Jorge Juan Plüss **Daniel Díaz***
Alicia G. Marchese **Ramiro Ingrassía***
Alicia M. Picco **Luciano Repetto***
Juan Carlos Scarabino

Instituto de Investigaciones y Asistencia Tecnológica en Administración, Escuela de Administración.

EL CONOCIMIENTO EN LOS SISTEMAS DE INFORMACIÓN

"...en unos cuantos decenios, la superficie de la tierra estará compartida por seres racionales, unos de los cuales habrán nacido de mujer y otros habrán sido manufacturados, por seres humanos o por otros individuos de su mismo género. ¿Afectará esta nueva perspectiva la autoimagen del hombre?" [i]

1. INTRODUCCIÓN

En este trabajo pretendemos resumir brevemente nuestras impresiones acerca del aporte que las máquinas pueden hacer al proceso de cambio que estamos viviendo. Esto se relaciona con el desarrollo futuro que prevemos para los sistemas de información y sus consecuencias en todo tipo de organizaciones humanas.

En el punto 2 se analiza el proceso de pensamiento en abstracto; a más de 30 años de los primeros intentos científicos tratando de hacer razonar a las máquinas y sabiendo que quizá este esfuerzo es el que les ha llevado a repensar muchas de las teorías psicológicas vigentes acerca de la creación de conocimiento.

Por último, hemos tomado de diversos autores sus experiencias y recopilaciones de distintos algoritmos desarrollados en el mundo, que tienden a dotar a las máquinas de la capacidad de razonar.

Si bien este tipo de herramientas, que requieren de la disposición de enormes cantidades de datos, exceden la posibilidad de una PyME, el presente trabajo constituye un análisis teórico que aportará nuevos elementos al prototipo de sistema experto ya desarrollado por nuestro equipo de investigación.

Entendemos que las herramientas descriptas a continuación se podrían aplicar naturalmente en organizaciones de tipo global. En nuestro caso, las reglas deberían ser enriquecidas con el aporte de la información proveniente de un cúmulo importante de datos globales para luego analizar su aplicación en el ámbito de las pequeñas y medianas organizaciones locales.

2. EL CONOCIMIENTO, SU ADQUISICIÓN Y APLICACIÓN

Los psicólogos conductistas, tomando como punto de partida el descubrimiento de los reflejos condicionados realizado por Pavlov a principios del siglo XX, analizan la conducta sin distinguir entre animales y seres humanos. "Su meta teórica es la predicción y el control de la conducta" [ii] y la conducta es definida como la respuesta de un organismo a los cambios que se producen en el medio, para lo cual se desconocen

* Colaboradores

los mecanismos que generan a la misma y se elimina toda explicación de los caminos invisibles que llevan a este efecto visible.

Quienes han tratado de emular el razonamiento humano utilizando computadoras, han encontrado que el problema es mucho más complejo y, hasta ahora no tienen una sola respuesta. Luego de haber logrado algunas aplicaciones exitosas basándose en reglas enunciadas por expertos, la comunidad científica ha notado que dichas reglas constituyen una estructura rígida e inaplicable a ningún otro dominio más que aquél para el cual fue diseñado. Se logró una compleja estructura de reglas, condicionadas por la consistencia lógica y las probabilidades (coeficientes de certeza), que sirven para resolver problemas sumamente acotados.

Luego, con el advenimiento de las redes neurales, los coneccionistas quisieron ver la posibilidad de crear sistemas que aprendan, a partir de un algoritmo lo más sencillo posible, que pudiera dar respuesta a problemas de diversos dominios. El inconveniente con que se encontraron, radica en que aún la estructura más simple puede fallar cuando se trata de llevar a símbolos conceptos que resultan elementales para un ser humano como, por ejemplo, determinar qué cosa debe presumirse como "similar"[iii]

En este sentido el aporte de Minsky resulta particularmente enriquecedor, ya que plantea que el camino a seguir no es único y que los sistemas del futuro deberán construirse con elementos provenientes de los dos razonamientos: simbólicos y coneccionistas (alineados y caóticos, basados en reglas y generadores de reglas):

"De esta manera, los sistemas de IA basados en el simbolismo están limitados para poder enfrentar excepciones a las reglas o lógica difusa, aproximaciones o a los fragmentos del conocimiento heurística. En reacción a esto, el movimiento conexionista, inicialmente, trato de diseñar sistemas más flexibles pero pronto llegaron a quedar aprisionados en su propia ideología, construir sistemas que aprendan dotados con la menor arquitectura estructural posible, esperando construir maquinas que puedan servir bien a todos lo expertos ecuánimemente, el problema con estas estructuras neutrales fueron que todavía encierran una suposición sobre que cosas son presumible de ser **similares"**[iv].

El campo de IA incluye muchas aspiraciones diferentes; algunos investigadores simplemente quieren máquinas que puedan hacer varios tipos de cosas que la gente llama "inteligentes", otros esperan entender qué características determinan que las personas puedan hacerlas. Aun más, otros investigadores buscan simplificar la programación y se preguntan; ¿por qué no construir una máquina, de una vez por todas, que pueda crecer y mejorarse aprendiendo de su experiencia?. ¿Por qué no podemos simplemente explicar lo que queremos, y después dejar que las máquinas hagan experimentos, o leer libros, o ir al colegio; las cosas que la gente hace?. Las máquinas actuales no hacen tales cosas: las redes de conexiones (neuronales) aprenden poco, aunque muestran algunos signos de convertirse en **inteligentes**, los sistemas simbólicos son **astutos** desde el principio, pero no han demostrado tener **sentido común**. Qué extraño que nuestro más avanzado sistema pueda competir con un experto humano pero sea incapaz de realizar cosas que hacen los niños. Minsky sugiere que esto proviene de la naturaleza de lo que llamamos especialización. Los especialistas generan modelos mediante reglas que realmente funcionan mientras que nos limitemos a ese campo específico. Pero cuando retornamos al mundo del sentido común, raras veces podemos encontrar reglas que puedan emplearse. En cambio, debemos reconocer cómo adaptar fragmentos del conocimiento a contextos y circunstancias particulares y debemos esperar necesitar muchas y diferentes clases de conocimiento en tanto que nuestras preocupaciones crezcan. Dentro de estos

campos “tontos”, encontramos mas y más excepciones y las ventajas iniciales de un contexto con reglas se convierten en fuertes limitaciones.

La investigación de IA debe evolucionar ahora de su foco sobre escenarios particulares, cuyo mejor exponente son los sistemas expertos. No hay una única forma de representar el conocimiento o de resolver problemas y las limitaciones de las actuales máquinas inteligentes, provienen de la búsqueda de teorías unificadas (simbólicos) o la búsqueda de reparar deficiencias de una teoría ordenada (conexionistas), pero conceptualmente se empobrecen al adoptar posiciones ideológicas. Las redes conexionistas numéricas puras son inherentemente deficientes en las habilidades de razonar bien; nuestros sistemas simbólicos lógicos son deficientes en las habilidades de poder representar lo más importante de la conexión heurística (lo incierto, aproximado y los enlaces que necesitamos para formular nuevas hipótesis). La versatilidad que necesitamos puede encontrarse sólo en arquitecturas complementarias que puedan explotar y manejar las ventajas de distintos tipos de representaciones al mismo tiempo. Para hacer esto, cada tipo formal de representación o inferencia de los conocimientos debe complementarse con una clase de máquina llamada *desaliñada (por ejemplo una red neuronal)* para que pueda incorporar la conexión heurística entre el conocimiento mismo y lo que esperamos hacer con el.

Esta discusión, que ha generado más dudas que aseveraciones nos lleva a hablar de un paradigma computacional, que por ahora es definitivamente superior al paradigma conductista como marco filosófico para interpretar la conducta humana. El conductismo es un heredero del empirismo inglés de Locke y Hume, que específicamente se caracteriza por insistir en que la única manera de conocer el comportamiento es por sus manifestaciones externas. La introspección cae en el descrédito: se generaliza la prohibición de hablar de procesos mentales; solo lo externo, lo observable, es materia de estudio científico.

En oposición a esto, el paradigma computacional reivindica los procesos internos, puesto que hay un estado interno en la máquina, además de haber entradas y salidas que corresponden a los estímulos y respuestas del conductismo. Los conductistas son ciegos a la etapa intermedia entre la entrada y la salida porque no tienen ninguna analogía empírica a qué apelar para esa etapa. Los computacionistas sí: tenemos un proceso interno en la máquina, si no directamente observable en el momento de su funcionamiento, empíricamente asegurado por resultar del proceso mismo de su construcción por los seres humanos.

Entre los más creativos exponentes del paradigma computacional debemos destacar a A. Newell y H. Simon. Ante todo, se les debe mencionar por su labor pionera de los años cincuenta, con la creación de un programa muy influyente, el Solucionador General de Problemas (*General Problem Solver*, mejor conocido como *GPS*); con él trataron de emular los métodos más generales de la inteligencia humana, lo que comúnmente identificamos como sentido común. Fracasarán en esa empresa, pero tal fracaso les dará ocasión para formular las primeras leyes del nuevo paradigma y, eventualmente, los llevará a formular su teoría fundamental. El *GPS* tiene tres objetivos. El primero es construir un modelo de la inteligencia humana; es un interés científico, de comprensión del fenómeno. El segundo es un interés tecnológico: crear herramientas inteligentes, donde lo importante es el provecho que se pueda obtener. El tercero, más bien filosófico, es descubrir en qué consiste la inteligencia en general, independiente de su incorporación en un ser humano o en cualquier otro organismo o mecanismo.

Resumiendo, hoy por hoy la situación está así: los métodos del pensamiento son o *métodos débiles*, pero aplicables a cualquier dominio (por ejemplo, el método de análisis de fines y medios), o *métodos fuertes*, pero aplicables solamente a cierto tipo

de problemas. De ahí han surgido dos ramas en la IA: la dedicada a simular la experticia de los expertos humanos, en base a la acumulación de reglas que representan conocimientos (herederas de las "tablas de diferencias" del *GPS*), y la dedicada a desentrañar el misterio del sentido común, una habilidad que aparentemente es completamente no especializada. Sin embargo, existe la grave sospecha, de que el sentido común como tal no existe: que lo que así llamamos no es otra cosa que una acumulación, en una memoria sumamente flexible y con poderes de recuperación de información excelentes, de experticias superficiales sobre miles de campos especializados, como relaciones humanas, física ingenua, y otras muchas dimensiones en que se desenvuelve el ser humano desde su infancia. Sabremos si eso es así cuando podamos replicar en un programa suficientemente versátil esa capacidad maravillosa del sentido común que pareciera no habernos costado nada, pero que por supuesto es resultado de una evolución milenaria y del aprendizaje espontáneo de muchos años de observación, ensayo y error, y reflexión por parte de los especímenes jóvenes del género *Homo*.

2.1. EMPIRISMO, SEMÁNTICA Y ONTOLOGÍA[v]

Si bien no es nuestra intención adentrarnos en cuestiones filosóficas, hemos notado que el problema del conocimiento es abordado a menudo con cierta ligereza, y creemos que es importante tener en cuenta que ciertas cuestiones, como la reproducción de procesos psicológicos por medio de máquinas, requieren de ciertos acuerdos acerca del lenguaje que se utiliza y sobre la significación que se da a los conceptos.

El lenguaje cotidiano tiene implicancias no dichas, que generan confusiones y malos entendidos, lo que obliga a todo aquél que desee desarrollar una labor científica a realizar el esfuerzo de realizar abstracciones con el fin de integrar eficientemente dichos elementos a los sistemas de información que se utilizan en las organizaciones.

El Problema De Las Entidades Abstractas

Tomando la definición de Carnap (1956), “el **empirismo** es el procedimiento o sistema basado únicamente en la práctica o rutina. Contra la especulación metafísica o idealista, aparece como doctrina filosófica. Es el eterno problema entre positivistas (realidad sensible) y los idealistas (la realidad fuera del alcance de los sentidos)”.

Según el autor, los empiristas son en general descreídos de todo tipo de entidades abstractas como propiedades, clases, relaciones, números, proposiciones, etc. pero no podremos obviarlas en ninguna de las disciplinas que nos propongamos, ya sean negocios, física o matemáticas. Partimos del hecho de que hay (existen) clases, números y proposiciones.

A modo de ejemplo tomamos elementos en el lenguaje cotidiano: el sistema ordenado de las cosas y eventos observables. Una vez que aceptamos el lenguaje en este contexto de cosas, podremos plantear y contestar preguntas como: ¿El monto de dinero disponible en caja es real? o ¿Es correcta la valuación de tal bien de uso? Los resultados de las observaciones son evaluados de acuerdo a ciertas reglas como la evidencia de confirmación o la no confirmación para las respuestas posibles. El concepto de realidad en esas preguntas internas es un concepto científico empírico. Reconocer algo como una cosa o evento real significa incorporarlo en el sistema de cosas en una particular posición de espacio y tiempo de forma que encajará junto con las otras cosas reales, de acuerdo a las reglas del entorno.

Debemos distinguir entre las preguntas y la realidad del mundo de las cosas. Esto no pertenece al campo del hombre de la calle ni de los científicos, sino al de los filósofos.

La decisión de aceptar el lenguaje de las cosas, en sí mismo no obsta para que usualmente esté influenciado por el conocimiento ya establecido, como alguna otra decisión concerniente a la aceptación de la lingüística u otras reglas. Los propósitos para los cuales el lenguaje se intenta usar, por ejemplo, el propósito de comunicación de conocimiento de los hechos, determinará qué factores son relevantes para esta decisión. Un ejemplo de sistemas que son de naturaleza más lógica que fáctica es el sistema numérico.

El sistema de proposiciones, en el cual nuevas variables son introducidas con el objeto de que una sentencia (declarativa) pueda ser substituida por una variable de este tipo; incluyen en adición a las sentencias del lenguaje original, también todas las sentencias generales con variables de todo tipo que pueden ser introducidas en el lenguaje.

La aceptación de un nuevo tipo de entidades es representada en el lenguaje por la introducción de un entorno de nuevas formas de expresión que serán usadas de acuerdo al nuevo conjunto de reglas. Habrá nuevos nombres para entidades particulares en cuestión; algunos de estos nombres pueden ya existir en el lenguaje antes de la introducción del nuevo entorno.

En sentido del análisis semántico ciertas expresiones del lenguaje son utilizadas para designar (o denotar o nombrar, o referirse a) ciertas entidades extra lingüísticas.

Para aquellos que desean desarrollar o utilizar métodos semánticos, la cuestión decisiva no es la cuestión ontológica de la existencia de entidades abstractas sino más bien la cuestión del uso de las formas del lenguaje abstracto o, en términos técnicos, el uso de variables a través de las cosas (o datos fenoménicos), es útil a los propósitos para los cuales el análisis semántico fue hecho: el análisis, la interpretación, la clarificación, o la construcción de lenguajes de comunicación, especialmente el lenguaje de las ciencias.

Esta cuestión no ha sido decidida ni aún discutida. No es una cuestión simple de sí o no, sino materia de gradación. Aún aquellos filósofos que han estudiado el análisis semántico y pensaron sobre herramientas utilizables para este trabajo, comenzando por Platón y Aristóteles y, en una forma más técnica en la base de la lógica moderna como Peirce y Frege, una gran mayoría acepta las entidades abstractas. Después de todo la semántica en el sentido técnico está todavía en la fase inicial de su desarrollo y debemos estar preparados para cambios posibles en los métodos.

No obstante, para poder acordar mínimamente en el desarrollo de un sistema, deberemos ser cautos en la formulación de aserciones y la crítica al examinarlos, pero tolerantes al permitir formas lingüísticas, y en muchos casos supondremos que existe un acuerdo sobre los significados, aunque el mismo no esté debidamente explicitado.

SEMÁNTICA: Ciencia que trata de la significación de las palabras. Trata de la relación entre los símbolos y lo que representan, así como de la forma de reaccionar el hombre ante los símbolos, de sus actitudes inconscientes, de las presunciones epistemológicas y lingüísticas y tiene por finalidad la sistematización del lenguaje científico y la unificación del conocimiento.

AGENTES: Son componentes que actúan autónomamente en función de objetivos dados. Forman parte de sistemas que proveen decisión, adaptación, asistencia personalizada, coordinación y soporte de negociación.

Qué Es La Ontología[1]

Respuesta corta: Es la especificación de una conceptualización

La palabra ontología parece generar muchas controversias en discusiones sobre IA. Tiene una larga historia en filosofía, en la cual se refiere al tema de la existencia. También suele confundírsela con epistemología, que se refiere al conocimiento y a lo conocido.

En el contexto del conocimiento, Gruber usa el término ontología para significar una especificación de una conceptualización. Esto es, una ontología es una descripción (como una especificación formal de un programa) de los conceptos y relaciones que pueden existir para un agente o una comunidad de agentes. Esta definición es consistente con el uso de ontología como conjunto-de-definiciones-conceptos, pero más general. Y es ciertamente un sentido diferente a la palabra utilizada en filosofía.

Lo importante es para qué es la ontología. Los especialistas han diseñado ontologías con el propósito de posibilitar la difusión (sharing) y reutilización del conocimiento. Pragmáticamente, una ontología es **un conjunto de definiciones del vocabulario formal**. Si bien esta no es la única forma de especificar una conceptualización, tiene algunas buenas propiedades para la difusión del conocimiento en el software de IA (por ej.: las dependencias semánticas entre el lector y el contexto). Prácticamente, **un compromiso ontológico es un acuerdo para usar un vocabulario (ej.: preguntas y afirmaciones) de una manera consistente (pero no completa) con respecto a la teoría especificada por una ontología.** Nosotros construímos agentes que se comprometen en ontologías. Nosotros diseñamos ontologías por lo tanto nosotros podemos compartir conocimientos con y entre esos agentes.

Ontologías Como Mecanismo De Especificación

El cuerpo del conocimiento formalmente representado en un sistema de información está basado en una conceptualización de objetos, conceptos, y otras entidades que son asumidas para existir en un área de interés y las relaciones que surgen de las mismas. Una conceptualización es una vista abstracta, simplificada del mundo que deseamos representar con algún propósito. Toda Base de conocimientos, sistema basado en conocimientos o agente de nivel de conocimientos se compromete con cierta conceptualización explícita o implícita.

Una ontología es una especificación explícita de una conceptualización. El término es tomado de la filosofía, donde la ontología es una cuenta sistemática de la existencia. **Para los sistemas de IA, lo que "existe" es aquello que puede ser representado**. Cuando el conocimiento de un dominio es representado en un formalismo declarativo, el conjunto de objetos que puede ser representado es llamado el universo del discurso. Este conjunto de objetos, y las relaciones descriptibles entre ellos, se reflejan en el vocabulario representacional con el que un programa basado en el conocimiento representa al conocimiento. En el contexto de la IA, podremos describir la ontología de un programa definiendo un conjunto de términos representacionales. En esta ontología, las definiciones asocian los nombres de entidades en el universo del discurso (ej.: clases, relaciones, funciones, u otros objetos) con un texto legible por humanos que describa lo que los nombres significan, y axiomas formales que limiten la interpretación y el correcto uso de esos términos. Formalmente, una ontología es un estadio de una teoría lógica.

---

[1] *GRUBER, Universidad de Stanford - http://ksl-web.stanford.edu/people/gruber/*

Nosotros usamos ontologías comunes para describir compromisos ontológicos de un conjunto de agentes, de forma que ellos puedan comunicar acerca de un dominio del discurso, sin operar necesariamente con una teoría globalmente compartida. Nosotros decimos que un agente se compromete con una ontología si sus acciones observables son consistentes con la definición en la ontología. La idea de compromiso ontológico está basado en la perspectiva del nivel del conocimiento. **El nivel del conocimiento es un nivel de descripción del conocimiento de un agente que es independiente del la representación a nivel simbólico utilizada internamente por el agente.** El conocimiento se atribuye a agentes observando sus acciones, un agente "sabe" algo si él actúa como si tuviera la información y está actuando racionalmente para lograr sus metas. Las "acciones" de los agentes, incluyendo servidores de base del conocimiento y sistemas basados en el conocimiento, pueden ser vistos a través una interface funcional relato y pregunta, donde un cliente interactúa con un agente haciendo aserciones lógicas (relatar), y cuestionando (preguntar)

Pragmáticamente, una ontología común define el vocabulario con el cual las preguntas y afirmaciones son intercambiadas entre agentes. Compromisos ontológicos son acuerdos para el uso de un vocabulario compartido en una manera coherente y consistente. Los agentes comparten el vocabulario necesario para compartir la base de conocimiento; cada uno conoce cosas que los otros no, y un agente que comparte una ontología no necesariamente debe contestar todas las preguntas que pueden ser formuladas en el vocabulario compartido.

Resumiendo, un compromiso hacia una ontología común es una garantía de consistencia, pero no completamente, con respecto a preguntas y afirmaciones utilizando el vocabulario definido en ontología.

3. DATAMINING

Partiendo del objetivo fundamental hacia el cual se orienta nuestra investigación, el procesamiento óptimo de toda la información disponible, surge la necesidad de integración de todos los datos de la organización, que algunos proveedores denominan ERP (Enterprise resource planning). Se trata de sistemas de gestión empresarial que recopilan todos los datos de la empresa, ya que integran todos los procesos de negocios, evitando la duplicación de registros y la transferencia de datos entre los sistemas. Esto facilita la conformación de nuevos modelos de negocios acordes con las demandas organizativas de hoy.

Esto no podría lograrse sin una adecuada planificación de los desarrollos, en forma concomitante con la evolución administrativa del ente. Es menester integrar el plan de negocios con la escalabilidad del sistema de información que permita incorporar nuevas filosofías, como por ejemplo, estructura cliente – servidor, reglas de negocio y estructuras de procesamiento de n - capas.

Aparece en este nuevo entorno el concepto de datawarehousing, que mejora los procesos al disponer de una base de datos independiente de los sistemas de gestión, cualquiera fueran sus fuentes de datos y desde las cuales los distintos usuarios pueden acceder a pedir información en el momento que la necesitan, sin afectar la performance normal del sistema.

Con esta estructura informática, cualquier empresa podrá desarrollar técnicas de DataMining, cuya idea subyacente es el descubrimiento de tendencias ocultas, que no podrían ser identificables por una persona, que existen dentro de grandes volúmenes de datos del datawarehouse.

DataMining es un proceso de inferencia de conocimiento a partir de una enorme cantidad de datos. Literalmente traducido significa “minería de datos”. Tiene tres componentes o etapas principales:

- Clasificación o clustering
- Reglas de asociación
- Análisis de Secuencia.

En **la clasificación**, analizamos un conjunto de datos y generamos un conjunto de reglas de agrupamiento para clasificar futuros conjuntos de datos. Clasificamos hechos y proveemos los síntomas que describen cada clase o subclase. Tiene mucho en común con el mecanismo estadístico y de aprendizaje de computadoras. El mayor problema radica en la etapa de aprendizaje de las reglas que agrupan los datos en clases predefinidas, fundamentalmente porque estamos tratando con millones de registros con un gran número de atributos involucrados, y los tiempos de ejecución se tornan prohibitivos.

Una **regla de asociación** implica cierta relación de asociación entre objetos de una base de datos, en distintos niveles de abstracción. Consiste, por ejemplo, en descubrir un conjunto de síntomas que generalmente se dan juntos en ciertas clases de hechos y luego estudiar las razones que subyacen a los mismos. Requiere de la búsqueda interactiva en grandes bases de datos de transacciones que es costosa de procesar.

El **análisis secuencial** consiste en la búsqueda con el fin de descubrir patrones que ocurren en secuencia de sucesos. Se relaciona con los momentos de ocurrencia de cada suceso.

3.1. ALGORITMOS DE CLASIFICACIÓN

Se trata de desarrollar una descripción o modelo para cada clase en una base de datos, basado en las características presentes en un conjunto de datos de entrenamiento etiquetados en clases.

MÉTODOS DE CLASIFICACIÓN DE DATOS.

Se trata de productos disponibles en el mercado que podrían ser elegidos como herramientas. Si bien son aceptado por algunos autores, mientras otros critican su utilización, entendemos que son aportes valiosos, a pesar de que aún se encuentran en etapa de maduración.

- Algoritmos estadísticos. Detectan patrones inusuales y patrones explicativos utilizando modelos estadísticos (modelos lineales).

- Redes Neurales. Emulan la capacidad del cerebro humano de encontrar patrones. Algunos desarrolladores han sugerido la aplicación de algoritmos de redes neurales para mapear patrones

- Algoritmos genéticos. Son técnicas de optimización que utilizan procesos como la combinación genética, mutación y selección natural en un diseño basado en conceptos de evolución natural.

- Método del vecino más cercano. Clasifica cada registro en un conjunto de datos basado en una combinación de las clases de los k registros más similares

- Inducción por reglas. Consiste en la extracción de reglas (if-then) basado en la significación estadística.

- Visualización de datos. Representación visual de relaciones complejas en datos multidimensionales.

Muchos algoritmos sugieren abstraer los datos de prueba antes de clasificarlos en varias clases. Esto se puede realizar de varias formas: Un conjunto de datos puede ser generalizado para niveles de abstracción mínimos, medios o altos. Si el nivel de abstracción es muy bajo, pueden resultar clases muy dispersas, árboles de clasificación demasiado tupidos y la consiguiente dificultad en concretar interpretaciones semánticas, mientras que un nivel muy alto puede resultar en la pérdida de precisión en la clasificación. El proceso de clasificación basado en generalizaciones de múltiples niveles ha sido implementado en DB_Minner system.

El aprendizaje mediante de reglas de clasificación involucra la búsqueda de reglas o árboles de decisión que particionan datos dados en clases predefinidas. Para el dominio de problemas realísticos, el conjunto de árboles de decisión posibles es muy grande para ser rastreado exhaustivamente. De hecho, la complejidad computacional de encontrar un árbol de decisión de clasificación óptima es un problema de difícil gestión.

3.1.1. ALGORITMOS ESTADÍSTICOS

Para el análisis de datos se requieren modelos. Dada una población finita de la cual proceden los datos, estos modelos deberán considerar la compleja estructura de la misma.

SOFTWARE PARA ANÁLISIS ESTADÍSTICO

Para llevar adelante el análisis estadístico de los datos relevados en una empresa es necesario seleccionar el software que permita a través de sus distintos módulos una mayor interpretación de los resultados. En primer término es imprescindible familiarizar a los usuarios en la estructura y manejo de los subprogramas de esta herramienta informática, con el objetivo de lograr la optimización de los recursos disponibles para el tratamiento de la información estadística.

En general, el análisis de las distintas experiencias, resume la información que contiene una muestra sobre la naturaleza de la población. Es decir, trabaja con un número limitado de casos.

La primera fase del análisis será el desarrollo (manejo) descriptivo de la información. La **estadística descriptiva** aborda el problema de sintetizar la información revelada por los datos, sin plantearse objetivos de naturaleza inductiva.

La extrapolación de resultados de la muestra a la población será el contenido de la **Inferencia Estadística**, cuyo objetivo es inferir conclusiones que se refieran a la población global, así como proporcionar medidas que permitan cuantificar el grado de confianza que podemos tener en tales conclusiones.

Como los posibles softwares para análisis estadístico parten de una descripción de la muestra, uno de los mayores inconvenientes es el de extrapolar los resultados a la población objeto de estudio. Para efectuar la extrapolación de los resultados observados a la población podemos recurrir a dos formas de actuar: la estimación y el contraste de hipótesis.

Se entiende por estimación de un parámetro, el cálculo de este valor a partir de la muestra.

El contraste de hipótesis se utiliza para decidir si cierta propiedad que suponemos posee la población es confirmada por la observación de la muestra.

La estimación y el contraste de hipótesis suponen dos formas complementarias para extrapolar los resultados observados en la muestra a la población. En muchos

casos el estudio estadístico tiene como única finalidad la comparación, aunque en general, el objetivo será mejorar la estimación o predicción del valor de la variable.

Decidir cual es la técnica estadística adecuada es el mayor inconveniente. Para nuestra experiencia podríamos considerar los siguientes métodos:

- Métodos relacionados con problemas de análisis de la varianza: Puede ubicarse en la estructura de un Modelo Lineal General, permiten analizar el efecto de los niveles de diversos factores (variables independientes) sobre una determinada característica.

- Métodos explicativos: Son todos aquellos en los que, como en la Regresión Lineal Múltiple , dentro del conjunto de variables observadas se distingue a una de ellas como dependiente. Su valor debe estimarse a partir de la información proporcionada por las restantes variables.

- Métodos descriptivos: No se distingue en este método entre variables dependientes e independientes, todas las variables se consideran al mismo nivel.

- Métodos para modelizar series temporales: Se puede considerar como un conjunto de observaciones de una variable, tomadas en intervalos regulares de tiempo.. Estos modelos explican la estructura y sirven para prever la evolución de la serie.

3.1.2. REDES NEURONALES

Hablar de redes neuronales implica entrar en un paradigma informático que es distinto al paradigma tradicional, ya que el computador ya no se programa paso a paso sino que se deja que el sistema aprenda solo. O sea que una red neural no es programada sino que debe enseñársele y con este proceso se logrará que resuelvan situaciones similares.

Como los sistemas expertos, no garantizan encontrar la mejor respuesta, sino una respuesta razonable.

El desarrollo de las redes neuronales se basa en tomar las características esenciales de la estructura neuronal del cerebro para crear sistemas que logren emularla, mimetizarla.

Pero la estructura del cerebro es diferente a la del computador. El cerebro humano no posee un único microprocesador, sino que está compuesto por miles de millones de neuronas, que realizan de modo impreciso y relativamente lento y con bastante imprecisión cálculos relativamente simples. Entonces este enfoque pretende sistemas compuestos por multitud de pequeños procesadores simples, que trabajan en paralelo, a los que se denomina neuronas artificiales.

O sea que básicamente la idea es tratar de hacer computadores que tengan la estructura física del cerebro humano.

Estas redes neuronales operan sobre la base de reconocimiento de patrones, y pueden captar, almacenar y utilizar conocimiento experimental, obtenido a partir de ejemplos. O sea que una de las principales características de las redes neuronales es el aprendizaje, ya que las mismas no se programan de forma directa, como por ejemplo si se realiza en los sistemas expertos, sino que se adquiere a partir de ejemplos, mediante un algoritmo de estimación, o de aprendizaje, para lograr encontrar los valores que corresponden a los parámetros del modelo. Estos valores reciben el nombre pesos sinápticos. En realidad estos algoritmos son modelos matemáticos multivariados que utilizan procedimientos iterativos hasta logra minimizar errores.

Las neuronas, cada una de las cuales realiza una función matemática, se agrupan en capas, constituyendo una red neuronal. Una determinada red neuronal está confeccionada y entrenada para llevar a cabo una labor específica. Si juntamos redes y les colocamos las interfases con el entorno obtenemos un sistema global.

¿Cómo aprenden?

Como dijimos, aprenden mediante ejemplos y el aprendizaje puede ser de dos tipos. Vemos cada uno:

- Aprendizaje supervisado:

Tiene la particularidad de parecerse al método tradicional de enseñanza, con un profesor que indica y corrige los errores del alumno hasta que éste aprende la lección. Si la red neuronal dispone de un tipo de aprendizaje supervisado debemos proporcionarle parejas de patrones entrada-salida y la red aprende a asociarlos. Es el equivalente estadístico de los modelos en los que hay vectores de variables independientes y dependientes: como ser técnicas de regresión, modelos de series temporales, etc.

- Aprendizaje no supervisado.

Aquí no existe un profesor que corrija los errores al alumno o sea que estaríamos en presencia de un autoaprendizaje, donde el alumno dispone del material de estudio pero nadie lo controla. En este caso solamente debemos suministrar a la red los datos de entrada para que extraiga los rasgos característicos esenciales. Es el equivalente estadístico de los modelos en los que sólo hay vectores de variables independientes y buscan el agrupamiento de los patrones de entrada: análisis de conglomerados, etc.

Características fundamentales de una red neuronal informática:

1. La cantidad de neuronas artificiales que la componen, o sea los elementos de procesamiento en sí.

2. Los grupos de neuronas que cumplen la misma función o sea la cantidad de niveles o capas que componen la red.

3. La función de transferencia entre las neuronas

4. Los pesos sinápticos, que son los valores de los coeficientes adaptativos capaces de cambiar con el aprendizaje. La denominación se debe a su similitud con la sinapsis entre las células nerviosas.

Funcionamiento:

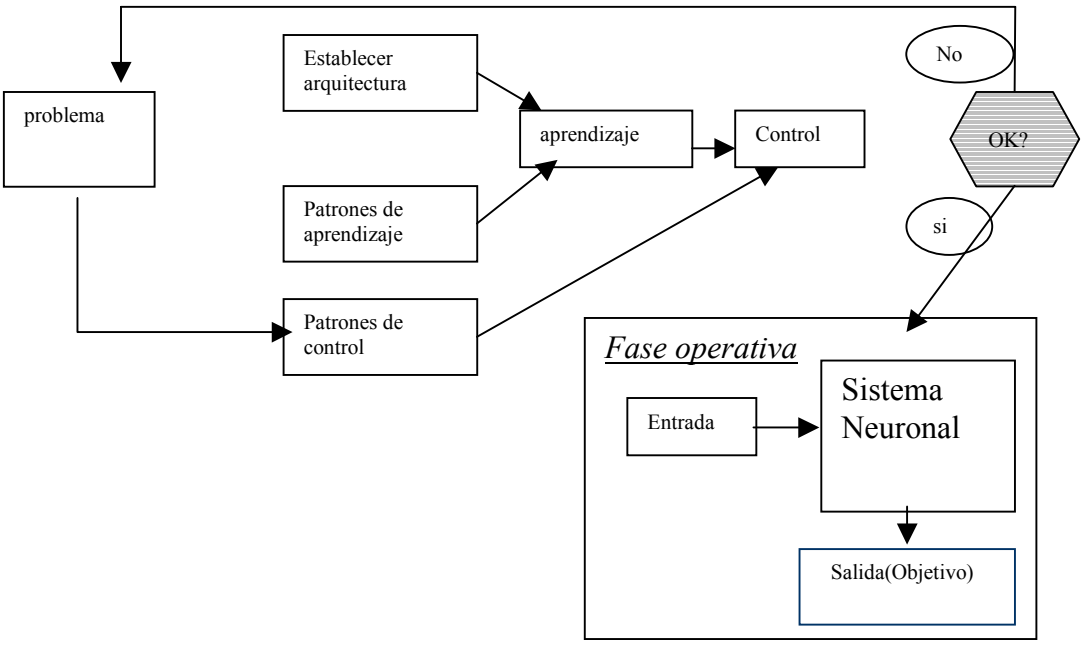
Cada neurona cuenta con numerosas entradas pero con una única salida. Esta salida se ramifica en varias direcciones para servir como entrada a otras neuronas.

De ahí que se conforme una verdadera red, dado que una neurona es afectada por el comportamiento de sus vecinas y a su vez afecta a las mismas, lo que implica que los pasos que se siguen tienen en cuenta el estado total de la red.

Los pesos sinápticos, son los coeficientes que amplifican o atenúan la señal de entradas. Si los valores de las entradas ponderadas por los mismos supera un determinado escalón, la neurona se activa, caso contrario permanece inactiva.

Emulación por Software:

Mediante software se puede emular el comportamiento de las redes neuronales en un computador convencional. De hecho existe una gran cantidad de programas de redes neuronales que funcionan incluso en un ordenador personal.



Modo de trabajo con redes neuronales.

Comparación con los sistemas expertos:

Si bien las redes neuronales se asemejan a los sistemas expertos en cuanto forman parte de la filosofía de la Inteligencia Artificial y tiene como objetivo el representar el conocimiento, son opuestos en cuanto a la forma de conseguir los mismos.

Los sistemas expertos pretenden emular la lógica del razonamiento humano, mientras s que las redes neuronales está orientadas hacia la parte física. Por otra parte los sistemas expertos tienen en sí un enfoque de tipo deductivo, ya que pretenden obtener reglas, en tanto las redes neuronales poseen un razonamiento de tipo inductivo, ya que se basa en aprendizaje mediante ejemplos. Como en nuestra vida diaria utilizamos ambos enfoques de razonamiento, ambas filosofías son aceptadas y ambos modelos son perfectamente compatibles, de forma que pueden incluso integrarse en un único sistema.

Como trabajan entonces las redes neuronales:

En un principio una red neuronal no posee ningún tipo de conocimiento almacenado. Entonces es preciso entrenarla para que ejecute una tarea. Debemos seleccionar los parámetros y el algoritmo.

Como en realidad se trata de un procedimiento estadístico, podemos decir que es necesario estimar los parámetros. Debe entonces seleccionarse el conjunto de datos a utilizar o sea los patrones de aprendizaje.

Una vez que obtenidos éstos, debe establecer la estructura neuronal, el número de neuronas en cada capa y las conexiones. O sea, definimos el tipo de red.

Estadísticamente hablando, se selecciona el modelo y el número de variables dependientes e independientes.

Luego vienen: El aprendizaje: Durante el mismo se produce una variación de los pesos sinápticos, o sea de los coeficientes que miden la intensidad de interacción entre las neuronas.

También en esta fase se puede producir la incorporación de nuevas neuronas o la pérdida de algunas de ellas. El programa comienza a iterar, por lo general mostrando gráficamente o tabularmente los resultados del aprendizaje para cada iteración.

La fase de test: Por medio de nuevos patrones de entrada, se comprueba la eficacia del sistema que se ha generado. Si no es aceptable, debe procederse a repetir la fase de desarrollo, ya sea utilizando un nuevo conjunto de patrones de entrenamiento, o modificando el sistema de aprendizaje o la arquitectura.

Superada esta fase, la arquitectura, neuronas y conexiones, y los pesos sinápticos quedan fijos pudiendo el sistema operar en modo recuerdo. El modo recuerdo es el modo de operación normal del sistema: dada una entrada proporcionará una salida concordante con el aprendizaje recibido.

La validación de los resultados: Aquí es conveniente destacar que es posible conocer el nivel de excitación de cada neurona, ya que puede obtenerse un gráfico o una tabla con los valores de los pesos sinápticos, que pueden ser usados con fines estadísticos.

3.1.3. ALGORITMOS GENÉTICOS

Son algoritmos de optimización global basados en los mecanismos de la selección natural utilizados por la genética.

Breve reseña histórica.

Hasta mediados del siglo XIX los naturistas creian que cada especie había sido creada por un ser supremo o habían surgido por generación espontanea.

Algunos trabajos como los de Carolus Linnaeus sobre la clasifcación biológica de los organismos, Jean Baptiste Lamark sobre el uso y desuso de órganos y Thomas Malthus sobre la incidencia de los factores ambientales, falta de alimentos y otros condicionantes que limitan el crecimiento de las poblaciones, fueron marcando el rumbo hacia la idea de la “evolución natural” y la necesidad de estudiar los procedimientos en que se basaba.

Esta idea fue tomando forma hasta que en 1858 Charles Darwin presentó su obra “On the Origin of Species by Means of Natural Selection”, en la cual enunció los principios que rigen la evolución natural de las especies.

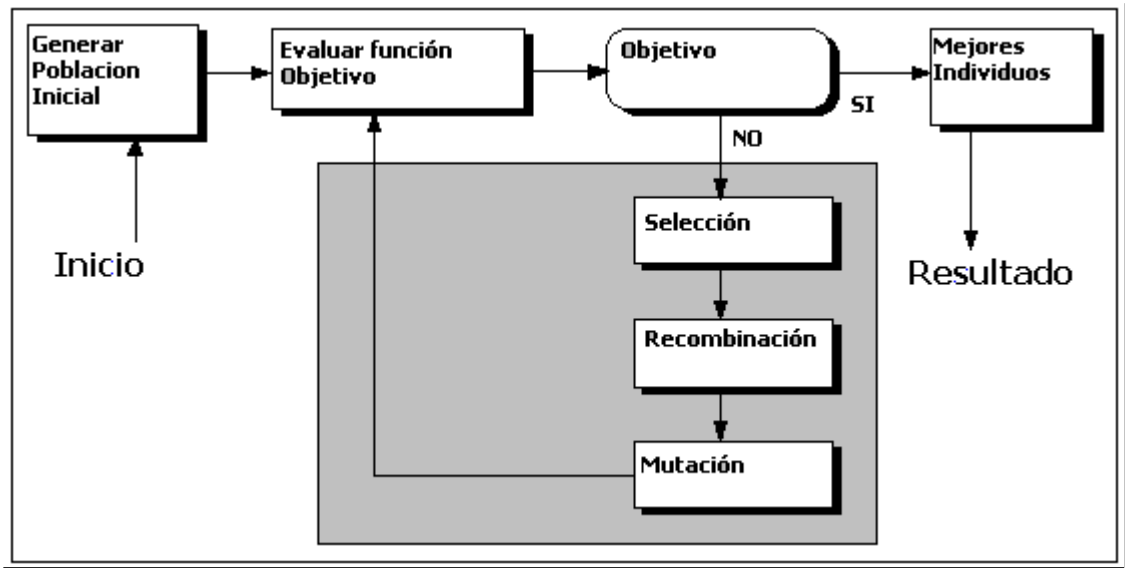
Ya en la década del setenta John Holland comenzó a desarrollar las primeras ideas sobre la técnica de búsquedas y optimización a través de Algoritmos genéticos, finalizando en 1975 con su libro “Adaptación en sistemas naturales y artificiales” considerada la Biblia de los Algoritmos Genéticos.

Los elementos y la técnica:

Los pasos básicos en que se basa un algorítmo génetico son:

Procedimiento	Explicación
Generación de una población inicial	Se determina en forma aleatoria un conjunto de individuos que conformarán la población. El numeró de individuos seleccionado será determinante de la perfomance del algoritmo.
Selección	Se seleccionan los individuos que sufrirán alteración o que intervendrán en la reproducción
Apareamiento	Reproducción de individuos
Mutación	Cambios en las características cualitativas de los individuos
Evaluación	Es el nudo central del algoritmo. Mediante esta función se determina cuales son los individuos que más se acercan al optimo y por lo tanto los que sobrevivirán

El algorítmo genetico Básico (BGA) se muestra en el siguiente diagrama



Parámetros del algoritmo genético

1. **Tamaño de la población:** Son el número de individuos que componen la muestra
2. **Tasa de cruzamiento:** Cuanto mayor sea la tasa más rápidamente los cambios se impondrán en la población.
3. **Tasa de Mutación:** Determina la cantidad de individuos que serán sometidos a este proceso.
4. **Intervalo de generación:** Determina el porcentaje de la población que será sustituida en la próxima generación.

Aplicaciones prácticas de los GA:

Entre otras funciones, los algoritmos genéticos se utilizan para:

- Control de Sistemas Dinámicos
- Generación y optimización de Bases de reglas
- Desarrollo de nuevas topologías de conexión:
  - Ingeniería de Redes Neuronales
  - Modelado de Estructuras Biológicas Neuronales
- Simulación de modelos Biológicos
- Composición musical

Aplicaciones en Administración:

Por sus características de ser algoritmos de búsqueda y optimización de funciones, creemos que se pueden dar múltiples aplicaciones de los GA. Entre ellas para resolver decisiones de composición óptima de capitales en empresas

3.2. ALGORITMOS DE REGLAS DE ASOCIACIÓN

Una regla de asociación es una regla que implica ciertas relaciones de asociación entre un conjunto de objetos, como "ocurrencia conjunta" o "uno implica el otro" en una base de datos. Dado un conjunto de transacciones, donde cada transacción es un conjunto de literales (llamados ítems), una regla de asociación es una expresión de la forma X, Y donde X e Y son conjuntos de ítems. El significado intuitivo de tal regla es que las transacciones de la base que contienen X tienden a contener Y. Un ejemplo de regla de asociación es "el 30% de las transacciones que contienen cerveza también contienen pañales, el 2% de todas las transacciones contienen ambos ítems". Aquí el 30% se denomina la confianza de la regla, y el 2% es el soporte de la regla. El problema es encontrar todas las reglas de asociación que satisfagan el mínimo de soporte y la mínima confianza requerida por el usuario.

El grupo "IBM´s Quest project team" desarrolló Apriori, un algoritmo de reglas de asociación para grandes bases de datos de transacciones . En este entorno, un "itemset" es un conjunto no vacío de ítems. Ellos descompusieron el problema de las reglas de asociación de minería en dos partes:

- Buscar todas las combinaciones de ítems que tienen soporte de transacciones por encima de un mínimo establecido. A éstos se le denomina itemsets frecuentes.

- Utilizar los itemsets frecuentes para generar las reglas deseadas. La idea general es que si dados ABCD y AB son itemsets frecuentes, podremos determinar si la regla AB CD contiene el ratio r = soporte(ABCD) / soporte(AB). La regla se soporta sólo si r >= que el mínimo de confianza. La regla tendrá un soporte mínimo porque ABCD es frecuente.

El algoritmo realiza múltiples pasadas sobre la base de datos. En el primer paso, simplemente cuenta las ocurrencias para determinar los itemsets con 1 item. Los pasos subsiguientes (k) consisten en dos fases. Primero, la frecuencia de itemsets L_{k-1} (el conjunto de todas las frecuencias de los k-1 itemsets) encontrados en el paso número k-1 que son usados para generar el itemset candidato C_k, utilizando el algoritmo anterior.

Algoritmos distribuidos / paralelos

Bases de Datos o Datawarehouses pueden guardar una enorme cantidad de datos para ser explotados. Las reglas de asociación de minería en tales bases de datos pueden requerir un sustancial poder de procesamiento. Una posible solución a este problema puede ser un sistema distribuido. Más aún, muchas grandes bases de datos son distribuidas y en función de las mismas se hace más necesario utilizar algoritmos distribuidos.

El mayor costo de asociación en reglas de minería es el procesamiento de ese conjunto de grandes conjuntos en la base de datos. La computación distribuida de grandes conjuntos implica nuevos problemas. Se puede computarizar localmente con facilidad, pero esto no implica el procesamiento global. Como es muy caro transmitir todos los datos a otro sitio, una opción es transmitir todas las cuentas de todos los conjuntos de datos, no importa el tamaño, a otros sitios. De cualquier forma, una base de datos puede contener enormes combinaciones de conjuntos de datos, y va a involucrar el traspaso de un gran número de mensajes.

Un algoritmo de datamining es FDM (Fast distributed mining of association rules) desarrolla las siguientes actividades:

1. La generación de conjuntos candidatos tiene el mismo espíritu que Apriori. Algunas relaciones entre grandes conjuntos globales y locales son exploradas para generar un conjunto más pequeño en cada iteración y por lo tanto reduce el número de mensajes para ser pasados.

2. Después de generar el conjunto de candidatos, dos técnicas de depuración, local y global, son desarrolladas para eliminar algunos conjuntos candidatos en cada sitio individual.

3. Para determinar si el conjunto candidato es grande, este algoritmo requiere sólo O(n) mensajes para soportar el intercambio de cuentas, donde n es el número de sitios en la red. Es mucho menos que el algoritmo de adaptación de A priori, que requiere $O(n^2)$ mensajes.

3.3. ANALISIS SECUENCIAL. Patrones secuenciales.

Los datos de entrada son un conjunto de secuencias, llamado secuencia de datos. Cada secuencia de datos es una lista ordenada de transacciones (o conjuntos de datos), donde cada transacción es un conjunto de ítems (literales). Típicamente, existe un tiempo de transacción asociado con cada transacción. Un patrón secuencial también consiste en una lista de conjuntos de ítems. El problema es encontrar todos los patrones secuenciales con un mínimo soporte especificado por el usuario, donde el soporte de un patrón secuencial es el porcentaje de secuencia de datos que contiene el patrón.

Un ejemplo de ese patrón es que los clientes de videoclubs típicamente alquilan “La guerra de las galaxias” y luego “El Imperio Contraataca” y luego “El regreso del Jedi”. Nótese que esos alquileres no necesariamente son consecutivos. Los clientes que alquilan algún otro video intermedio, también contienen ese patrón secuencial. Por otro lado, los elementos de un patrón secuencial no necesariamente son ítems simples. Este problema fue inicialmente motivado por aplicaciones en la industria minorista, incluyendo mailing, ventas add-on, o satisfacción al cliente. Los resultados se aplican a dominios científicos y de negocios y, por ejemplo, también al área médica, en la cual la secuencia de datos puede corresponder a los síntomas o muertes de un paciente, con una transacción correspondiente a los síntomas mostrados o muertes diagnosticadas durante la visita al médico. Los patrones descubiertos usando estos datos pueden ser usados en investigaciones sobre defunciones para ayudar a identificar síntomas/muertes que preceden ciertas muertes.

Algoritmos de búsqueda de patrones secuenciales

Varios grupos de trabajo en este campo proponen algoritmos para detección de patrones secuenciales. Dos de ellos propuestos por el IBM´s Quest data team.

El problema de minería de patrones secuenciales puede ser dividido en las siguientes fases:

- Fase de ordenamiento. Implícitamente convierte la base de datos de transacciones original en una base de secuencias.

- Fase de "litemset" en la cual encontramos el conjunto de todos los conjuntos de ítems L. Simultáneamente estamos buscando el conjunto de todos las secuencias l, desde que ese conjunto es justamente l.

- Fase de Transformación. Necesitamos determinar repetidamente cuál de un conjunto dado de secuencias están contenidas en una secuencia de clientes. Para hacer esta prueba rápido, transformamos cada secuencia de cliente en una representación alternativa. En una transformación de secuencia de cliente, cada transacción es reemplazada por el conjunto de todos los litemset contenidos en esa transacción. Si una transacción no contiene ningún litemset, la misma no se retiene en la secuencia transformada, se saca de la base de datos transformada. De cualquier forma, la misma aún contribuye a la cuenta del número total de clientes. La secuencia de clientes queda ahora representada por la lista de los conjuntos de los litemsets.

- Fase de secuencia. Usa el conjunto de los litemsets para encontrar las secuencias elegidas.

- Fase máxima. Encuentra las secuencias máximas a través el conjunto grande de secuencias. En algunos algoritmos esta fase se combina con la fase de secuencia para reducir el tiempo invertido en el conteo de secuencias no máximas.

La estructura general de los algoritmos para la fase de secuencia es que ellos pueden hacer múltiples pasadas sobre los datos. En cada pasada, empezamos con un conjunto inicial de secuencias. Usamos el conjunto inicial para generar nuevas secuencias potenciales, llamadas secuencias candidatas. Encontramos el soporte para esas secuencias candidatas durante el paso sobre los datos. Al final de la pasada, determinamos cuál de las secuencias candidatas son actualmente generales. Estas candidatas generales se convierten en la semilla para el próximo paso, todas las l-secuencias con soporte mínimo, obtenidas en la fase litemset, conforman el conjunto inicial.

Existen dos familias de algoritmos: contar-todo y contar-algunos. El algoritmo contar-todo cuenta todas las secuencias generales, incluyendo las no-máximas. Las no-máximas deben entonces ser podadas (en la fase máxima). AprioriA11 es un algoritmo de contar-todo, basado en el algoritmo Apriori para encontrar itemsets generales presentados antes. A priori-Some es un algoritmo de contar-algunos. La intuición que subyace a estos algoritmos es que desde que estamos interesados en secuencias máximas, podemos evitar la cuenta de secuencias que estén contenidas en secuencias mayores, siempre y cuando primero contemos las secuencias mayores. De cualquier forma, debemos ser cuidadosos y no contar demasiadas secuencias grandes que no tengan soporte mínimo. De otra forma, el tiempo que ahorremos no contando algunas secuencias contenidas en otras mayores puede ser menor al tiempo invertido contando secuencias que no tengan soporte mínimo que nunca debieron ser contadas debido a que sus subsecuentes no fueron grandes.

4. CONCLUSION

Luego de estudiar una vasta cantidad de papeles de desarrollo, algunas de las conclusiones son las siguientes:

- Comparación de algoritmos:

Ninguno de los algoritmos puede reemplazar a los otros cuando la medida de la performance tiende a una ocurrencia generalizada. Este resultado, generalmente

denominado “No free luncj theorem or conservatios law” (Wolpert 1994, Schaffer 1994) asume que todos los posibles objetivos son igualmente deseados.

Promediando la performance de un algoritmo a través de todos los objetivos, asumiendo que son igualmente deseados, sería como promediar la performance de un automóvil en todos los tipos de terreno, asumiendo que todos son igualmente deseados. Esta asunción es claramente incorrecta en la práctica; dados determinados dominios, es claro que no todos los conceptos son igualmente probables.

En el ámbito de la medicina, muchas medidas (atributos) que los médicos han desarrollado a lo largo de los años tienden a ser independientes. Si los atributos están altamente correlacionados, sólo uno será elegido. En tales dominios una cierta clase de algoritmo de aprendizaje puede superar a los otros. Por ejemplo, Naive-Bayes parece tener una buena performance en dominios médicos. Quinlan identifica familias de dominios paralelos y secuenciales y reclama que las redes neurales son buenas para desarrollos en dominios paralelos, mientras que los algoritmos de árboles de decisión son mejores para desarrollos en dominios secuenciales. Por lo tanto, un algoritmo de inducción simple no puede considerarse el mejor clasificador en todas las situaciones.

Este campo se encuentra aún en su infancia y está en constante evolución. Las primeras personas que pensaron seriamente el problema de datamining fueron aquellos que desarrollaron el campo de las bases de datos, ellos fueron los primeros que enfrentaron el problema. Muchas de las herramientas y técnicas utilizadas en Datamining provienen de otros campos relacionados como reconocimiento de patrones o teorías de estadística y complejidad. Sólo recientemente los desarrolladores de estos ámbitos han estado interactuando para resolver el uso de minería de datos.

Muchos de los intentos de datamining fallaron por el vasto **tamaño de los datos** . Las nuevas técnicas para acumularlos se encuentran aún en proceso de desarrollo. Por otra parte, todos los algoritmos que se proponen para minería de datos deberán actuar por fuera de las estructuras centrales de datos, para no alterar la actividad normal de la organización. Muchos de los algoritmos existentes no están orientados en este sentido. Algunos de los propuestos últimamente, como algoritmos paralelos están comenzando a ver este aspecto, que será de gran importancia para los procesos de toma de decisiones estratégicas.

Debemos mencionar además que las nuevas Bases de datos (sistemas de gestión integral de datos) surgidas en el mercado, contemplan el desarrollo de técnicas de datamining, si bien aún rudimentarias.

Muchos de los algoritmos asumen que los datos están **libres de ruidos,** como resultado, el tiempo más largo de la solución de problemas se vuelve el procesamiento de dato. La conformación de datos y el gerenciamiento de la experimentación y resultados son frecuentemente los que más tiempo y frustraciones conllevan.

El concepto de datos ruidosos puede ser comprendido por el ejemplo de mining logs. Un escenario real puede ser la situación en que uno desea buscar información a partir de los ingresos (logs) en la web, como por ejemplo cuáles han sido las páginas más visitadas. Pero un usuario puede acceder a un web site por error en la URL o haber apretado incorrectamente un botón. En tal caso, esa información le resultará inútil en el caso de que estemos tratando de deducir una secuencia en la cual el usuario accede a páginas web. Los logs pueden contener tal enorme cantidad de items de datos que constituyen ruidos . Una base de datos cualquiera puede constituirse con el 30 o 40% de datos ruidosos y pre - procesar estos datos puede llevar más tiempo que el de ejecución de un algoritmo.

Más allá de los problemas que tengamos para la aplicación de estos algoritmos, es innegable que los mismos ya se encuentran entre nosotros y tarde o temprano serán de uso generalizado. Esto es así no sólo por la gran utilidad que prestan, sino también por envergadura de las decisiones que deben ser tomadas, en cuanto a velocidad de respuesta y en cuanto a cantidad de información necesaria para las mismas.

El aporte de esta nueva concepción del conocimiento, en cuya generación intervienen hombres y máquinas, es innegable a nivel científico ¿Podremos ignorarlo cuando pensamos a las organizaciones?

---

[i] GUTIERREZ, Claudio, "Psicología de las computadoras", Costa Rica 1999

[ii] BRAUNSTEIN, Néstor, "Psicología, Ideología y Ciencia", Ed. Siglo XXI, Buenos Aires, 1975, pág.37

[iii] WINSTON, Patrick, "Logical vs. Analogical or Symbolic vs. Connectionist or Neat vs. Scruffy", in Artificial Intelligence at MIT., Expanding Frontiers, (Ed.), Vol 1, MIT Press, 1990. Reprinted in AI Magazine, 1991

[iv] MINSKY, Marvin, "Logical vs. Analogical or Simbolic vs. Connectionist or Neat vs. Scruffy", http://www.ai.mit.edu/people/minsky/minsky.html

[v] RUDOLF CARNAP ***Revue Internationale de Philosophie*** **4 (1950): 20-40. Reprinted in the Supplement to *Meaning and Necessity: A Study in Semantics and Modal Logic,* enlarged edition (University of Chicago Press, 1956).**