Lenguaje de Consulta Basado en Conocimiento para la selección de Técnicas Estadísticas y de Data Mining

Héctor Oscar Nigro, Daniel Xodo, José Francisco Zelasco
INCA/INTIA – Departamento de Computación y Sistemas
Facultad de Ciencias Exactas – Unicen – Tandil
Campus Universitario – Paraje Arroyo Seco s/n
TE +54-2293-432466 – FAX: +54-2293-444431
Email:[onigro;dxodo]@exa.unicen.edu.ar

Resumen

Este artículo describe una aplicación de un Sistema Estadístico y Data Mining, en el marco del proyecto de investigación MERAIS (Métodos de Razonamiento Aproximado para la Investigación Socioeconómica) desarrollado en el Area Informática de Gestión perteneciente al grupo INTIA de la Facultad de Ciencias Exactas de la Unicen. Esta aplicación consiste en diversos módulos con los que se pretende avanzar en la construcción de un Sistema de Análisis de Datos basado en Conocimiento. Es una aproximación a los procesos continuos, cuyo objetivo es mejorar los programas de análisis de datos. No sólo desde el punto de vista de su utilización, sino desde el de la selección de la técnica estadística y data mining, más apropiada a ciertos tipos de datos.[8]

1) INTRODUCCION

Data Mining y Estadísticas tienen diferentes tradiciones intelectuales. Ambas atacan problemas de colección y análisis de datos. Data Mining tiene orígenes recientes. Se encuentra en la tradición de la Inteligencia Artificial, Aprendizaje automatizado, Sistemas de Información Gerencial y Metodología de Bases de Datos. Trabaja generalmente con conjuntos grandes de datos. La Estadística tiene una tradición más larga, favorecida por modelos probabilísticos y acostumbrados a trabajar con conjuntos de datos relativamente pequeños. Ambas tradiciones utilizan herramientas informáticas, pero a menudo son diferentes [4].

Un Sistema de Análisis de Datos basado en Conocimiento se debe comportar con el usuario, de una forma parecida a como lo haría un consultor estadístico y data mining. Intenta guiar al analista hacia modelos adecuados a sus datos y necesidades, verificando si los datos se adaptan a las hipótesis de los modelos a emplear, realizando análisis primarios, gráficos, etc., antes de aplicar los modelos complejos [9].

El Sistema de Análisis de Datos basado en Conocimiento propuesto es un sistema abierto, relacionado y universal [7]; a) abierto: supone que se trata de un sistema con bases de conocimiento potencialmente ilimitadas, que se irán nutriendo en el tiempo de conocimiento experto existente o que se vaya desarrollando; b) relacionado: advierte la interrelación entre conocimiento experto de tipo estadístico con conocimiento experto de tipo computacional sobre los distintos lenguajes de consulta, y c) universal: capaz de incluir cualquier tipo de técnica estadística y de Data Mining, y poderla aplicar sobre unos datos en cualquier tipo de lenguaje estadístico y de consulta de Data Mining. Podríamos decir que esta aplicación aspira integrar ambas tradiciones.

El sistema comporta tres bloques. En el punto siguiente se describen aspectos del Núcleo del Sistema. En el posterior el módulo interactivo. A continuación las áreas temáticas vinculadas al proyecto. Finalmente se presentan las conclusiones.

2) MODULO NÚCLEO

2.1. Base de Conocimiento Estadístico y Data Mining (B.C.E.DM)

La base de conocimiento estadístico y de Data Mining es una biblioteca de operaciones de análisis de datos, donde se encuentran las técnicas de análisis con sus descripciones y condiciones

asociadas. Organizada como una jerarquía en archivos que constituirán el conocimiento en sí mismo, que desciende desde análisis generales (análisis descriptivo, análisis causal, etc.) hasta técnicas particulares, que conforman modalidades de las anteriores (descripción detallada de datos, análisis gráfico, contrastes de dos muestras, etc.). Conformado por un analizador semántico de condiciones sobre los distintos tipos de análisis incluidos en la base sirven de apoyo a esta base de conocimiento [7].
Dentro de los atributos que posee la B.C.E.DM. se encuentran los parámetros de evaluación.. El dominio de dichos parámetros puede responder a diferentes paradigmas, como también al nivel de medición de las variables involucradas, tipo de conocimiento que se quiere analizar, y cantidad de variables involucradas. Por ejemplo, en caso de técnicas univariadas, interesa conocer de la distribución de la variable [medidas de tendencia central, Dispersión, Frecuencias]. También pueden plantearse interrogantes como: si se quiere tratar ciertos casos de una manera diferente de otros [Sí, No], cual es la forma de la distribución [Simetría, Sesgo], etc. [11]

2.2.Base de Lenguaje Estadísticos y de consulta de Data Mining (B.L.E.DM)

La base de conocimiento para los lenguajes estadísticos y de consulta de Data Mining (B.L.E.DM) es una base de conocimiento experto que almacena información sobre los distintos lenguajes estadísticos, de consulta de data mining, y subrutinas propias que completen aquellas. Puede constituirse en un lenguaje propio generador de lenguajes estadísticos y de control, convirtiéndose de esta forma en una biblioteca de rutinas de programación dentro de un metalenguaje de creación propia. El objetivo final será disponer del suficiente conocimiento como para poder crear macroinstrucciones que den respuesta a los distintos problemas de Análisis de Datos planteados.
Se trata de aprovechar el conocimiento estadístico de la B.C.E.DM configurando la B.L.E.DM como una base relacional, asociando a cada modalidad de un análisis una o varias macroinstrucciones en uno o varios lenguajes, que serán las que constituirán, junto con las subrutinas propias, el conocimiento de esta base. Se intenta crear un metalenguaje semántico, configurado con las generalidades de los lenguajes estadísticos, que permita construir macroinstrucciones en todos los lenguajes de soporte. [7]

2.3.Diseño de formularios configuración del instrumento de medición, meta-dato de la matriz de datos.

Toda vez que el paso previo para poder realizar análisis de datos es la determinación de las consideraciones oportunas sobre las variables que se van a analizar, completamos el sistema integrado con un módulo de diseño de formularios. Existen distintas formas de proceder a la hora de la captura de los datos: 1. La introducción de las respuestas a los cuestionarios que hayan servido de soporte para entrevistas. 2. La captura directa interactiva (encuestas telefónicas, etc.), inteligente. 3. La lectura directa sobre ficheros externos, procedentes de otros programas y formatos, igualmente con validación automática inteligente. [7]
En esta parte figuran las definiciones de una lista de dimensiones, con las que se describe un objeto de estudio. Se detalla una lista con las principales características de cada variable, nivel de medición, raíz a la que pertenece, valores de esta variable, gráfico o tabla de representación, nombre de la variable y etiqueta. Dentro de la configuración de un instrumento de medición, se encuentran las reglas de medición, estructuradas en una taxonomía.

3) MODULO INTERACTIVO

3.1. Manejo de Análisis de Datos Experto (M.A.D.E)

Un modelo de datos estadísticos es conveniente a los efectos de representar la complejidad de las estructuras de datos estadísticos y para procesar operadores de alto nivel. Esto se consigue describiendo un mecanismo basado en conocimiento para deducir operaciones elementales que se encuentran debajo de un operador de alto nivel. [1]

Es deseable un lenguaje de alto nivel para que el usuario pueda especificar solamente los datos de comienzo y la estructura lógica del resultado. Tal interacción conceptual requiere un mapeo automático, transformando un operador de alto nivel en varias operaciones elementales.
El Manejo de Análisis de Datos Experto (M.A.D.E.) es un sistema basado en conocimiento capaz de validar una expresión de consulta y verificar la viabilidad de la misma por medio de varios tipos de conocimiento, a saber: conocimiento de análisis de datos, en términos de criterios de aplicabilidad de operadores; conocimiento del dominio; conocimiento del procedimiento y la estrategia requerida para entender el estado actual de la Base de Datos que se desea analizar.
El razonamiento M.A.D.E. ataca la estructura de resolución de problemas mediante una heurística basada en la descomposición del problema: empezando con la formulación conceptual de una consulta, se procede a la búsqueda de un plan de resolución más económico de primitivas, asumiendo las limitaciones del esquema del resultado y del estado actual de la Base de Datos.

3.2.Lenguaje de Consulta Estadístico y de Data Mining

El proceso de Análisis de Datos es un proceso interactivo, el usuario establece los datos que desea analizar, y los mismos deben estar provistos con un conjunto de primitivas para comunicarse con el sistema de Análisis de Datos. Incorporando estas primitivas en un lenguaje de consulta Estadístico y de Data Mining se logra mayor flexibilidad en la interacción del usuario a partir de una base para diseñar una interfaz gráfica. Se logra un standard para una mejor comercialización. Entre las Primitivas; se pueden mencionar:

a) Datos relevantes de las tareas a analizar (descripción y especificación de la Unidad de Análisis),
b) Tipo de conocimiento a analizar (morfismo que va de la hipótesis estadística -hipótesis de investigación + parámetros de evaluación- a la operación de análisis de datos, ya sea una operación Estadística o de Data Mining),
c) Conocimiento "previo" del dominio (reglas jerárquicas que se instancian en el momento de construcción del instrumento de medición, cuestionario, fichero de datos, etc.),
d) Medidas de interés y umbrales para analizar los patrones, medidas y test estadísticos,
e) visualización de los patrones descubiertos.

La Sintaxis se puede resumir en:

a) datos relevantes de las tareas,
b) el tipo de conocimiento a analizar,
c) especificación del concepto de jerarquía,
d) medidas de interés y umbrales, medida estadística y prueba estadística,
e) presentación de patrones y visualización.

3.2.1. Primitivas 'Conocimiento Previo' o 'unidad de análisis'

Como Primitivas de Conocimiento Previo, podemos mencionar la Jerarquía, con diferentes características: a) Esquema de Jerarquía, Ej. calle < ciudad < provincia_o_estado < país; b) Jerarquía de conjunto, Ej. {20-39} = joven, {40-59} = adulto; c) Jerarquía de operación derivada, dirección_email : nombre-clave < departamento < universidad < país; d) Jerarquía de reglas, margen_bajo_beneficio (X) <= precio(X, P1) y costo(X, P2) y (P1 - P2) < $50.[6]

3.2.2.Primitivas 'Datos relevantes de la tarea a realizar' o 'hipótesis de investigación'.

Las Hipótesis de Investigación son proposiciones tentativas acerca de las posibles relaciones entre dos o más variables, requisitos: a) deben referirse a una situación real, b) los términos o variables de la hipótesis deben ser comprensibles, precisas y concretas, c) la relación entre variables propuestas por una hipótesis debe ser clara y lógica, d) los términos de la hipótesis y la relación planteada entre ellas, deben ser observables y mensurables, e) las hipótesis deben estar relacionadas con términos disponibles para probarlas.

Sea Hipótesis de Investigación, una n-upla conformada por <V_i, con i = 1 ...n, R_{ij} con j = 1...m, NM_i, Tipo_hipótesis, Categoría_variable$_i$, m, n >,

V_i = variables involucradas en la hipótesis de investigación,

R_{ij} = valores que puede tener la variable i,

m = cantidad de valores de la variable i,

NM_i = nivel de medición de la variable i, nominales, ordinales e intervalares.

Tipo_hipótesis = tipo de hipótesis de investigación, descriptiva, correlativa, causal, distinción de grupos.

n = cantidad de variables involucradas en la relación,

Categoría_variable$_i$ = categoría de la variable i en función del tipo de hipótesis de investigación, independiente, dependiente, interdependiente o descripción.

3.2.3. Primitivas: Tipos de Conocimiento

Una Hipótesis de Análisis de Datos (Estadística o de Data Mining) es la transformación de las hipótesis de investigación, nulas y alternativas en símbolos estadísticos y de algoritmos de Data Mining. De esta manera una Operación de Análisis de Datos (Estadística o de Data Mining) es un morfismo sobre la hipótesis de Análisis de Datos que involucra la selección de estadísticos, tiene como codominio una operación de Análisis de Datos, que a su vez es una n-upla conformada por < Medidas de interés, umbrales para analizar los patrones medida estadística, prueba estadística>, según corresponda.

Los diferentes tipos de conocimiento ayudan a construir decisiones sobre el tipo de relación permitida entre las variables, y por consiguiente a la selección de la técnica de análisis de datos correspondiente. Los diferentes tipos de conocimiento son [10]:

a) Nivel de la descripción que consiste en caracterizar un fenómeno o situación concreta indicando sus rasgos más peculiares o diferenciadores. Se trata de una enumeración en la que se hace una especie de inventario de las cuestiones precedentemente indicadas;
b) Se pasa de la descripción a la explicación a través de un nivel intermedio, el de la clasificación. Para definir las relaciones entre varias categorías de fenómenos es preciso que estas categorías hayan sido determinadas con precisión. La clasificación agrupa armónicamente fenómenos semejantes y de este modo reduce la innumerable variedad de hechos concretos a cierto número de “tipos”, de ahí el nombre de tipología. Los datos y fenómenos se ordenan, disponen o agrupan en clases sobre la base del descubrimiento de propiedades comunes. Este nivel exige un mayor nivel de sistematización, categorización y ordenación. Es una tarea de categorización, consistente en agrupar objetos discriminándolos, dentro de un conjunto, en una serie de subconjuntos. Esta discriminación se hace de acuerdo a ciertas actitudes, características, cualidades o propiedades en común. Agrupar una determinada clase de hechos o fenómenos y conocer su distribución es una forma de facilitar la manipulación de los mismos, pero no es explicarlos;
c) Nivel de la Explicación también abarca el de la previsión. Puesto que la explicación científica consiste en comprobar la dependencia entre dos fenómenos, A y B, se puede predecir la aparición de B si A se produce. Nivel que se identifica con explicar la causa de un fenómeno, y/o insertar el fenómeno en un contexto teórico, de modo que permita incluirlo en una determinada generalización o legalidad. Es el nivel más profundo de la investigación.

3.2.4. Primitivas: Visualización de Patrones

Con respecto a las Primitivas de Visualización de Patrones se puede mencionar que diferentes dominios requieren diferentes formas de representación, v.g. -reglas, tablas, gráficos de torta o de barras, etc. Aquí el concepto de jerarquía es también importante- , el conocimiento descubierto

podría ser más fácil de entender cuando se representa en un nivel de abstracción alto. Drill up/down, pivoting, slicing and dicing proveen diferentes perspectivas para los datos [6].

4) TEMAS INVOLUCRADOS EN EL PROYECTO

Las principales áreas involucradas en el proyecto son las siguientes: a) Descubrimiento de Conocimiento en Bases de Datos, b) Data Mining, c) Lenguajes de consulta de Data Mining, d) Estadística Computacional, e) Lenguajes estadísticos, f) Análisis Inteligente de Datos, g) Sistemas Estadísticos Expertos, h) Tratamiento Informático de encuestas, i) Inteligencia Artificial, j) Métodos Empíricos para la Inteligencia Artificial, k) Aprendizaje automatizado, l) Sistemas de Información Gerencial y m) Metodología de Bases de Datos

5) CONCLUSIONES

Este trabajo, en particular respecto a la conformación de parámetros de evaluación que incluye técnicas de Data Mining y Estadísticas; (el M.A.D.E.), procura entrar en una fase de utilización de herramientas de Análisis de Datos, más reflexiva y racional. El principal objetivo en su desarrollo consiste en aprovechar, desde el Data Mining, la experiencia de la tradición estadística con análisis de datos.

Con esta integración de disciplinas instrumentadas en esta aplicación se espera mejorar la calidad de servicio de las herramientas de las que dispone el usuario confrontado al análisis de importantes volúmenes de información, incrementando sus recursos para una mejor toma de decisiones.

REFERENCIAS

1. Carla Basili y Leonardo Meo-Evoli, *A deductive processor for statistical databases*, Diciembre 1995, Instituti di Studi sulla Riserca e sulla Documentazione Scientifica, CNR, Roma, Italia.
2. Bonifacio Martín Del Brío y Alfredo Sanz Molina, *Redes Neuronales y Sistemas Borrosos, Introducción teórica y práctica*, Ed.Ra-Ma, 1997, Universidad de Zaragoza, España.
3. W.S.Sarle, *Neural Network and Statiscal Models*, Proceeding of te 19th. Annual SAS Users Group International Conference, Cory NC, Abril 1994, USA.
4. John Maindonald, *Data Mining from Statistical perspective*, Statistical Consulting Unit of the Graduate School, Australian National University, s/f., Australia.
5. Paul R. Cohen, *Empirical Methods for Artificial Intelligence*, MIT Press, Cambridge, Massachusets, 1995, USA.
6. Jianei Han y Micheline Kamber, *Data Mining, Concepts and Techniques*, Morgan Kaufmann Publishers, 2001, San Francisco, USA.
7. Andrés González Carmona et.al., *Desarrollo de un Sistema Experto estadístico para imputación automática de Datos y su posterior tratamiento estadístico*, Instituto de Estadística de Andalucía, Año de Edición 2000, Andalucía, España.
8. R. Haux, *Statistical expert systems – some problems and some new views*, 9th German workshop on Arificial Intelligence, 313-322, Berlin. Springer. 1985.
9. Félix Aparicio Pérez, *Tratamiento Informático de Encuestas*, Ed.Ra-Ma, Abril 1991, Madrid, España.
10. Roberto Hernández Sampieri y otros, *Metodología de la Investigación*, Mc Graw Hill, 2da.Edición, Enero 2000, México.
11. Andrews, F.M., Klem, L. Davidson, T.N., O'Malley, P.M., And Rodgers, W.L., *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data*, 2nd Ed., Survey Research Center, Institute for Social Research, The University of Michigan, 1981, USA.