo con un formato pre-definido, es decir con una nomenclatura pre-establecida para la cabecera de cada una de las columnas.

Si el proceso a ejecutar es el **entrenamiento**, el archivo MS-Excel que se preparará debe contener dos hojas:

a) Una hoja con los datos de entrada para cada patrón, con las siguientes características:

Nombre de la Hoja	**Nombres de las cabeceras de columnas (1ra fila)**
EENTRADA	IEE1,..., IEEm,VEE1, VEE2, VEE3, ..., VEEn

Donde:
IEEm = Identificador Entrenamiento Entrada "m"-ésima
VEEn = Variable Entrenamiento Entrada "n"-ésima

En la segunda fila, y debajo de cada identificador se guarda la descripción de cada variable según el analista tributario o usuario que prepare los datos de entrada.

Los valores de cada una de las variables de entrada consideradas se deben incluir a partir de la tercera fila.

b) Una hoja con los datos de salida para cada patrón, con las siguientes características:

Nombre de la Hoja	**Nombres de las cabeceras de columnas (1ra fila)**
ESALIDA	IES1,..., IESm,VES1, VES2, VES3, ..., VESn

Donde:
IESm = Identificador Entrenamiento Salida "m"-ésima
VESn = Variable Entrenamiento Salida "n"-ésima

La segunda fila en forma análoga a la hoja de entrada.
Los valores de cada una de las variables de salida consideradas se deben incluir a partir de la tercera fila.

Si el proceso a ejecutar es el **funcionamiento o predicción**, el archivo MS-Excel que se preparará debe contener solo una hoja:

Los datos de entrada para cada patrón, debe tener las siguientes características:

Nombre de la Hoja	**Nombres de las cabeceras de columnas (1ra fila)**
PENTRADA	IPE1,..., IPEm, VPE1, VPE2, VPE3, ..., VPEn

Donde:
IPEm = Identificador Predicción Entrada "m"-ésima
VPEn = Variable Predicción Entrada "n"-ésima

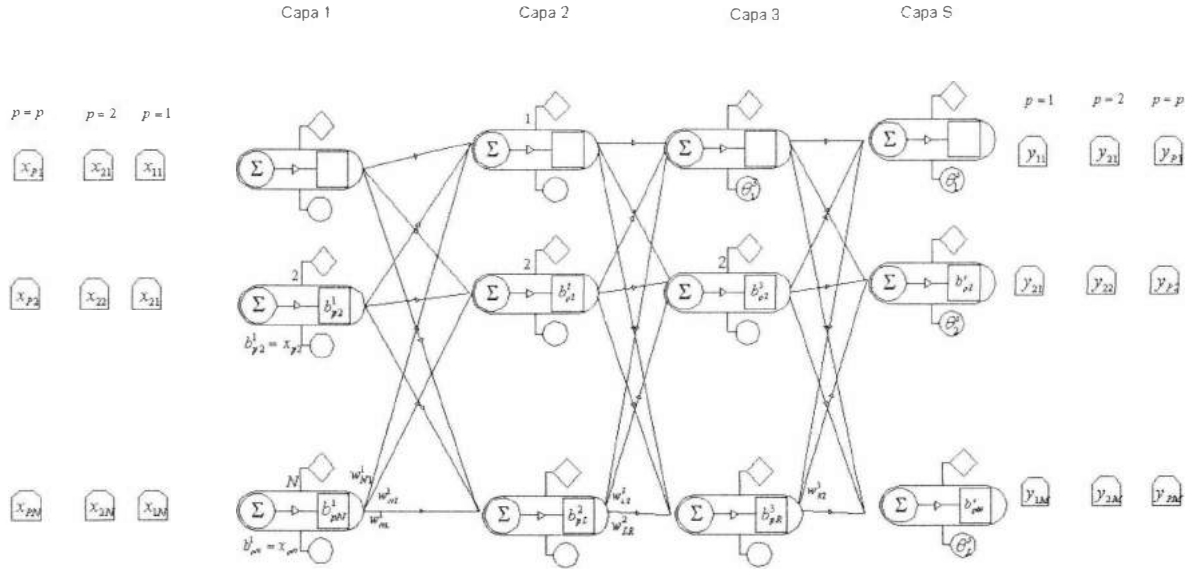
En la segunda fila poner la descripción de las variables según notación del usuario. Los valores de cada una de las variables de entrada consideradas se deben incluir a partir de la tercera fila.

En este caso no se suministra un archivo de salida, por que precisamente es lo que se quiere obtener como resultado del proceso de predicción. El sistema generará una hoja adicional en el mismo archivo MS-Excel con la nomenclatura que se indica en el numeral "salidas del sistema".

6.2.2 Elementos de procesamiento

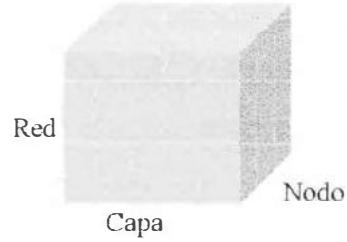
6.2.2.1 Esquema general de una red neuronal

En el siguiente gráfico se muestra un modelo general de una red neuronal, donde se puede apreciar la interrelación entre las capas y sus nodos correspondientes. Observar especialmente los componentes de cada nodo.



6.2.2.2 Objetos del sistema

Para el procesamiento, el sistema predictor se sustenta básicamente en la interrelación de tres elementos fundamentales: el objeto red, un objeto capa y un objeto nodo. La relación entre ellos se puede apreciar en el siguiente gráfico:



Una Red Neuronal debe tener varias capas (entrada,ocultas, salida).

Cada capa debe tener una o mas neuronas (nodos).

Una neurona (nodo) es la unidad atómica de procesamiento.

6.2.2.3 Esquema del entrenamiento

Los datos antes de ser sometidos al entrenamiento, se someten a un tratamiento previo: un escalamiento. Teniendo en cuenta que el valor de las variables insumo de la red pueden tener valores muy grandes y con una dispersión muy grande, lo primero que se realiza es una normalización de los datos, empleando el promedio y la desviación estándar de la muestra contenida en el archivo de datos. Luego todos los valores son convertidos proporcionalmente a su magnitud, a una nueva escala en el intervalo [0,1].

Después del escalamiento de los datos, se procede a configurar la red neuronal, esto significa indicar la cantidad de capas ocultas y las neuronas para cada una de ellas; asimismo, los valores del error máximo permitido, el numero de iteraciones limite para detener el entrenamiento y otros parámetros como la tasa de aprendizaje, el factor momento, entre otros.

El entrenamiento propiamente dicho, empieza con la inicialización de pesos con valores aleatorios entre 0 y 1. Luego se efectúa el aprendizaje de la red en forma iterativa hasta que el error calculado por la red neuronal sea menor o igual al error

deseado o que el número de iteraciones haya llegado a la cantidad limite de iteraciones, evitando con este ultimo criterio una iteración perpetua, en el caso no se alcance el error deseado.

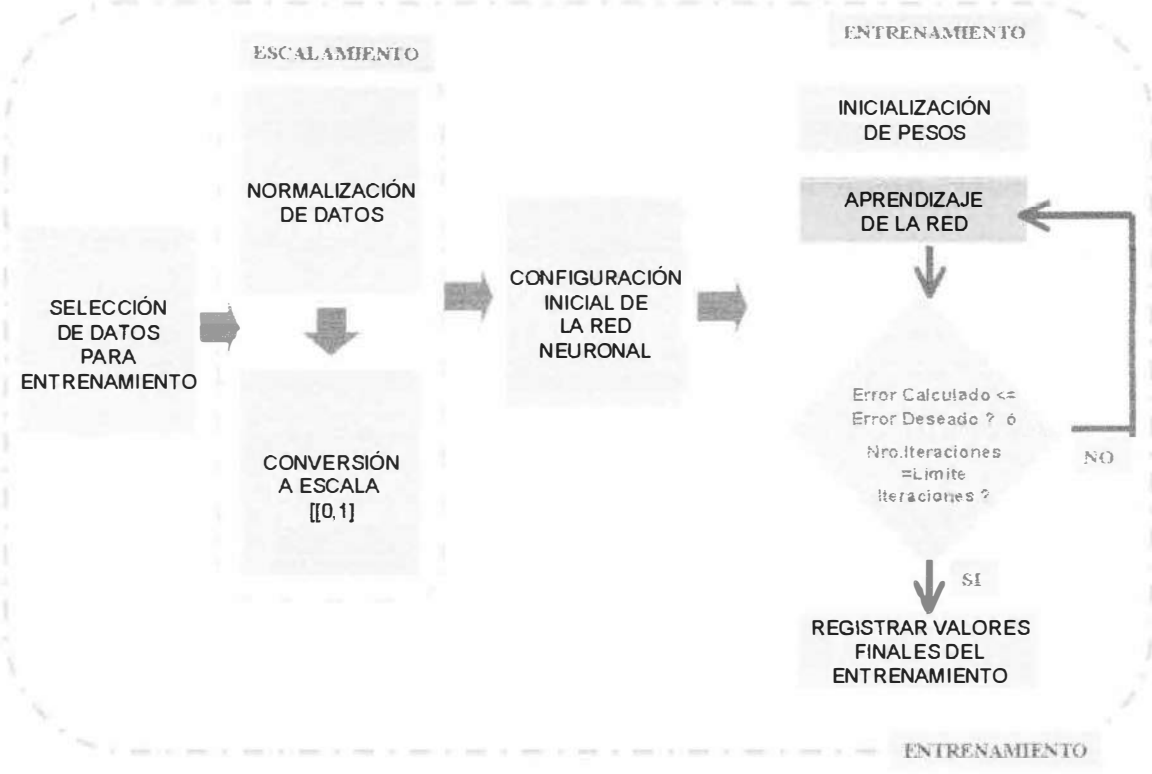
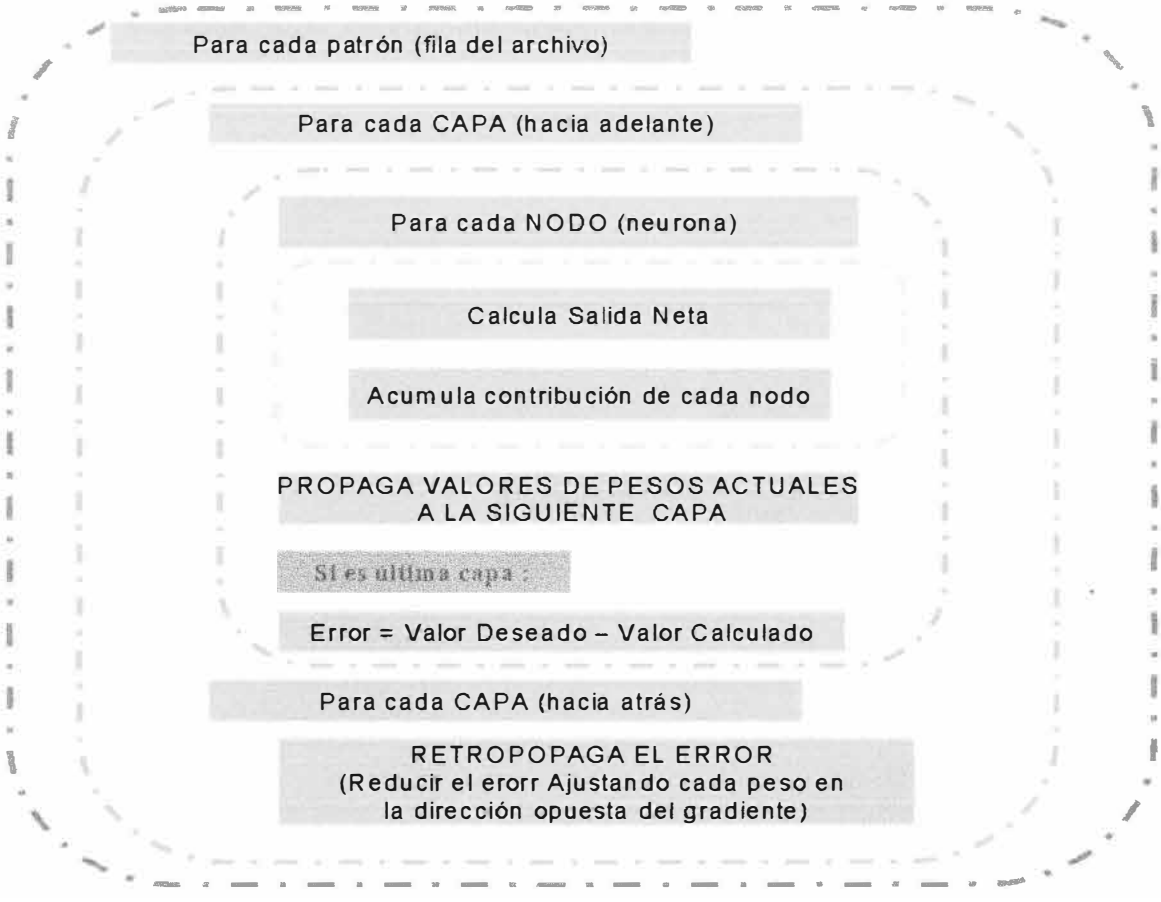


Diagrama del aprendizaje

El aprendizaje es la piedra angular para el entrenamiento de la red neuronal. La "magia" del aprendizaje se hace posible a través de la modificación continua de los pesos asociados a las conexiones entre las neuronas de una capa y las neuronas de la capa adyacente (anterior y/o posterior, según el caso).

La secuencia de aprendizaje se da en forma iterativa, procesando cada patrón del archivo de entrada. Dentro de cada patrón de entrenamiento, se realiza primero un "recorrido hacia delante" que consiste en propagar los datos del patrón a través de las neuronas de las capas intermedias (ocultas) hasta llegar a la capa de salida. Para cada neurona se calcula un valor de salida aplicando el valor de entrada a la función de activación.

Al llegar a la última capa, se calcula el error de esa iteración, como la diferencia entre el valor deseado y el valor calculado. Luego se comienza la propagación hacia atrás (retropropagacion) que consiste en calcular un "delta" para cada neurona. Para la ultima capa, el "delta" se obtiene a partir del error hallado y la aplicación del valor de la neurona a la derivada de la función de activación. Para las capas anteriores, se usan los "deltas" de la capa inmediata siguiente, asociando para el calculo la aplicación del valor de la neurona a la derivada de la función de activación. Para mas detalle puede ver el Anexo "seudocódigo de métodos del sistema neuro y el Anexo "Ejemplo numérico del aprendizaje"

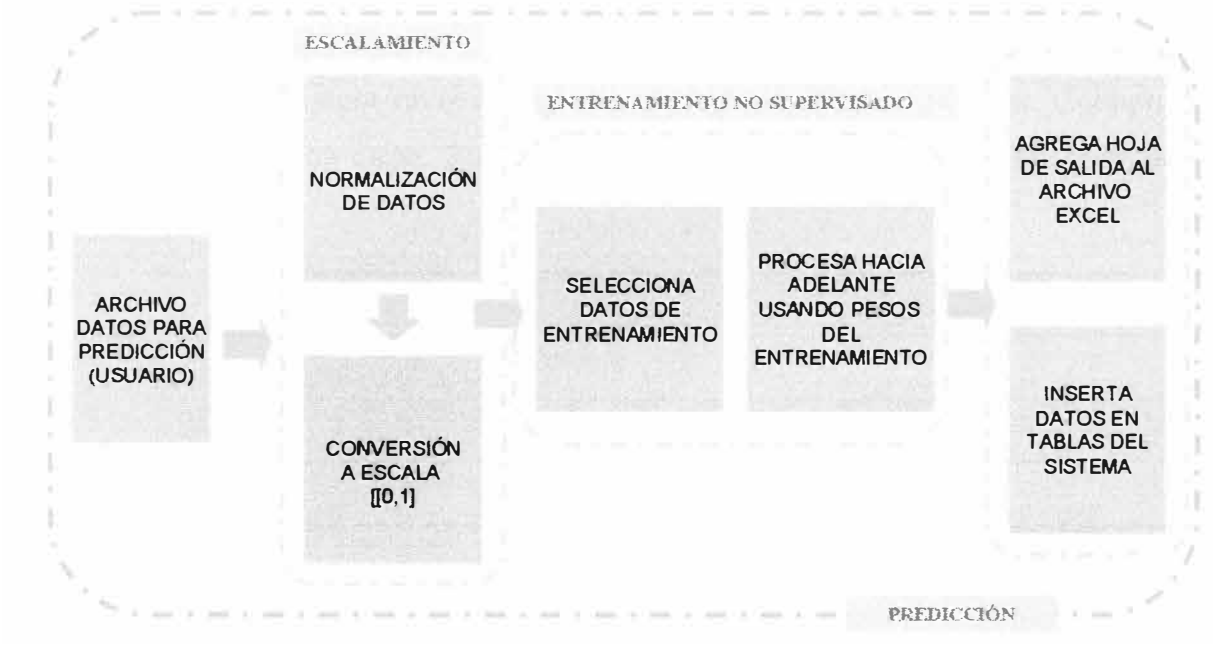


6.2.2.4 Esquema de la predicción

Para realizar una predicción, se suministra al sistema un archivo MS-Excel, pero con la particular característica de tener solo la hoja de entrada (la hoja de salida es precisamente lo que se desea obtener). En forma analoga al entrenamiento se

realiza el escalamiento de los datos y validación de los datos de entrada. Luego se procede a seleccionar de la lista de archivos de entrenamiento, uno que se haya realizado con el mismo tipo de variables cuyos valores ahora se quiere predecir.

La predicción es un proceso mas sencillo, dado que no hay recalcular pesos, ni retropropagar valores de errores calculados. En una predicción, se usan los valores del archivo de entrada y estos se propagan hacia delante usando los pesos finales obtenidos en el entrenamiento, de tal forma que se calculan los valores de salida a través de las capas intermedias, hasta llegar a la capa de salida. Finalizado el proceso de predicción se crea una hoja de salida en el mismo archivo MS-Excel que se suministró con la hoja de entrada. Además, se graban los datos obtenidos, en las tablas del sistema.



6.2.2.5 Descripción detallada de los Objetos

En los cuadros siguientes se podrá observar una descripción detallada de los objetos que soportaran el sistema predictor. En el cuadro 6.1, se observa la descripción de cada uno de los tipos de objetos.

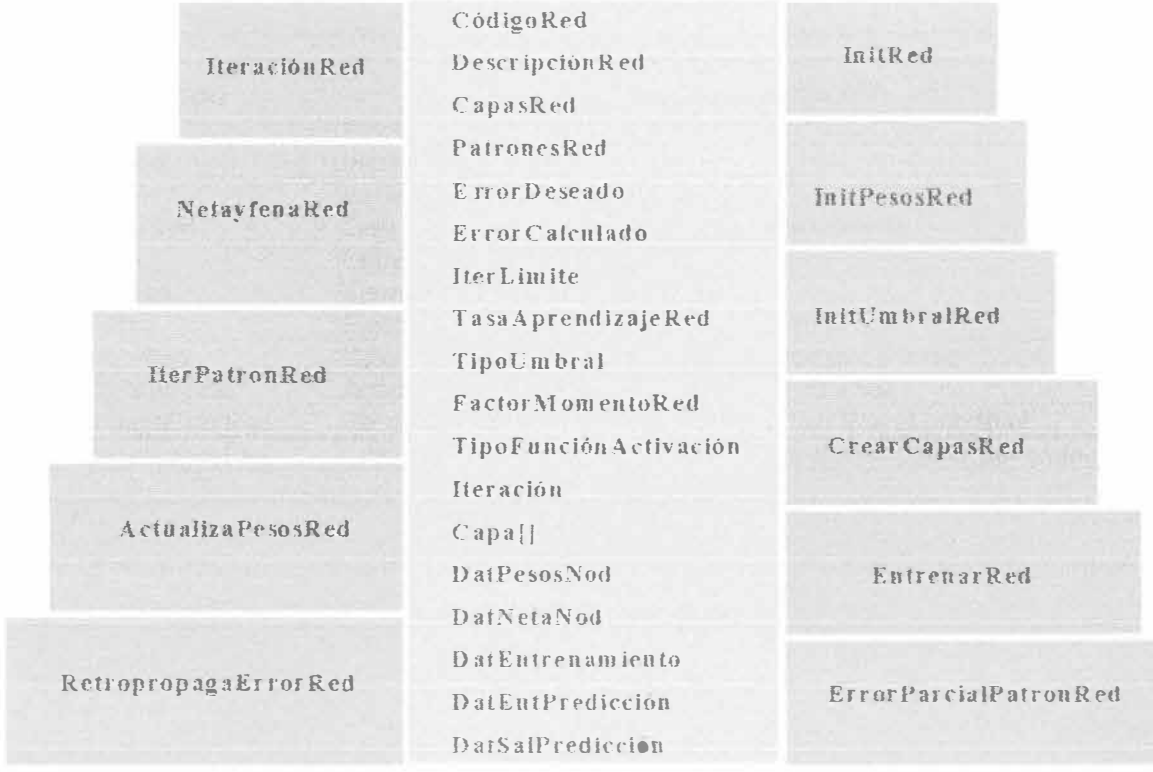
Predictor - Neuro - Objetos Descripción	
Objeto	**Descripción**
Uo_Red	Es el objeto de mayor nivel. Contiene un conjunto de objetos capa
Uo_Capa	Es el objeto de nivel intermedio. Contiene un conjunto de objetos nodo
Uo_Nodo	Es el objeto de menor nivel. Es la unidad atómica o básica del sistema

Cuadro 6.1

Los atributos y métodos de estos tres objetos se irán describiendo detalladamente en las secciones siguientes.

6.2.2.5.1 El Objeto Red

Es el objeto de mayor nivel de agregación en el modelo neuronal. Contiene un conjunto de objetos capa. Sus atributos y métodos se muestran en el siguiente gráfico:



El objeto Red

La descripción de cada uno de los atributos del objeto red se puede apreciar en el siguiente cuadro:

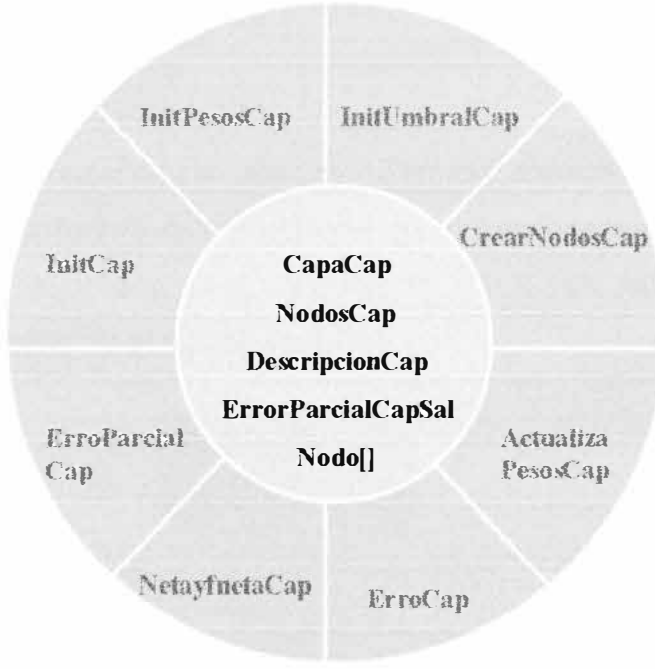
IPESAT - Objeto Red	
Atributo	**Descripción**
CódigoRed	Código de identificación de la Red
DescripciónRed	Descripción de la Red
CapasRed	Número de capas en la Red (Entradas,Ocultas y Salida)
PatronesRed	Número de patrones de Entrada o Salida (la cantidad de la misma)
ErrorDeseado	Error ideal o el que se toma de referencia
ErrorCalculado	Error calculado en una iteración (error luego de recorrer todos los patrones)
IterLimite	Limite de iteraciones (ingresado por el usuario)
TasaAprendizajeRed	Valor numérico de la tasa de aprendizaje
TipoUmbral	Tipo de umbral
FactorMomentoRed	Valor numérico del factor momento
TipoFunciónActivación	Tipo de función de activación (SL,SH,LI)
Iteración (*)	Número de iteración en la que se encuentra el proceso
Capa[]	Cantidad de capas en la red
DatPesosNod	Datos de los pesos de todos los nodos de la red
DatNetaNod	Datos de umbral, valor neto, salida neta y error por nodo
DatEntrenamiento	Datos de entrada y datos de salida del entrenamiento
DatEntPredicción	Datos de entrada para predicción
DatSalPredicción	Datos de salida de la predicción

Los métodos del objeto red, se describen en el siguiente cuadro:

IPESAT - Objeto Red	
Método	**Descripción**
InitRed	Inicialización de la red
InitPesosRed	Inicialización de pesos de la red
InitUmbralRed	Inicialización del umbral de la red
CrearCapasRed	Crea las capas necesarias para el entrenamiento
EntrenarRed	Entrena la red una vez inicializada
IterPatronRed	Ejecuta una iteración con un patrón
IteraciónRed	Ejecuta paso a paso el entrenamiento de la red
NetayFnetaRed	Calcula Neta y Funcion Neta para las capas de la red
ActualizaPesosRed	Actualiza los pesos de la red (de principio a fin)
RetropropagaErrorRed	Recorre o propaga el error hacia atrás (del final al principio)
ErrorParcialPatronRed	Calcula el error en la capa de salida (es el error parcial por cada patrón)

6.2.2.5.2 El Objeto Capa

Es el objeto de nivel intermedio, y permitirá manejar las ocurrencias de la capa de entrada, las capas ocultas y la capa de salida. Cada objeto capa contendrá un conjunto de objetos nodo. Sus atributos y métodos se muestran en el siguiente gráfico:



La descripción de cada uno de los atributos del objeto capa se puede apreciar en el siguiente cuadro:

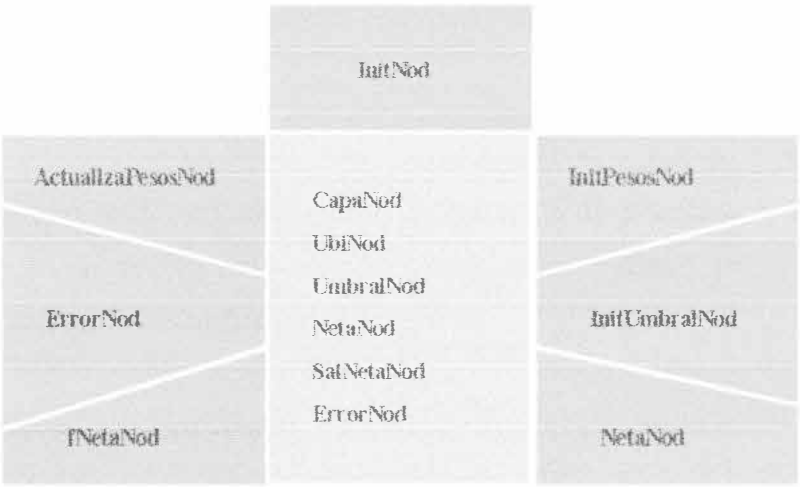
IPESAT - Objeto Capa	
Atributo	**Descripción**
DescripciónCap	Información de la capa (Entrada,Oculta, Salida)
CapaCap	Es el orden (posición) de la capa en la red
NodosCap	Número de nodos en la capa actual
ErrorParcialCapSal	Error parcial en la capa de salida
Nodo[]	Cantidad de nodos en la capa actual

Asimismo, el detalle de los métodos del objeto capa se describen en el siguiente cuadro:

IPESAT - Objeto Capa	
Método	**Descripción**
InitCap	Crea los nodos e inicializa los atributos
InitPesosCap	Inicializa los pesos de una capa
InitUmbralCap	Inicializa el umbral de una capa
CrearNodosCap	Crea los nodos necesarios para el entrenamiento
ActualizaPesosCap	Actualiza los pesos de los nodos en la capa actual
ErroCap	Ejecuta la retropropagación del error en la capa
NetayfnetaCap	Calcula Neta y Función Neta en la capa
ErroParcialCap	Calcula error parcial solo para la capa de salida

6.2.2.5.3 El Objeto Nodo

Es el objeto de nivel elemental o atómico. Permite almacenar información individualizada de cada neurona que interviene en el funcionamiento de la red neuronal. Asimismo es la unidad de procesamiento de la red. Sus atributos y métodos se pueden apreciar en el siguiente gráfico:



La descripción de cada uno de los atributos del objeto nodo se puede apreciar en el siguiente cuadro:

IPESAT - Objeto Nodo	
Atributo	**Descripción**
CapaNod	Capa a la que pertenece el nodo
UbiNod	Ubicación del nodo en la capa
UmbralNod	Umbral del nodo (dato inicializado)
NetaNod	Neto del nodo actual
SalNetaNod	Salida Neta del nodo actual
ErrorNod	Error en el nodo actual

Asimismo, el detalle de los métodos del objeto nodo se describen en el siguiente cuadro:

IPESAT - Objeto Nodo	
Método	**Descripción**
InitNod	Inicializa los parámetros del nodo
InitPesosNod	Inicializa los pesos de un nodo vs los nodos de la capa siguiente
InitUmbralNod	Inicializa el umbral en cada nodo
NetaNod	Calcula el valor neto de cada nodo
fNetaNod	Evalúa la función de activación para el nodo actual
ErrorNod	Calcula el error para un nodo vs los nodos de la capa siguiente
ActualizaPesosNod	Actualiza los pesos de un nodo vs los nodos de la capa siguiente

6.2.3 Salidas del sistema

Tanto el proceso de entrenamiento como el proceso de predicción registran los datos obtenidos en las tablas del sistema. Estos datos pueden ser consultados a través de las opciones habilitadas con este fin.

6.2.3.1 Esquema de consultas y reportes

Para el entrenamiento, se puede seleccionar y visualizar los datos que se cargaron en la hoja de entrada del archivo MS-Excel, tanto los datos de identificación como los valores de las variables para cada patrón de entrada. Análogamente, se pueden visualizar los datos que se cargaron en la hoja de salida. Adicionalmente, para un entrenamiento específico se puede conocer la configuración de la red utilizada, es decir, el numero de capas ocultas, la cantidad de neuronas para cada una de ellas, la cantidad de iteraciones, el error final obtenido, los parámetros de entrenamiento, entre otros.

6.2.3.2 Nomenclatura de los datos de salida en MS-Excel

Cuando el sistema se ejecuta con fines de entrenamiento, no registra datos en la hoja MS-Excel, simplemente todo se almacena en la base de datos. Recordar que para el entrenamiento, todos se le ha proporcionado al sistema un archivo con dos hojas, una de entrada y otra de entrada.

Si la ejecución es con propósitos de predicción, el sistema genera una hoja con valores de salida para la(s) variable(s) que se predice(n), en una cantidad equivalente a los patrones suministrados como entrada.

La hoja MS-Excel con los datos de salida para cada patrón, tendrá las siguientes características:

Nombre de la Hoja	Nombres de las cabeceras de columnas (1ra fila)
PSALIDA	IPS1,..., IPSm,VPS1, VPS2, VPS3, ..., VPSn

Donde:
IPSm = Identificador Predicción Salida "m"-ésima
VPSn = Variable Predicción Salida "n"-ésima

Los valores de cada una de las variables resultado de la predicción, serán registradas a partir de la tercera fila, pero el sistema también arma la primera y segunda fila a partir de la hoja de entrada.

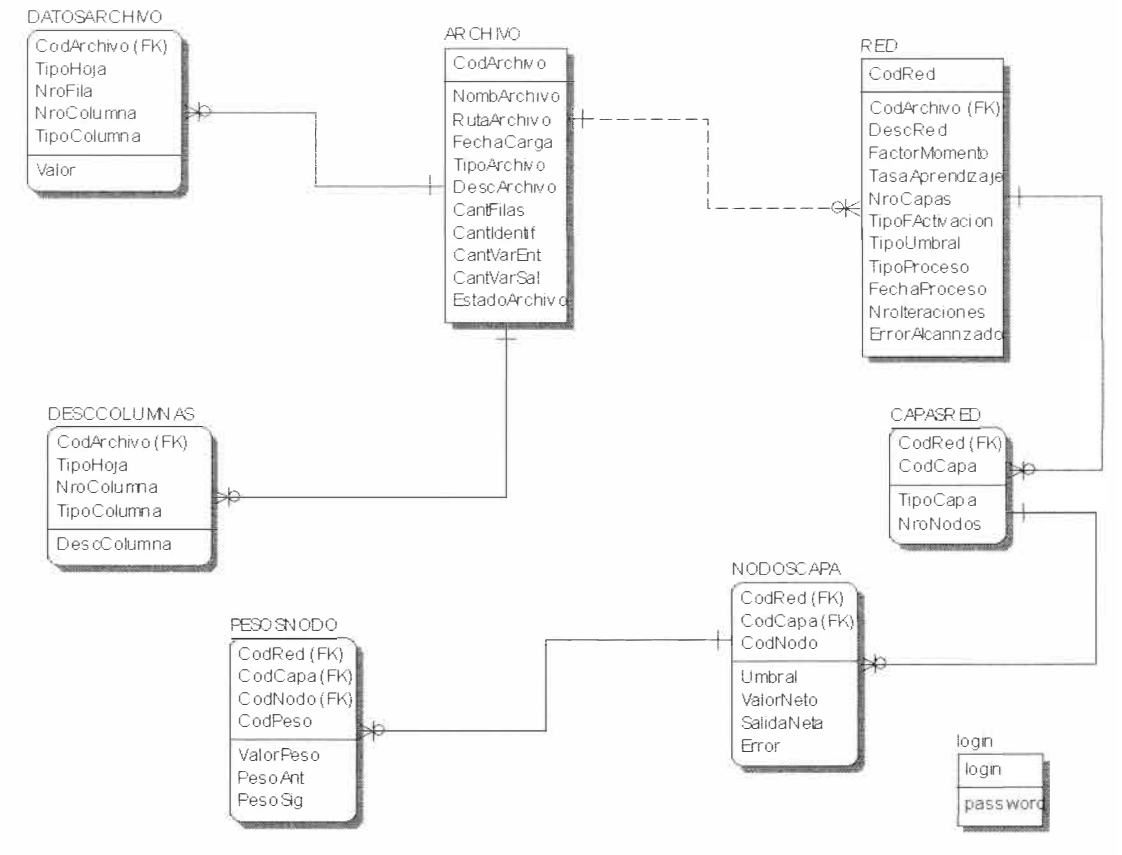
6.2.4 Repositorio de datos

Uno de los grandes rasgos distintivos del sistema predictor sobre cualquier otro prototipo que se haya implementado usando redes neuronales, es la capacidad de almacenar los datos en forma permanente, en un conjunto de tablas interrelacionadas que forman parte de un modelo relacional.

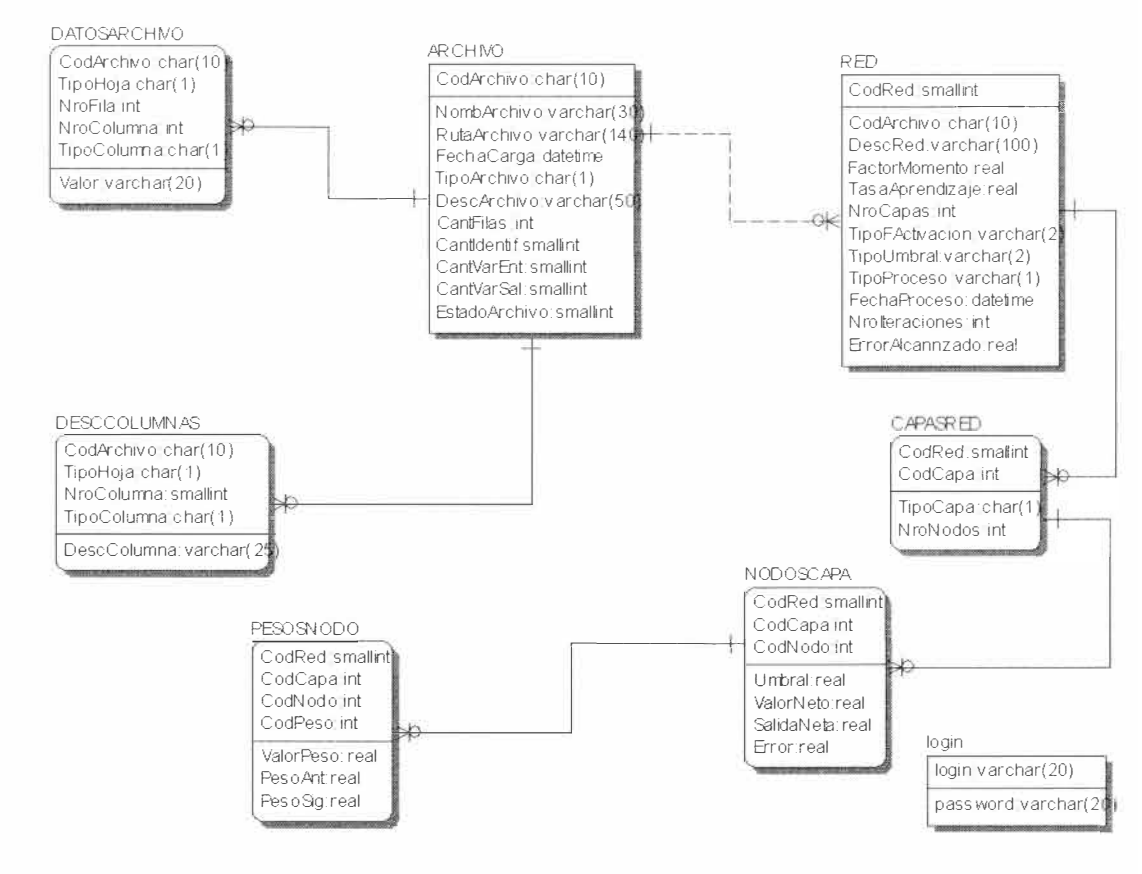
El sistema almacena los datos de un entrenamiento óptimo para un conjunto de variables de entrada, de tal manera que pueda ser directamente aprovechado posteriormente en una situación similar, es usando los mismos tipos de variables, evitando realizar en nuevo entrenamiento. Dado que el sistema guarda los datos en tablas y no solo en memoria RAM, la disponibilidad de dichos datos no requiere el funcionamiento permanente del sistema.

Con la posibilidad de disponer de datos de entrenamientos anteriores, la etapa de funcionamiento o predicción se facilita y fortalece enormemente. Es suficiente obtener un entrenamiento con valores adecuados, y utilizar ese "aprendizaje" en tareas de predicción.

Se muestra a continuación el modelo lógico de datos que subyace al sistema para los fines de almacenamiento antes descritos:



Para mayor detalle de los tipos de dato de cada campo, se incluye a continuación el modelo físico del sistema.



6.3 Factores de innovación del sistema predictor

a) Es multipropósito

- Puede predecir variables de salida de cualquier actividad que maneje información.
- Permite definir las variables con las que se va trabajar.

b) Es multidimensional

- Trabaja en forma irrestricta y dinámica con varias capas y varios nodos.
- El usuario define interactivamente la cantidad de capas y el número de nodos que usará el sistema.

- Permite definir las variables con las que se va trabajar.

c) Implementa algoritmos neuronales con programación orientada a objetos (OOP).

- Objetos Red, Capa y Nodo (atributos y métodos).

d) Acepta múltiples formas de entrada/salida

- En línea (para configurar parámetros de la red neuronal).
- En batch (archivos de texto, archivos Excel).

e) Permite almacenamiento de datos no volátil (permanente).

- Usa un conjunto tablas de una base de datos relacional.
- Guarda información histórica de los procesos ejecutados.

f) Facilita reutilización de "experiencia" predictiva.

- Aprovecha la combinación de valores óptimos de entrenamientos anteriores para nuevas predicciones.
- Extrae y actualiza el "conocimiento" en forma acumulativa.

CAPITULO VII

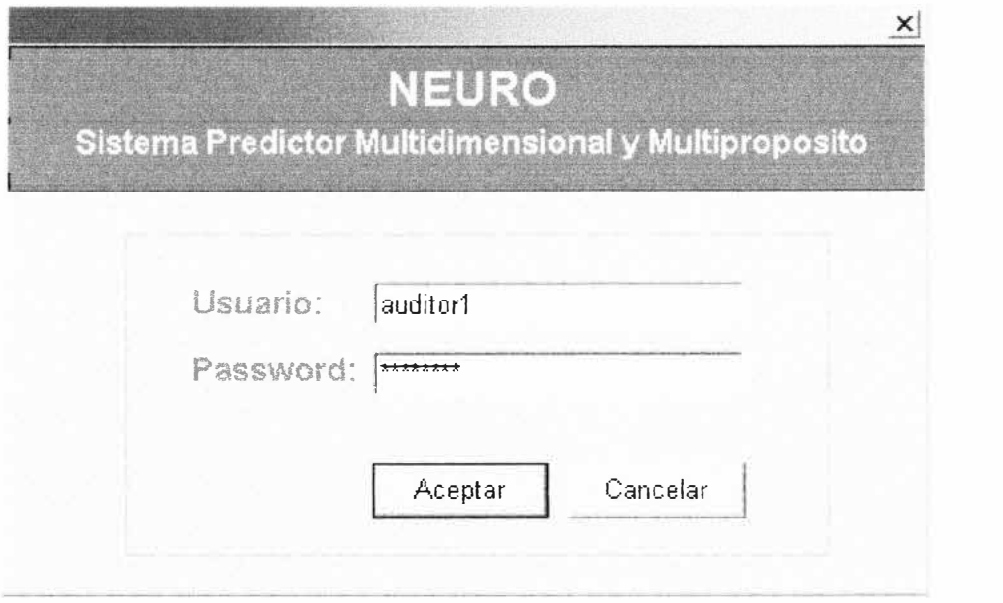
EL SISTEMA PREDICTOR COMO ALTERNATIVA PARA LA IDENTIFICACION DE PATRONES DE EVASION TRIBUTARIA.

Como se mencionó en el capitulo anterior, el sistema desarrollado es multipropósito, es decir puede servir para tareas de predicción en distintas clases de negocio, por ejemplo en el sector financiero, educación, manufactura, agricultura, etc. Solo depende de la selección de las variables relevantes de parte de los especialistas en dichos negocios para definir cuales serian los insumos del sistema predictor neuronal.

Para el caso del presente proyecto de investigación, luego de haber explorado el componente tributario del problema de la evasión, y con el apoyo de algunos analistas tributarios se han seleccionado un primer grupo tentativo de variables relevantes que alimentaran el modelo neuronal, los cuales se han indicado en el capitulo 3.

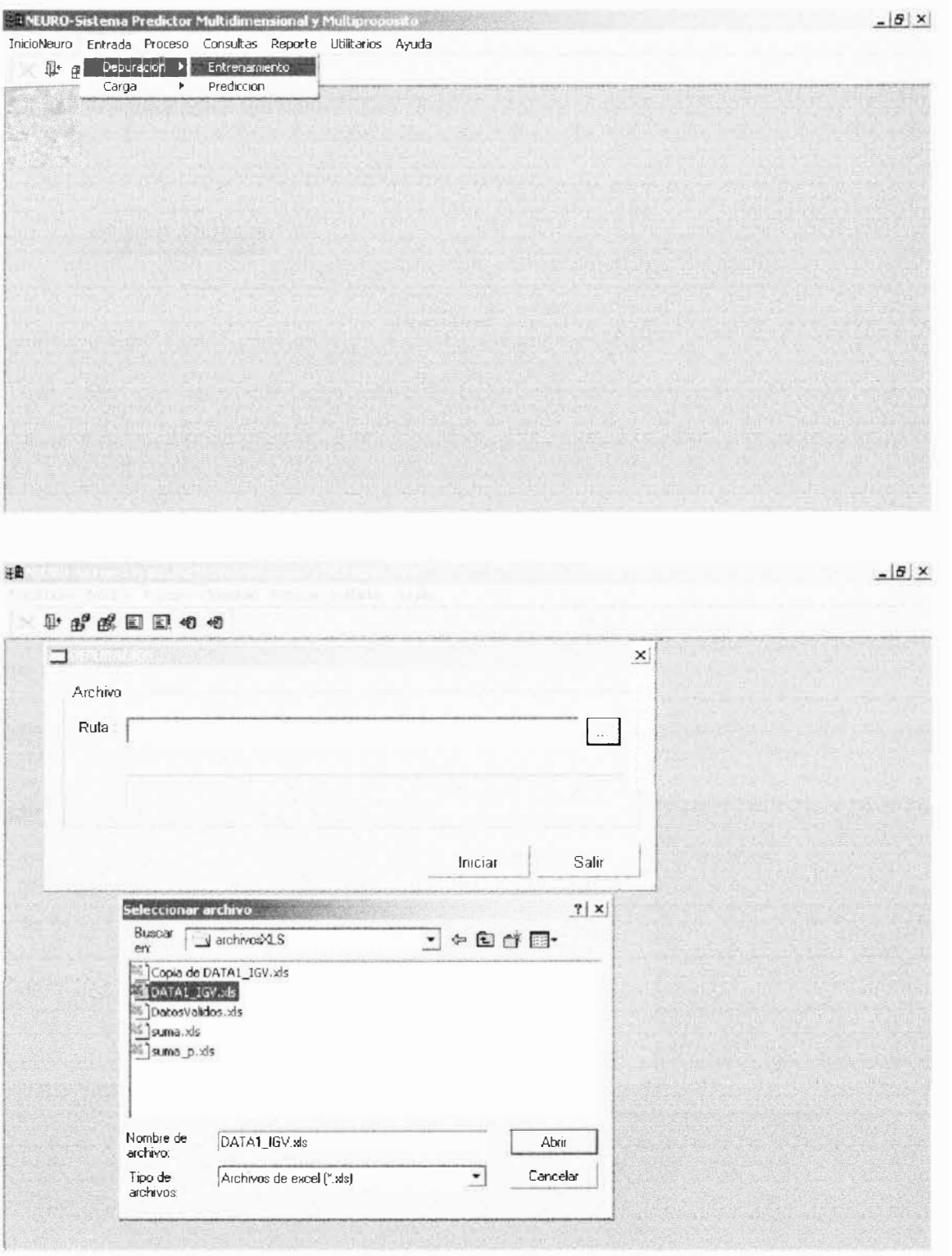
7.1 Uso del sistema predictor para identificar patrones de evasión en el IGV

En las secciones siguientes se mostrará un caso de aplicación del sistema Neuro con propósitos de predicción. El insumo lo constituye un archivo MS-Excel con la nomenclatura ya comentada en el capitulo anterior. Los datos corresponden al tributo IGV para un subconjunto de las variables relevantes: el monto de las ventas totales, el monto de los ingresos no declarados y el monto de ventas netas gravadas. Para mas detalle puede ver el Anexo: "Datos de Entrenamiento" y el Anexo "Datos de Predicción".

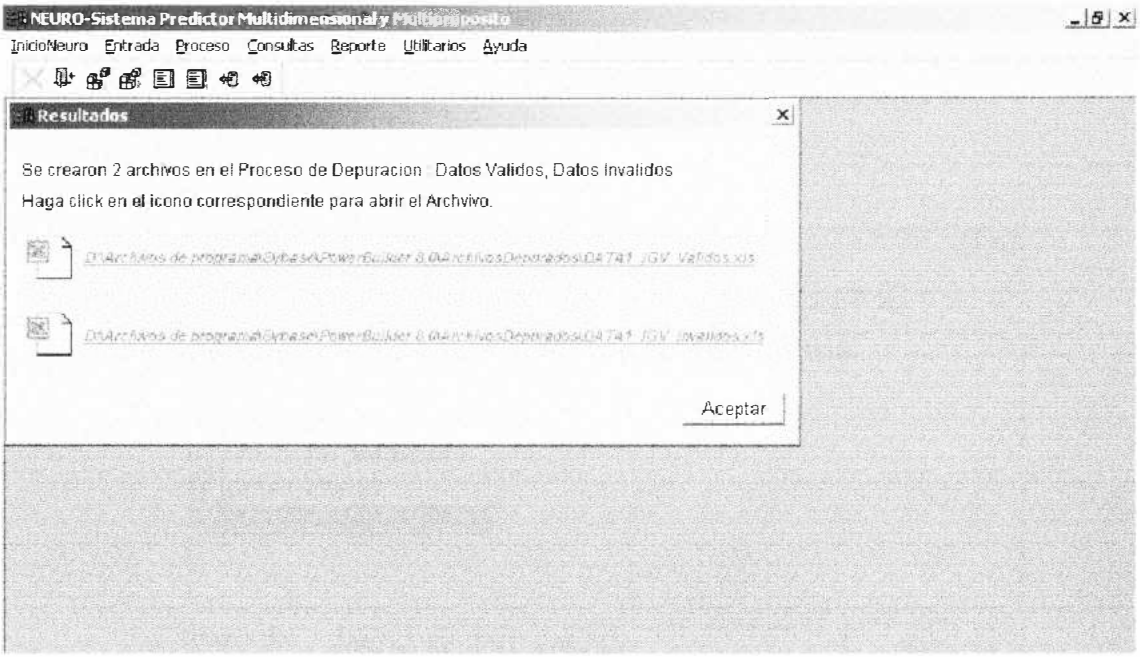


Los ítem que contiene el menú están claramente diferenciados en entrada, proceso (entrenamiento y predicción), consultas, reportes y utilitarios. El primero generara todas las ventanas para el proceso de carga de los datos de un archivo MS-Excel a la base de datos. Para el proceso, en la opción entrenamiento recupera los datos de la base de datos para su posterior utilización, asimismo se guardan aquí los valores del entrenamiento en la base de datos, y para la predicción se recupera los datos cargados para estimar resultados.

7.1.1 Depuración del contenido del archivo de entrenamiento

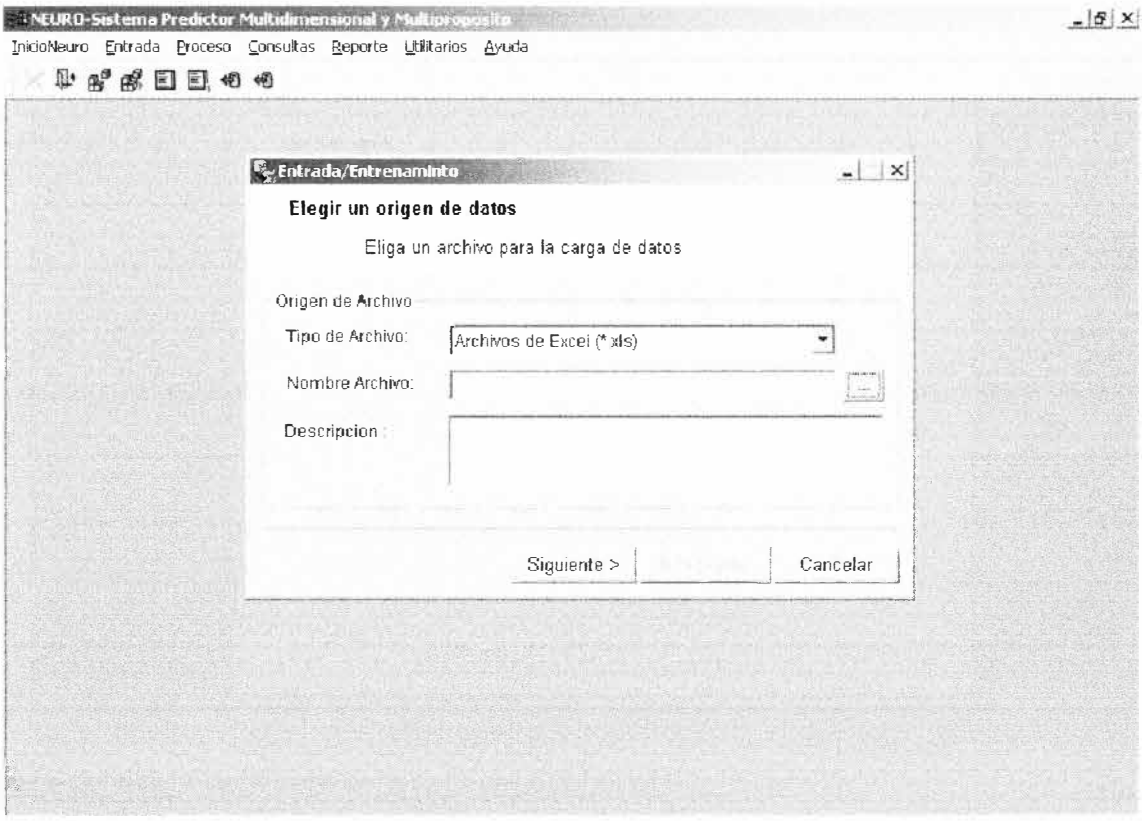

NEURO-Sistema Predictor Multidimensional y Multiproposito
InicioNeuro
Entrada
Proceso
Consultas
Reporte
Utilitarios
Ayuda
Depuracion
Carga
Entrenamiento
Prediccion
Archivo
Ruta
Iniciar
Salir
Seleccionar archivo
Buscar en:
archivosXLS
Copia de DATA1_IGV.xls
DATA1_IGV.xls
DatosValidos.xls
suma.xls
suma_p.xls
Nombre de archivo:
DATA1_IGV.xls
Abrir
Tipo de archivos:
Archivos de excel (*.xls)
Cancelar

Aquí se muestra el termino del proceso de Depuración del archivo, se generan 2 archivos MS-Excel (en la figura mostrada, se puede acceder a estos haciendo clic en el icono correspondiente), un archivo con los datos validos y otro con los datos no validos.

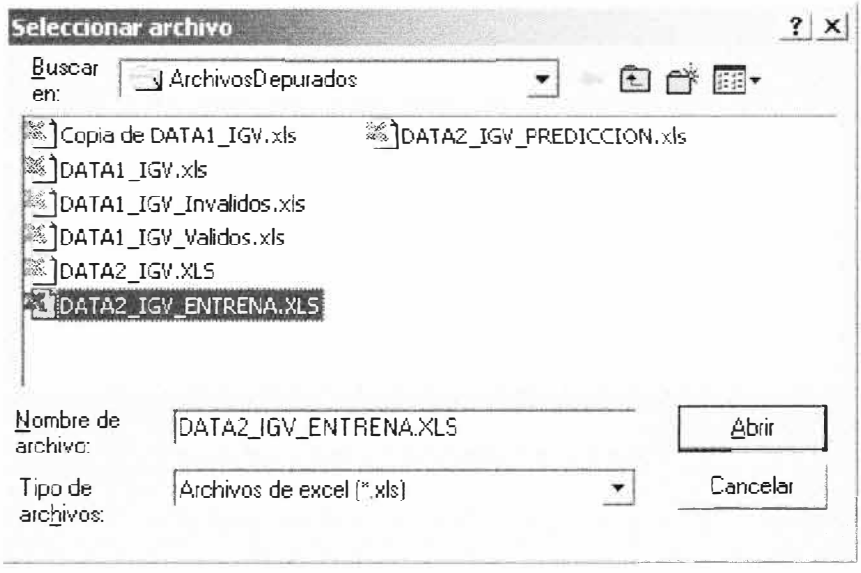


7.1.2 Opción carga del archivo

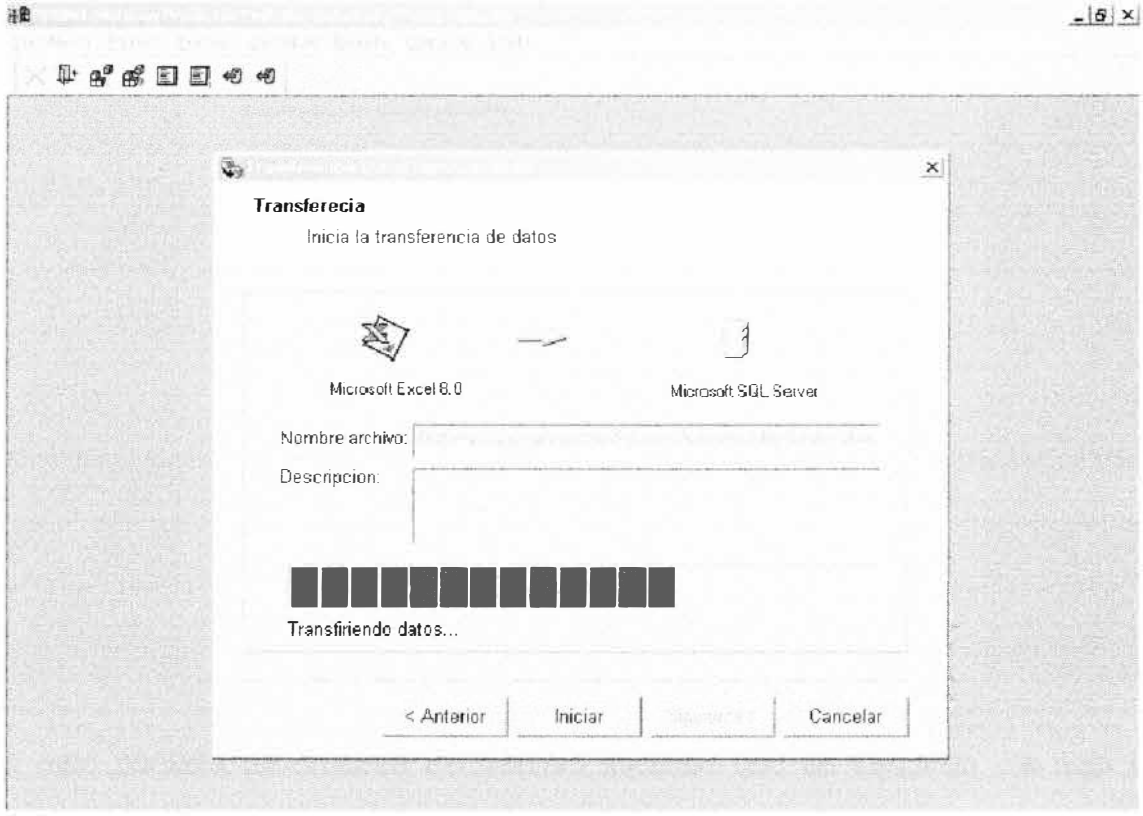
Se inicia el proceso de carga del archivo depurado.

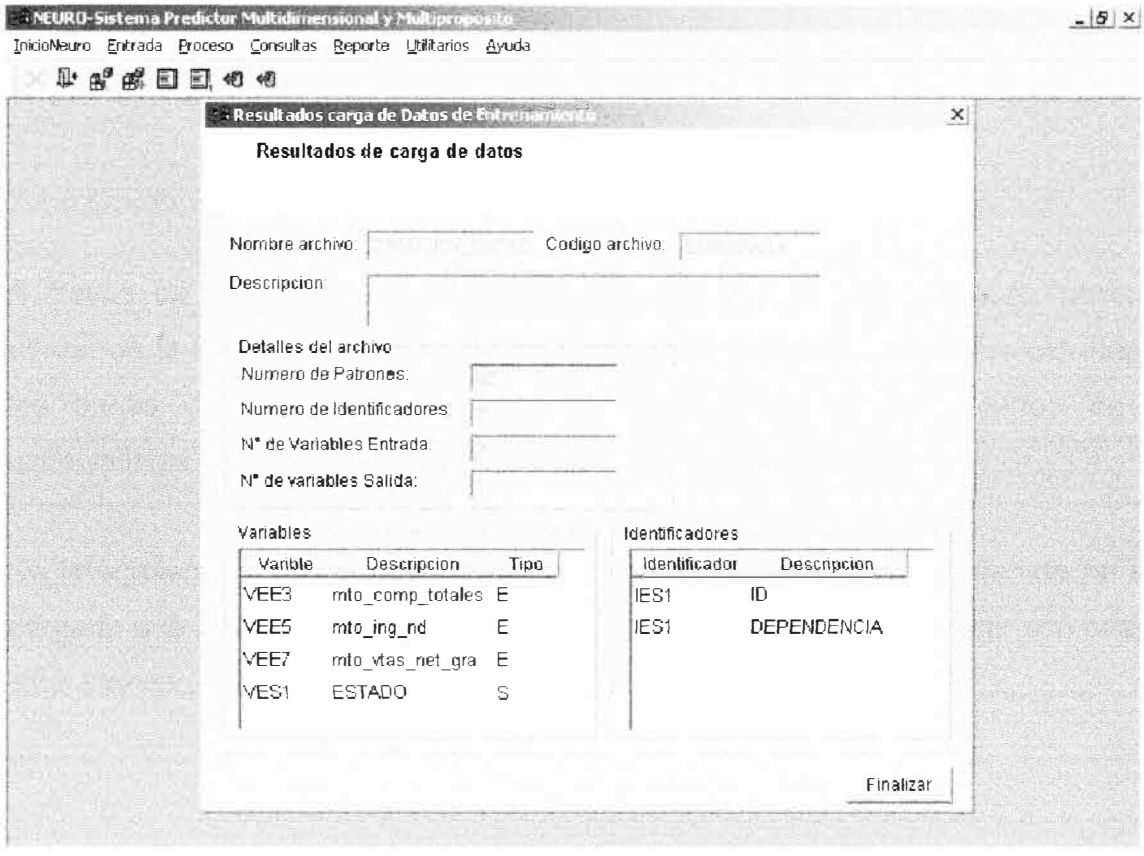


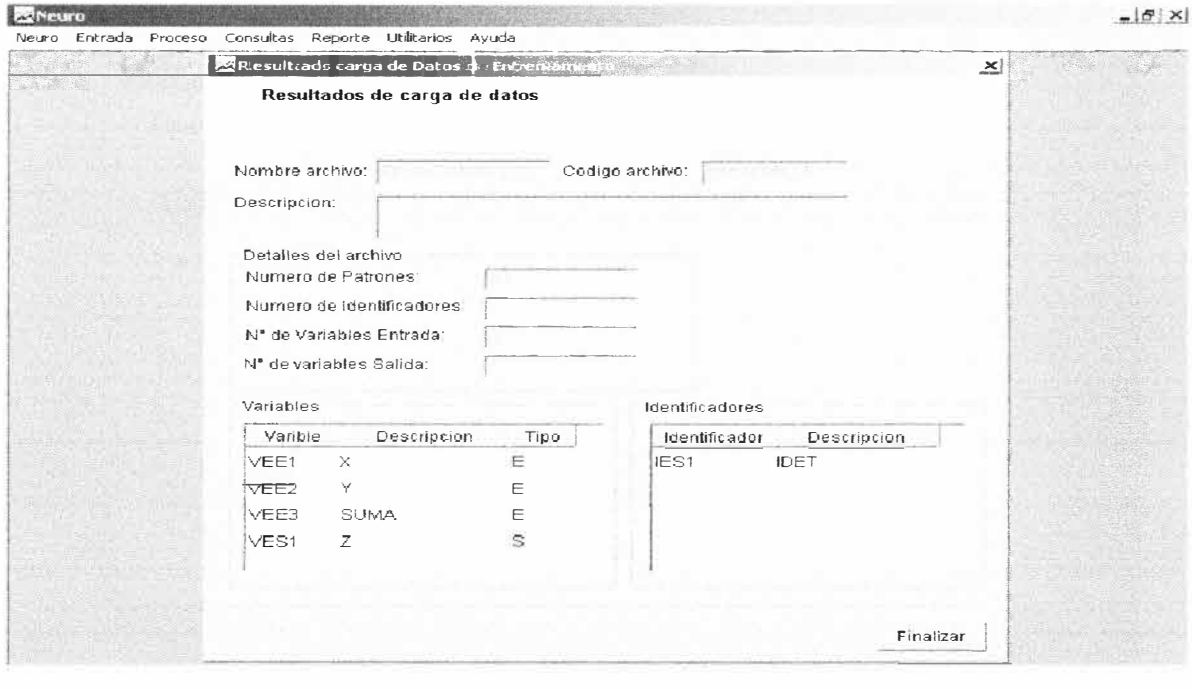
Se selecciona el archivo:



Se inicia el proceso de carga del archivo





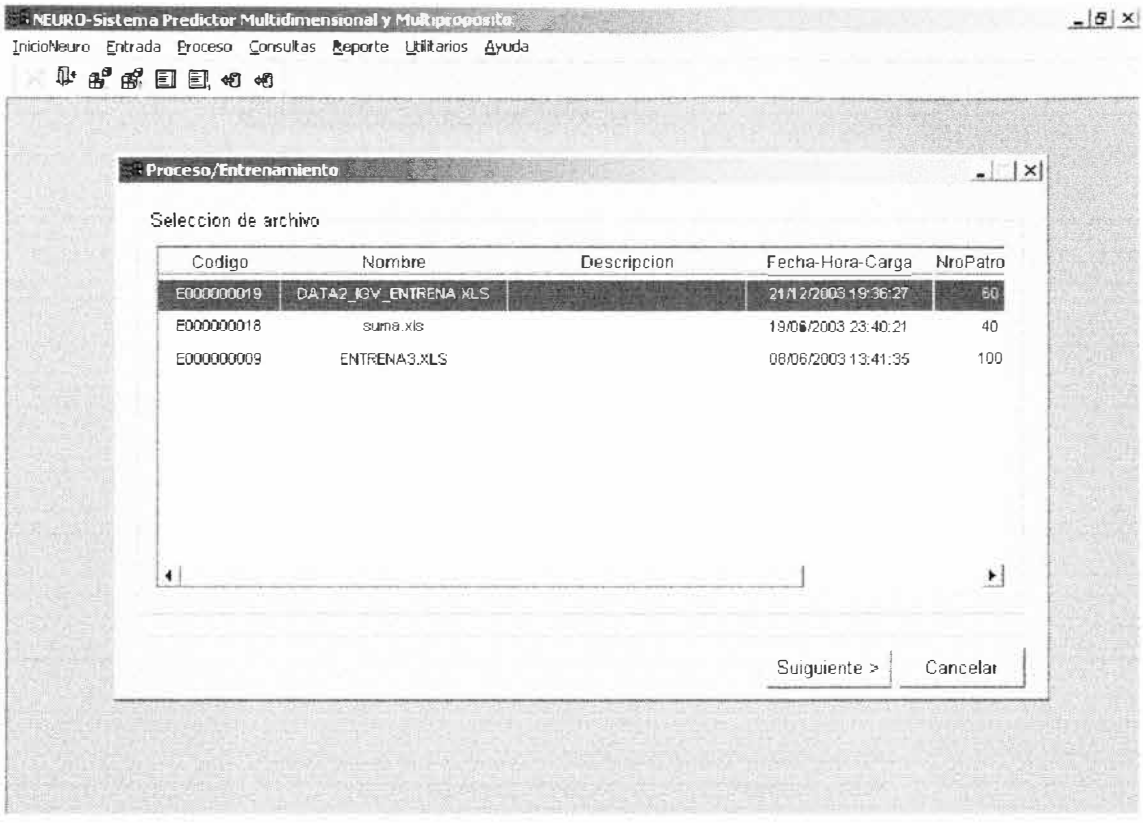


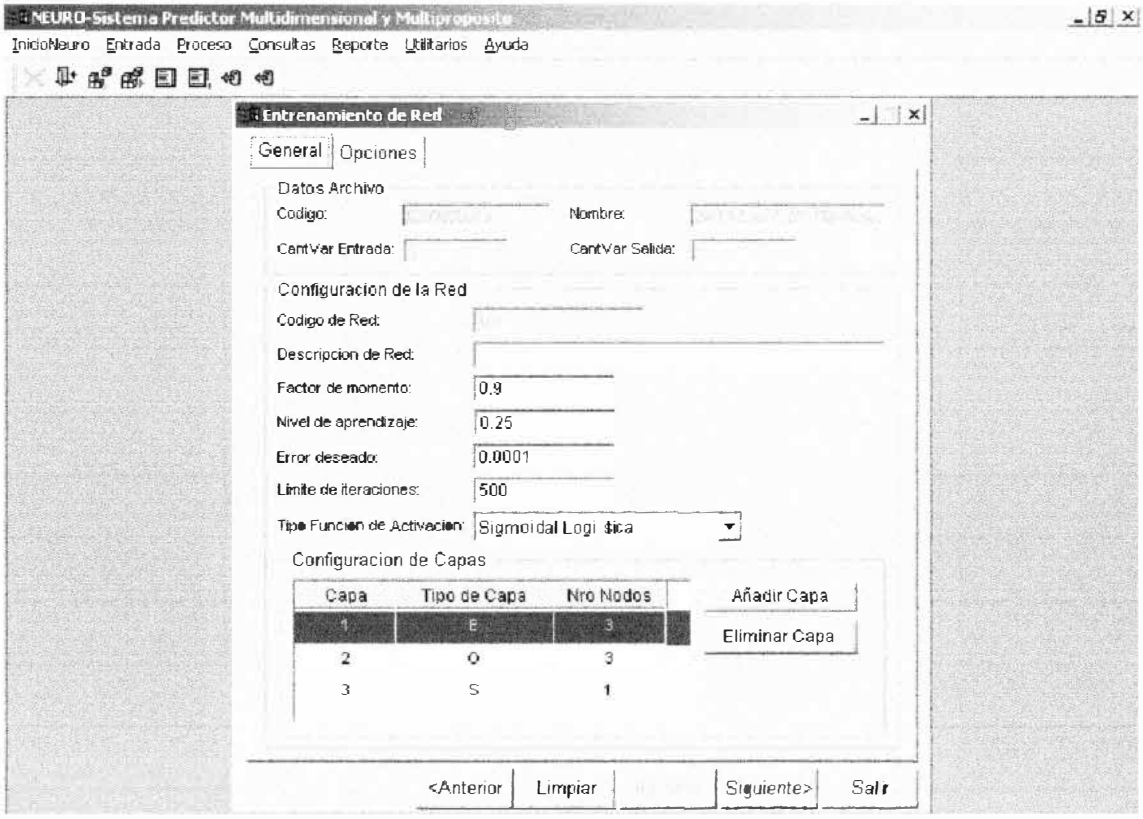
En esta pantalla se muestra los valores iniciales que se cargaron vía hoja de cálculo Excel para el entrenamiento asi como las constantes, las cuales son importantes. Los datos del entrenamiento están dados por "default" sin embargo se pueden cambiar

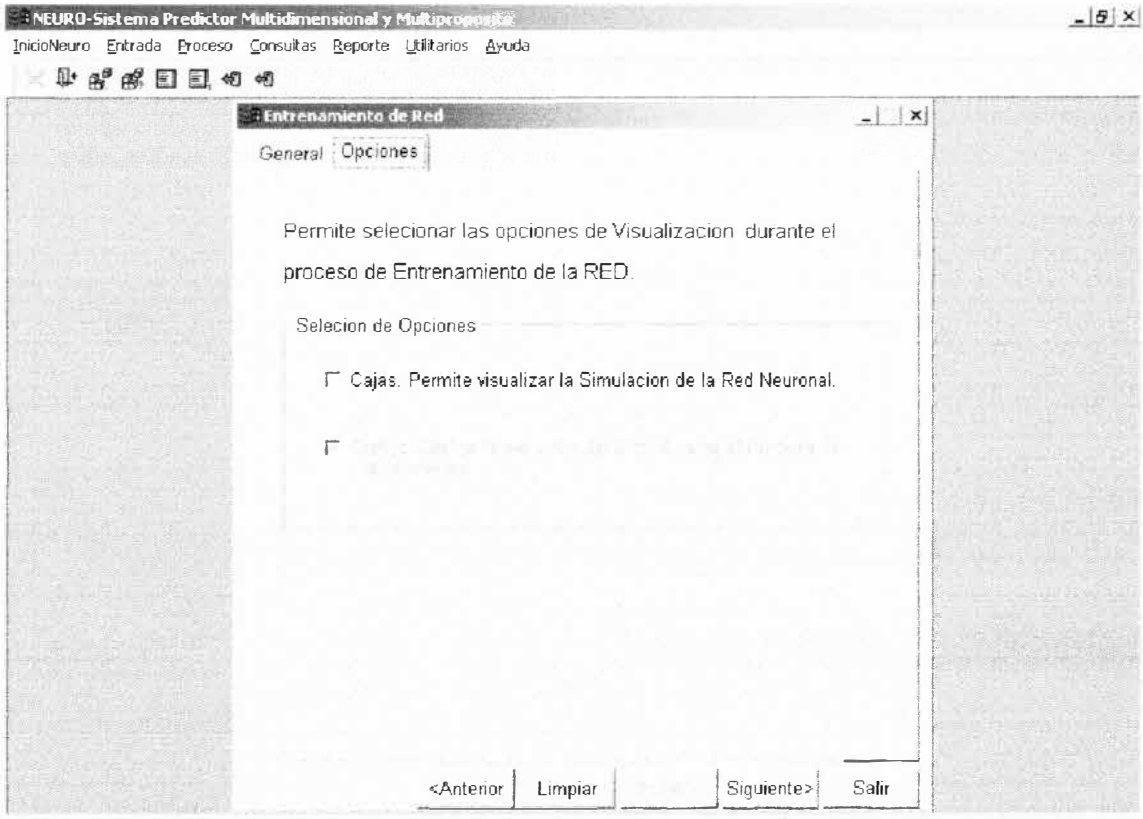
7.1.3 Opción Entrenamiento

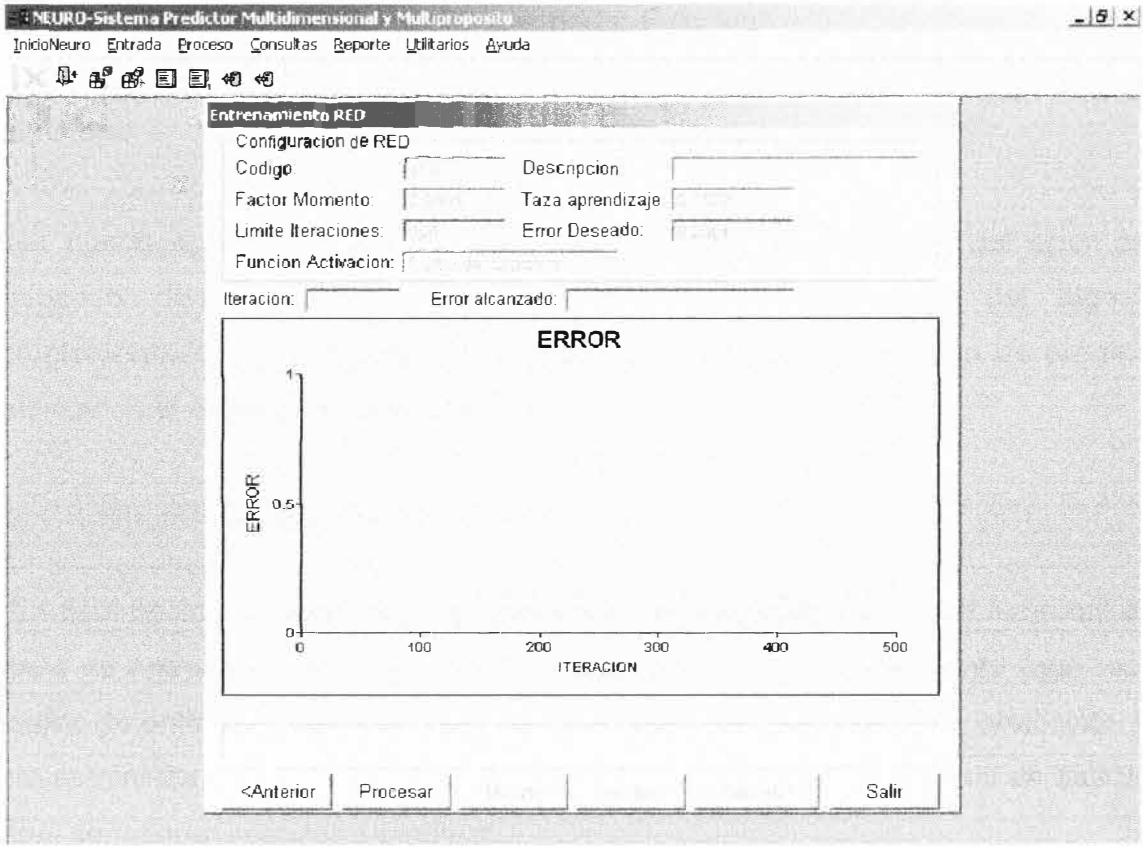
A través de esta opción podremos conseguir que la red neuronal "aprenda" utilizando la información almacenada en la base de datos, la que fue extraída de las bases de datos de la institución basándonos en los criterios de los especialistas tributarios y cargados a través de archivos Excel.

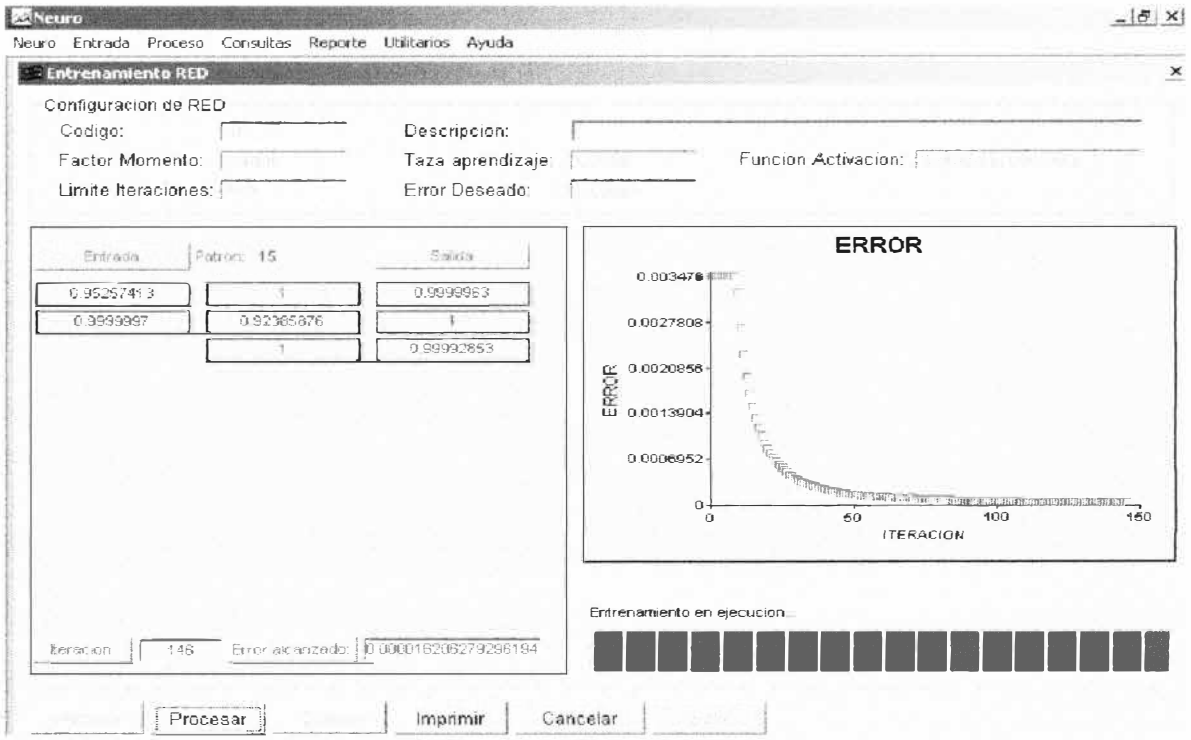
En la siguiente pantalla visualizamos la lista de archivos cuyo contenido ha sido cargado a la base de datos. Por lo tanto, lo que corresponde es elegir con cual de ellos vamos a entrenar a la red.


NEURO-Sistema Predictor Multidimensional y Multiproposito
InicioNeuro Entrada Proceso Consultas Reporte Utilitarios Ayuda
Proceso/Entrenamiento
Seleccion de archivo
Codigo Nombre Descripcion Fecha-Hora-Carga NroPatro
E000000019 DATA2_IGV_ENTRENA.XLS 21/12/2003 19:36:27 60
E000000018 suma.xls 19/06/2003 23:40:21 40
E000000009 ENTRENA3.XLS 08/06/2003 13:41:35 100
Suiguiente > Cancelar


NEURO-Sistema Predictor Multidimensional y Multiproposito
InicioNeuro Entrada Proceso Consultas Reporte Utilitarios Ayuda
Entrenamiento de Red
General Opciones
Datos Archivo
Codigo:
Nombre:
CantVar Entrada:
CantVar Salida:
Configuracion de la Red
Codigo de Red:
Descripcion de Red:
Factor de momento: 0.9
Nivel de aprendizaje: 0.25
Error deseado: 0.0001
Limite de iteraciones: 500
Tipo Funcion de Activacion: Sigmoidal Logi tica
Configuracion de Capas
Capa Tipo de Capa Nro Nodos
1 E 3
2 O 3
3 S 1
Añadir Capa
Eliminar Capa
<Anterior Limpiar Siguiente> Salir


NEURO-Sistema Predictor Multidimensional y Multiproposito
InicioNeuro Entrada Proceso Consultas Reporte Utilitarios Ayuda
Entrenamiento de Red
General Opciones
Permite selecionar las opciones de Visualizacion durante el proceso de Entrenamiento de la RED.
Selecion de Opciones
Cajas. Permite visualizar la Simulacion de la Red Neuronal.
<Anterior
Limpiar
Siguiente>
Salir


NEURO-Sistema Predictor Multidimensional y Multiproposito
InicioNeuro Entrada Proceso Consultas Reporte Utilitarios Ayuda
Entrenamiento RED
Configuracion de RED
Codigo:
Descripcion:
Factor Momento:
Taza aprendizaje:
Limite Iteraciones:
Error Deseado:
Funcion Activacion:
Iteracion:
Error alcanzado:
ERROR
ERROR
1
0.5
0
0
100
200
300
400
500
ITERACION
<Anterior
Procesar
Salir

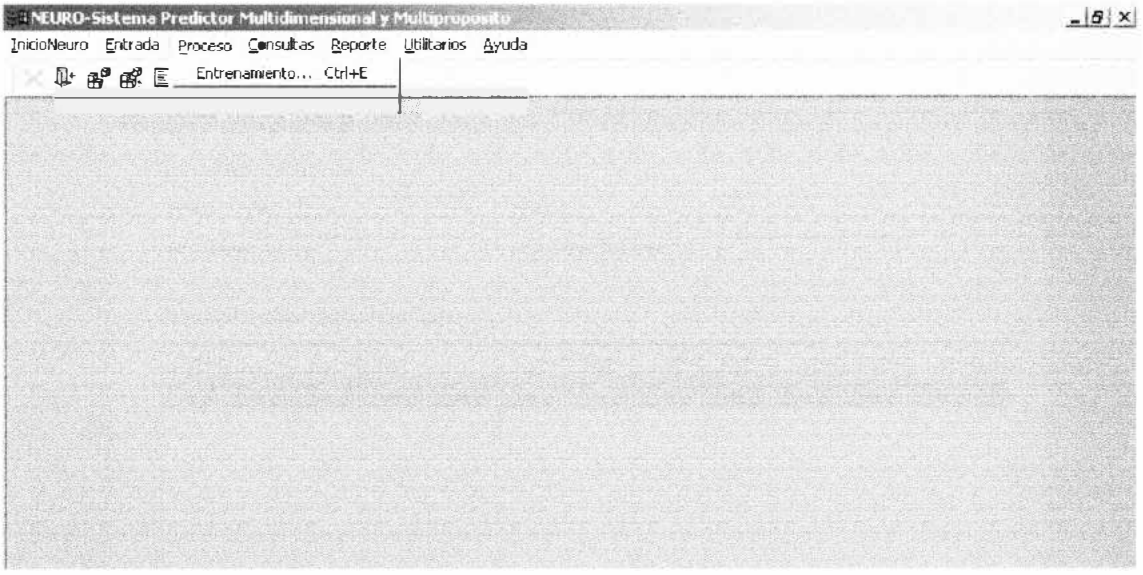


En la pantalla anterior se generan las cajas que muestran la evolución de los valores asociados a cada neurona o nodo. Esto significa la modificación continúa de los pesos, a partir de valores iniciales que fueron inicializados en forma aleatoria.

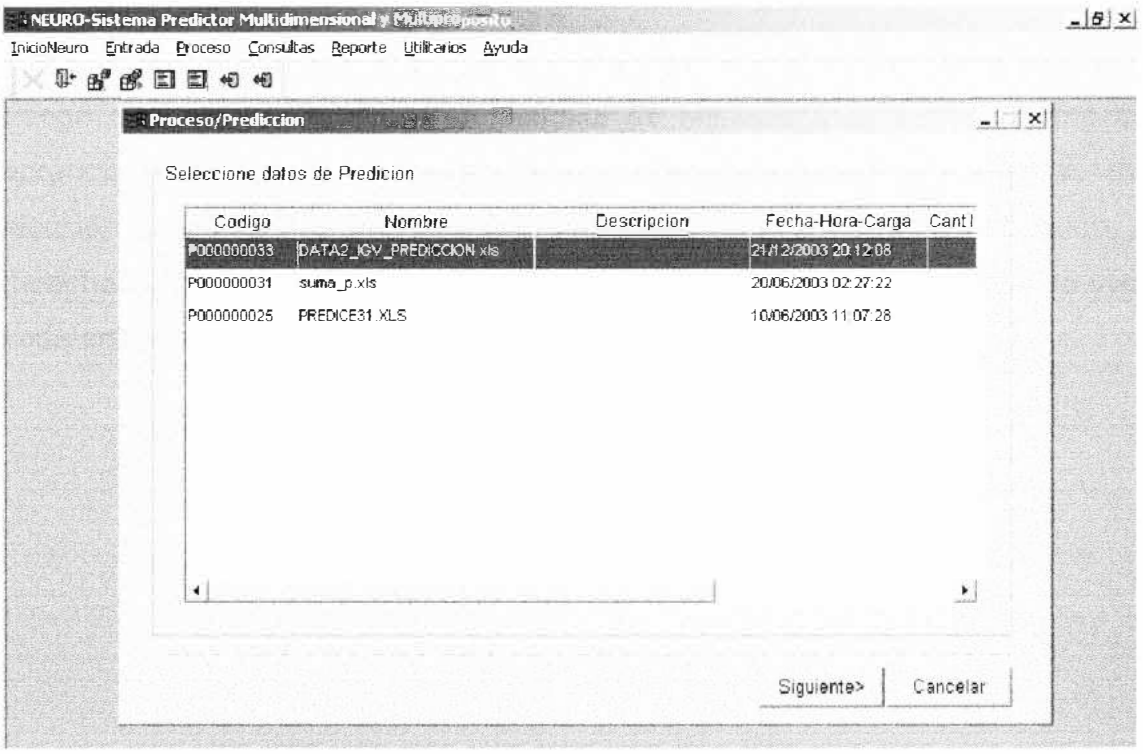
Se puede apreciar el tipo de curva que describe la evolución del error en el proceso de aprendizaje. Esto es señal de convergencia del algoritmo implementado, y por lo tanto, está preparado para cualquier tarea de predicción utilizando el aprendizaje realizado.

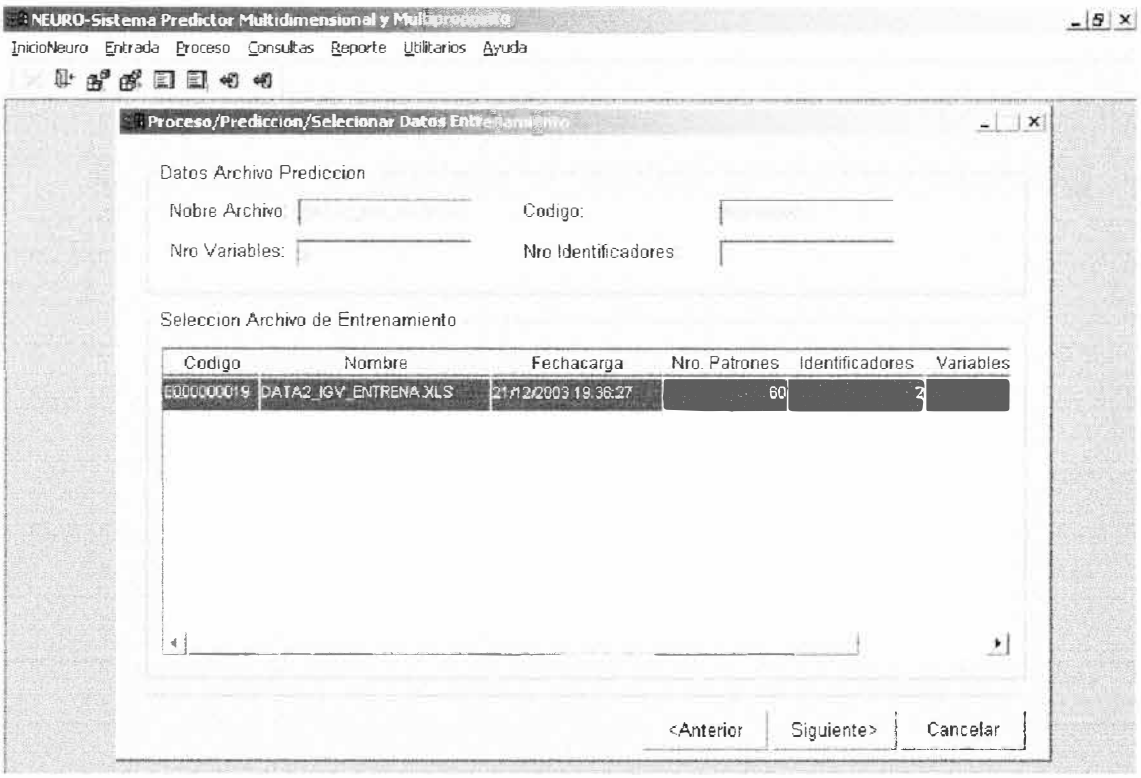
7.1. 4 Opción Predicción

En esta opción se aprovecha "el conocimiento" adquirido por la red neuronal en la fase de entrenamiento. En contraste con la opción de entrenamiento (que recibía datos de entrada y de salida para su entrenamiento), al proceso de predicción solo se suministra datos de entrada, por que es precisamente los valores de salida los que se quieren calcular o predecir.

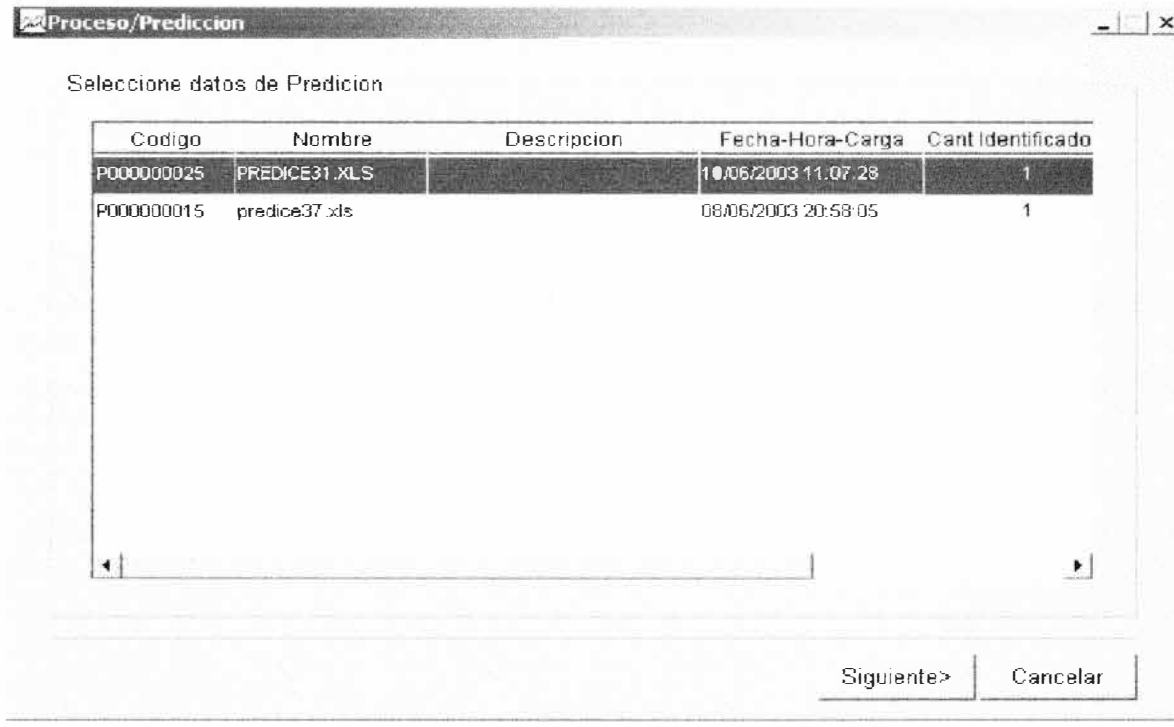


El proceso empieza con la elección de uno de los archivos que han sido cargados para fines de predicción.

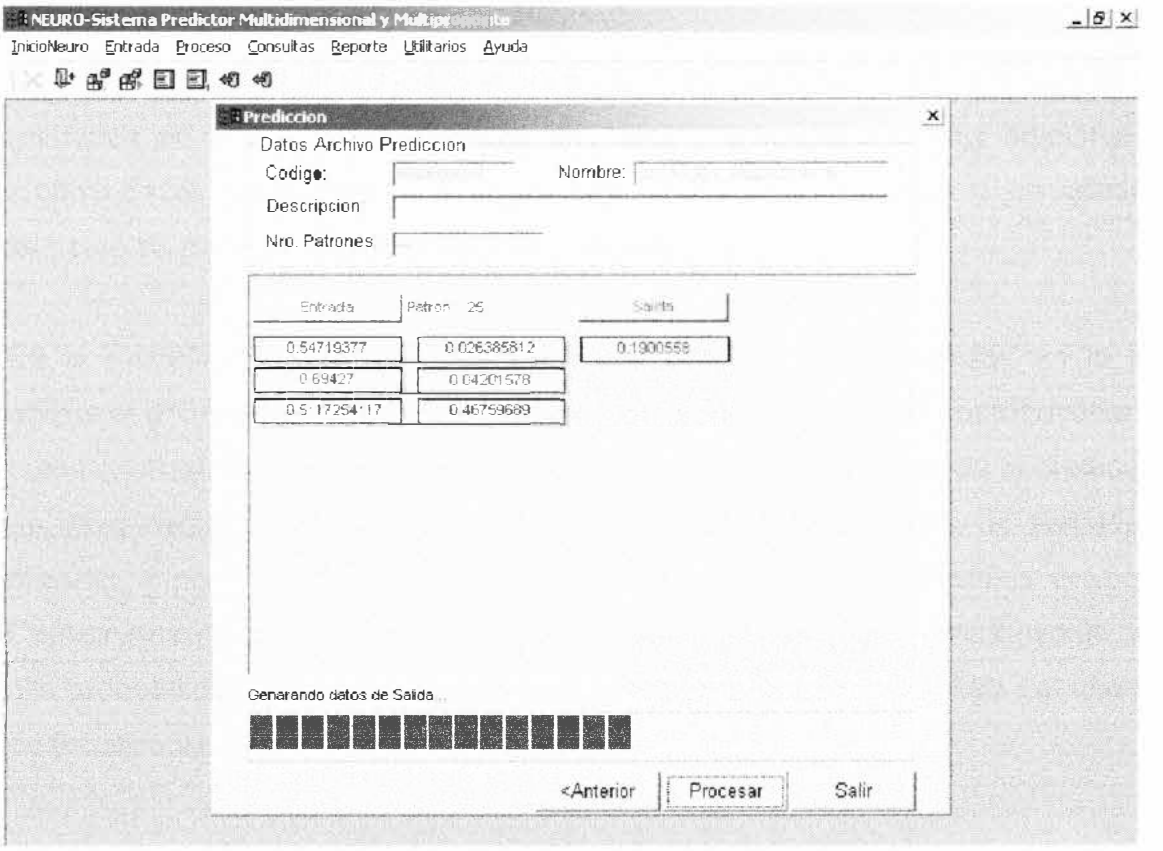


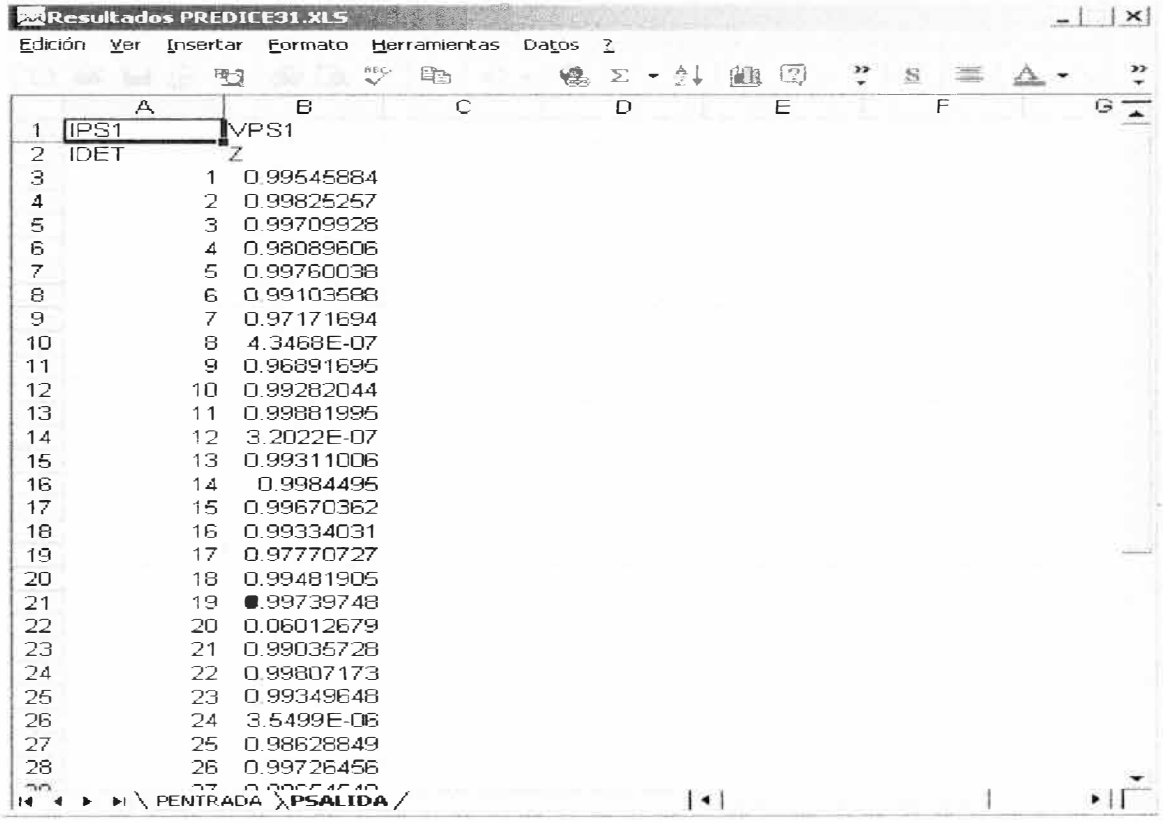


Luego a través de la siguiente pantalla se selecciona el entrenamiento cuyo aprendizaje se quiere aprovechar para predecir. Esto se logra a través de la elección del código de entrenamiento apropiado. Para fines de identificación, a la derecha del código también se muestra el nombre del archivo Excel que se usó en cada entrenamiento:



Luego de la selección del archivo para predicción y del entrenamiento correspondiente, se da inicio a la ejecución de la predicción, mostrándose durante el proceso la evolución de los datos que se quiere calcular y una barra que muestra el porcentaje de avance.





	A	B	C	D	E	F	G
1	IPS1	VPS1					
2	IDET	Z					
3	1	0.99545884					
4	2	0.99825257					
5	3	0.99709928					
6	4	0.98089606					
7	5	0.99760038					
8	6	0.99103588					
9	7	0.97171694					
10	8	4.3468E-07					
11	9	0.96891695					
12	10	0.99282044					
13	11	0.99881995					
14	12	3.2022E-07					
15	13	0.99311006					
16	14	0.9984495					
17	15	0.99670362					
18	16	0.99334031					
19	17	0.97770727					
20	18	0.99481905					
21	19	0.99739748					
22	20	0.06012679					
23	21	0.99035728					
24	22	0.99807173					
25	23	0.99349648					
26	24	3.5499E-06					
27	25	0.98628849					
28	26	0.99726456					

Cuando el proceso de predicción ha concluido, todos los datos calculados son grabados en las tablas de la base de datos y además una hoja adicional del archivo Excel que se uso en la carga. Esto facilita su portabilidad y visualización tal como se muestra en la pantalla precedente.

En la pantalla se muestran dos columnas que han sido grabadas en la hoja adicional (PSALIDA). La primera columna es el identificador del contribuyente (un número correlativo) y la segunda columna representa el resultado de la predicción: un valor cercano a cero o cero se interpreta como la ausencia de señales de evasión, y por lo tanto no amerita incluirlo en la lista de potenciales evasores. Contrariamente, un valor cercano a uno o uno, significa que el contribuyente tiene alta probabilidad de ser evasor, por lo que debe ser incluido en la lista del universo de fiscalizables.

Nótese lo que representa la posibilidad de predecir potenciales evasores, sin recurrir a reglas del negocio tributario que en el esquema vigente significa establecer muchas relaciones con abundantes conectores lógicos, y una combinación intensiva de datos de las variables relevantes. En el sistema basado en una red neuronal, se proporciona los datos de las variables y se deja que el sistema haga el trabajo en sustitución del especialista.

CAPITULO VIII

ANÁLISIS COSTO / BENEFICIO

8.1 Costos de investigación para la incorporación de la tecnología Data Mining con Redes Neuronales

La exploración de tecnologías emergentes para innovar esquemas vigentes, tiene un costo inherente a todo un ciclo de investigación. La aventura, inicialmente incierta, de examinar alternativas tecnológicas, pasa inevitablemente por evaluar varias opciones, con un despliegue de esfuerzo y recursos, hasta culminar con la elección de la mejor opción para la consecución de los objetivos planteados.

Mencionaré escuetamente los costos explícitos e implícitos asociados al proyecto. Cabe enfatizar que el proyecto solo cubre la fase de exploración, ya que es precisamente un proyecto de investigación, y lo que se ha elaborado es un prototipo, con perspectivas de implantación futura. Por lo tanto, solo se indicaran los costos incurridos en la fase exploratoria.

Mencionaremos aquí todos aquellos costos "visibles" y que en cierta medida son cuantificables. Se enunciará los costos clasificados en software, hardware, recursos humanos (investigadores, desarrolladores) y elementos complementarios (materiales de consulta, materiales de escritorio y otros).

8.1.1 Software

Durante la fase de investigación, cada analista (tributario o de sistemas) y el desarrollador y probadores del software utilizan una PC asignada en forma

exclusiva. La plataforma de software base, está conformado por los sistemas operativos Windows NT, Windows 2000 y UNIX. El software adicional, para el caso de los analistas tributarios es básicamente de documentación: MS Office. El software adicional empleado por los analistas de sistemas es de modelamiento y documentación (MS Office, Visio, etc); algunos de estos analistas con frecuencia asume el rol de coordinador o líder de proyecto. Para el caso de los programadores el software adicional es esencialmente de programación y documentación (Informix-4GL, Reflect, MS Office). Es necesario precisar, que una misma persona con frecuencia alterna como analista y como programador, empleando por lo tanto, una combinación de ambos grupos de software. Las personas involucradas con la prueba del software (probadores) usan esencialmente el Reflect y el prototipo desarrollado, y para la documentación (preparar manuales de usuario, informes) usan básicamente el MS-Office.

El costo por lo tanto, esta constituido por la inversión en software necesario para el desarrollo de la investigación. En términos estrictos no se ha adquirido nada nuevo, por que ya se disponía de licencias de tales herramientas. Sin embargo, para fines de simplicidad, incluiremos una parte del costo de la licencia de cada software, como "consumido" por el proyecto. Tener presente que justamente esa "porción" del precio de la licencia es la que directamente se indica en los cálculos, y no el precio total, dado que hay otros proyectos y aplicaciones que los usan.

Power Builder : (US$ 200 / licencia) * (3.5 soles / US$) * 2 licencias = 1400 soles

SQL- Server: (US$ 300 / licencia) * (3.5 soles / US$) * 3 licencias = 3150 soles

Reflect: (US$ 150 / licencia) * (3.5 soles / US$) * 1 licencia = 525 soles

Sistemas Operativos (Windows, UNIX): US$ 200 * (3.5 soles / US$) = 700 soles

MS Office, diagramadores,etc: US$ 100 * (3.5 soles/US$) * 2 usuarios = 700 soles

Total Costo Software = 6475 soles

Todo el software mencionado tiene una "vida útil" variable. El ritmo de obsolescencia es mayor en unos que en otros. Se asumirá para fines de simplificar el análisis de costos, una tasa de depreciación lineal promedio de 3 años. Esto significa que al cabo de ese tiempo el software podrá ser removido y/o actualizado por estar obsoleto. También, es necesario hacer mención que la adquisición de software original proporcionada el beneficio de obtener actualización de versiones a costos ventajosos, y descuentos especiales por adquisición de gran cantidad de licencias y también por instalación en red.

Cuadro 8.1

Tipo de Software	Costo Unitario (US $)	Tiempo Deprec. (años)	Tiempo Deprec. (meses)	Deprec. Mensual (US $)	Deprec. Mensual (Soles *)	Deprec. Diaria ** (Soles)
Software de Base	600	4	48	12.5	31.25	1.42
Software de Análisis y Diseño	400	3	36	11	27.5	1.25
Software de Programación	500	3	36	13.8	34.5	1.56
Software de Prueba	500	3	36	13.8	34.5	1.56
Software de Documentación	300	3	36	8.3	20.75	0.94

(*) Para fines de conversión de dólares a soles, se está asumiendo un tipo de cambio promedio de 2.50 para los años 2001 y 2002, y 3.5 para el 2003, intervalo del análisis de costos.

(**) Se considera un mes de 22 días laborables.

Para no complicar el cálculo de los costos asociados al software, que como se explicó tienen usos diversos dependiendo de la función que desempeña la persona que lo utiliza, procederemos a calcular la Depreciación Diaria Promedio del Software (DDPS):

DDPS = Promedio de las depreciaciones individuales por unidad de los tipos de software

DDPS = (1.42 + 1.25 + 1.56 + 1.56 + 0.94) / 5

DDPS = 1.35 (Soles x día)

8.1.2 Hardware

En el caso del hardware empleado, para el caso de los servidores se usara un criterio de "tasa de uso" (porcion utilizado por el proyecto) similar al del software. Los montos que aparecen a continuación son directamente la "porcion" usada por el proyecto.

Servidor NT: US$ 100 * (3.5 soles / US$) = 350 soles

Servidor UNIX: US$ 50 * (3.5 soles / US$) = 175 soles

Para el caso de las computadoras personales (PCs) se considera que el proyecto en promedio usa 3 PCs en su totalidad, pero como el intervalo fue de dos años aproximadamente, y asumiendo que el hardware tiene una tasa de depreciación de 4 años, se aplicará un factor de 0.5

PCs: 3 * US$ 1200 * (3.5 soles / US$) * 0.5 = 6300 soles

Total Costo Hardware = 6825 soles

8.1.3 Recursos Humanos

Para la estimación del costo en recursos humanos, debe tenerse en cuenta que solamente ha habido una dedicación a tiempo parcial al proyecto. Por lo tanto, lo mas apropiado será utilizar una medición por hora, considerando salarios promedio por hora y la cantidad total de horas acumulada para el proyecto.

Una estimación gruesa del tiempo, se basa en la utilización de fines de semana principalmente. Para el caso del investigador, en los dos años, con un promedio de 50 semanas, tendríamos 100 semanas. En cada fin de semana se utilizó en

promedio 12 horas, por lo que en forma acumulada son 1600 horas del investigador.

Investigador principal = 100 semanas*(12 horas/semana)*(35 soles/hora)= 42000 soles

Los analistas tributarios, han tenido una intervención intermitente, por lo que el total de horas a considerar no excederán las 80 horas.

Analistas tributarios = 80 horas * (30 soles/hora) = 2400 soles

Para las dos personas que apoyaron en el prototipeo, en forma alternada, y en épocas especificas, la dedicación ha sido también generalmente los fines de semana, con una ratio de horas de aproximadamente 10 horas por semana, considerando un total de 50 semanas (1 semestre para cada uno).

Prototipeadores= 50 semanas * (10 horas / semana)*(10 soles/hora)=5000 soles

Total Costo Recursos Humanos = 49,400 soles

8.1.4 Otros Gastos

Toda fase de investigación, consume recursos adicionales al hardware y software propiamente dicho. Se mencionaran individualmente los que tengas relevancia, y los restantes se agruparan en varios:

Internet

Se han utilizado muchas horas. Se ha estimado un aproximado de 800 horas. Asumiendo el costo promedio por hora de 1.5 soles, el costo asociado sería:

Costo Internet = 800 horas * (1.5 soles / hora) = 1200 soles

Libros

Además de usar la opción de consulta en las bibliotecas, por comodidad de consulta, se ha tenido que adquirir un total de 6 libros referidos a temas entre tecnológicos, tributarios y sistémicos. Para fines de simplificación se indicara el promedio de los precios de los libros; este promedio es US$ 30. Por lo tanto, el costo o inversión asociado a los libros seria:

Costo Libros = 6 libros * (US$ 30 / libro) * (3.5 soles / US$) = 630 soles.

Varios

Papel, tinta, útiles de escritorio, etc = 500 soles

Total Otros Gastos = 2,330 soles

8.1.5 Costo total

Costo Total = Costo Software + Costo Hardware + Costo Recursos Humanos + Otros Gastos

Costo Total = 65,030 soles

8.2 Beneficios tangibles e intangibles de contar con un nuevo esquema de identificación de patrones de evasión.

8.2.1 Beneficios tangibles potenciales

Los beneficios potenciales que se pueden cuantificar son aquellos derivados del ahorro que significará la disminución de auditorias no exitosas (sin reparos). Por ejemplo, si antes del 100% de casos revisados por el auditor, solo encontraba reparos ("indicios" de evasión) en el 80% de los casos, y en el 20% restante no

encontraba ningún reparo, definitivamente este último 20% fue incorrectamente seleccionado, y se dedicaron recursos innecesariamente para concluir que "todo es conforme".

El nuevo esquema definitivamente no reducirá a cero la tasa de auditorias sin reparos ("fracasos"), pero si es una promesa para reducir el margen a una tasa de 10% en promedio. Una mejoría sustancial sin duda, sin consideramos que en algunos casos pueda llegar a niveles del 5% o incluso menos. La clave es el uso de nueva tecnología soportado por un nuevo esquema.

Ahorro en salarios de auditoria y revisión de casos.

Considerando que la labor de auditoria no solo involucra al auditor, sino también (dependiendo del monto a niveles jerárquicamente superiores), se muestra el grado de intervención de cada uno de ellos.

Monto de Referencia	Auditor	Supervisor	Jefe de Sección	Jefe de División
< 300,000 (Caso1)	SI	SI	NO	NO
< 500,000 (Caso2)	SI	SI	SI	NO
<1,000,000 (Caso3)	SI	SI	SI	SI

Nota: Para montos mayores a 1000000 interviene también el Intendente (firma)

Por otro lado, tomando como referencia que el tiempo promedio otorgado a un auditor para revisar un caso de IGV es de 3 días, es decir 24 horas, y teniendo presente que ha medida que se asciende en la escala jerárquica hay participación en la revisión de algunos casos pero con una menor cuota de tiempo, se han

estimado en el siguiente cuadro la cantidad de horas que cada involucrado dedica, según el tipo de caso (son horas promedio).

Monto de Referencia	Auditor	Supervisor	Jefe de Sección	Jefe de División
(Caso1)	24 horas	4 horas	NO	NO
(Caso2)	32 horas	5 horas	2 horas	NO
(Caso3)	40 horas	6 horas	3 horas	2 horas

Considerando salarios promedio por hora de: 20 (auditor), 30 (supervisor), 40 (jefe de sección) y 50 (jefe de division), se tiene el siguiente cuadro de costos por caso:

Monto de Referencia	Auditor	Supervisor	Jefe de Sección	Jefe de División
(Caso1)	480	120	NO	NO
(Caso2)	640	150	80	NO
(Caso3)	800	180	120	100

El costo total por caso seria:

	Caso 1	Caso 2	Caso 3
Costo	600	870	1200

Son cifras conservadoras, considerando que tampoco se ha considerado las horas del personal de programación operativa para la selección de los casos.

Aprecie el costo involucrado a un caso que luego de toda la labor de auditoria resulta sin reparos ("fracaso"). En estos casos, la administración tributaria no recuperara absolutamente nada. Distinto es el caso del costo desplegado para obtener reparos que pueda cubrir en parte el costo en algunos casos o en otros casos en forma holgada y con creces.

Lógicamente con el nuevo esquema, también habrá revisiones similares, pero la gran diferencia es que la tasa de auditorias con fracaso será menor, y por lo tanto, los costos irrecuperables serán menores.

8.2.2 Beneficios intangibles potenciales

La implantación masiva (a nivel nacional) del nuevo esquema de selección de contribuyentes fiscalizables permitirá:

Generar sensación de riesgo en el contribuyente. El sentirse controlado se traduce en una necesidad de cumplir cabalmente con las obligaciones tributarias.

Lograr oportunidad en la identificación de los casos (contribuyentes y montos) resultantes de la diferencia entre lo declarado y lo pagado.

Coadyuvar al incremento de la recaudación tributaria, a través de la modalidad de inducción al pago.

Disminuir la cantidad de analistas tributarios dedicados a la selección de contribuyentes fiscalizables. Las personas excedentes son reubicadas para desempeñar otras funciones.

Mejorar la imagen de la Administración Tributaria. La recurrente percepción del contribuyente, como ente burocrático y desinformado, se va disipando progresivamente.

CAPITULO IX

CONCLUSIONES Y RECOMENDACIONES

9.1 Conclusiones

- Entre los múltiples problemas que tiene que afrontar la Superintendencia Nacional de Administración Tributaria (SUNAT), una de particular importancia es la evasión tributaria. El primer reto ha sido definir adecuadamente el "problema evasión". Fue muy útil la aplicación de parte de la Metodología de Sistemas Blandos (MSB) para la definición del problema e identificación de variables, a partir de la recolección, conciliación e integración de las múltiples percepciones de los diferentes agentes tributarios.

- La evasión es un problema recurrente en la Administración Tributaria. La tecnología Data Mining es una interesante opción para reducir sustancialmente la tasa de desaciertos en la selección de contribuyentes evasores potenciales y por lo tanto reducir los costos de los programas de fiscalización y auditoria.

- Los márgenes de error de las predicciones hechas con el sistema basado en redes neuronales son menores a los obtenidos por métodos estadísticos de predicción.

- No existe una arquitectura ideal de red neuronal para todo tipo de aplicaciones. El mejor modelo se va obteniendo a través sucesivos entrenamientos, por ensayo y error. Sin embargo, en la mayoría de las veces una red con una o dos capas ocultas, y dos o tres neuronas en cada una de ellas, ha sido suficiente para obtener resultados satisfactorios.

- Un componente crítico para la adecuada utilización de un sistema basado en redes neuronales es el tratamiento de los datos que servirán de insumo del modelo neuronal. Es fundamental la consistencia, el escalamiento y la normalización de los datos antes del entrenamiento y predicción.

9.2 Recomendaciones

- Todo proyecto que intente implementar soluciones con tecnología Data Mining debe evaluar la técnica o técnicas que mas se ajustan a las necesidades de la organización, teniendo en cuenta sus restricciones de recursos y sobre todo el nivel de conocimientos que sobre el particular poseen sus cuadros profesionales.

- Si una organización considera atractiva implementar Data Mining usando redes neuronales, debe asegurarse de contar con bases de datos con suficiente información histórica que permita un adecuado entrenamiento del sistema que construirán. Adicionalmente, debe identificarse los intervalos en los que cambian las reglas del negocio, para tenerlo en cuenta en las condiciones de extracción de datos.

- La implementación de un sistema de predicción usando redes neuronales requiere la interacción con expertos en las reglas del negocio para que seleccionen las variables relevantes que serán usadas como insumo de la red neuronal. Lo anterior puede pasar por la definición de indicadores o ratios que serán calculados a partir de los datos almacenados en las bases de datos históricas.

GLOSARIO DE TERMINOS

- IGV: Impuesto General a las Ventas
- MSB: Metodología de Sistemas Blandos
- SCP: Sistema Contenedor de Problemas
- SSP: Sistema Solucionador de Problemas
- IPESAT: Identificación de Patrones de Evasión en el Sistema de Administración Tributaria.
- IA: Inteligencia Artificial
- RNA: Redes Neuronales Artificiales
- COA: Confrontación de Operaciones Autodeclaradas
- DAOT: Declaración Anual de Operaciones con Terceros.

REFERENCIAS BIBLIOGRÁFICAS

TOPICOS TRIBUTARIOS

REVISTA DEL INSTITUTO PERUANO DE DERECHO TRIBUTARIO – Nro. 32
Apuntes sobre la vigencia y limites del llamado "delito contable" - Junio 1997

EVALUACIÓN DE LA CAPACIDAD RECAUDATORIA DEL SISTEMA TRIBUTARIO Y DE LA EVASIÓN TRIBUTARIA
Michael Jorrat De Luis
Servicio de Impuestos Internos – Chile
Conferencia Técnica Centro Interamericano de Administradores Tributarios – CIAT - 1996

LA EVASIÓN TRIBUTARIA
Jorge Cosulich Ayala
Proyecto Nacional de Política Fiscal – CPAL – PNUD
Serie Política Fiscal – Nro 39 – Nov. 1993

ESTIMACIONES DE LA EVASIÓN EN COLOMBIA
V Seminario Regional de Política Fiscal, Estabilización y Ajuste
Conferencia Técnica Centro Interamericano de Administradores Tributarios – CIAT - 1993

ESTIMACIÓN DE LA EVASIÓN EN EL IVA EN CHILE (1980-1993)
Juan Toro y Jorge Trujillo
Comisión Económica para América Latina y el Caribe CPAL / PNUD

EVASIÓN FISCAL EN MÉXICO
Alma Rosa Moreno
Mario Gabriel Bubedo
Proyecto Nacional de Política Fiscal – CPAL – PNUD
Serie Política Fiscal – Nro 41 – Dic. 1993

TOPICOS SISTÉMICOS

TEORIA GENERAL DE SISTEMAS APLICADA
John P. Van Gigh
Prentice Hall - 1979

METODOLOGIA DE SISTEMAS SUAVES EN ACCION
Peter Checkland
Jim Scholes
Megabyte Noriega Editores – México 1994

METODOLOGÍA DE LA INGENIERIA DE SISTEMAS
Arthur Hall

LA QUINTA DISCIPLINA
Peter Senge
1994

LA DANZA DEL CAMBIO
Peter Senge, A Kleiner
1999

TÓPICOS TECNOLÓGICOS

DATA WAREHOUSING
La Integración de información para la mejor toma de decisiones
Harjinder S. Gill
Prakash C. Rao
Prentice Hall Hispanoamericana S.A. - 1998

BUILDING THE DATA WAREHOUSE.

Inmon

QED Technical Publishing Group 1992

ADVANCED DATA MINING

Heller

Data Management Review – Set 1995

DATABASE MINING: A PERFORMANCE PERSPECTIVE

Agrawal, Swami

IEEE – Transaction on Knowledge and Data Engineering – Dic 1993, Vol 5, Nro. 6

APLICACIONES INDUSTRIALES DE DATA MINING

Jorge Bocca y Richard Weber

XXI – Taller de Ingeniería de Sistemas

Santiago de Chile – 1998

WEB SITES

Redes Neuronales
http://www.gc.ssr.upm.es/inves/neural/ann2/anntutorial.htm

Aplicaciones convencionales de Redes Neuronales
http://hpus.u-aizu.ac.jp/hppd/hpux/NeuralNets/NeurDS-3.1/
ftp://genesis/bbb.caltech.edu/pub/genesis
ftp://ftp.cs.colorado.edupub/cs/misc/Mactivation-3.3.sea.hqx
http://www.dendronic.com/beta/htm
http://www.attrasoft.com
http://www.electronica.com/neural
http://ciberconta.unizar.es/leccion/redes/
http://www.gc.ssr.upm.es/inves/neural/ann2/author.htm

Aplicaciones Financieras de las Redes Neuronales
http://www.lania.mx/spanish/actividades/newsletters/1998-otono-invierno/tecnicas.html
http://www.ciberconta.unizar.es/Biblioteca/0004/SerGall96.html
http://www.emsl.pnl.gov:2080/proj/neuron/