- Introducción -

La fusión de datos o emparejamiento estadístico (statistical matching) es una técnica que completa conjuntos de datos de diferentes fuentes para conseguir un archivo, aunque artificial, con todas las variables de interés. Es un medio de limitar la recolección de datos, reconstruyendo la información faltante. Es una estimación estadística de los datos faltantes. No es un problema de análisis estadístico con datos faltantes en el cual se consideran los mecanismos que conducen a la ausencia de datos[1].

En el caso de la fusión de datos, se presentan bloques completos de datos ausentes, en general muestras independientes.

El objetivo de la fusión de datos es por lo tanto, el obtener un solo archivo que pueda ser analizado posteriormente con herramientas de minería de datos (data mining tools).

Existe sin embargo un riesgo al usar esta nueva información, pues se debe tener cuidado dado que son en realidad estimaciones y no observaciones y por tanto, conllevan una incertidumbre asociada.

Con frecuencia es imposible obtener toda la información de la misma muestra cuando son muchas las variables cuyos valores deben ser medidos o recolectados. En ocasiones es imposible tener la muestra completa, se trata entonces de aprovechar la información de datos secundarios. Por ejemplo, para eliminar la molestia en los encuestados y por lo tanto evitar sesgo, se trabaja con dos muestras independientes diferentes, en las cuales se separan las variables de interés en dos partes con un conjunto común de variables.

Una representación gráfica general del problema puede ser la siguiente:

N_0 {	X_0	Y_0	?
N_1 {	X_1	?	Z_1

Figura 1 Problema general

[1] Little, R.J.A; Rubin, D.B. 2002

El primer archivo contiene N_0 unidades (individuos) con información sobre p+q variables. El segundo archivo contiene N_1 unidades con información sobre p+r variables. Con frecuencia N_0 es mayor que N_1 aunque no es condición necesaria. Bajo este esquema, se llamarán a las p variables, **variables comunes** y a las q + r variables, **variables específicas**. El problema es entonces llenar las partes faltantes de la tabla.

La fusión cae dentro de la metodología Hot-Deck, dentro de la cual se tienen dos grandes familias. La primera conocida como el Método Alemán, que consiste en realizar primero una tipología conjunta de los individuos donantes y receptores según las variables comunes. A continuación se asignan los valores faltantes de los individuos receptores en función de la clase a la que pertenecen (Wiegand 1986). La segunda familia es la del vecino más próximo: para cada individuo receptor se obtiene el vecino o vecinos más próximos del archivo donante y a continuación se transfieren los valores faltantes (G. Santini 1984). Santini explica que la fusión estadística se presenta en muchas situaciones que dependen de la problemática en concreto. Por ejemplo, encuestas paralelas, cuestionarios alternados, paneles decalados en el tiempo etc. Estas situaciones tienen una característica común o "forma canónica", y es que están formadas por dos subconjuntos; receptores y donantes. Se trata de hacer una imputación masiva de todo un bloque de información. No se debe confundir por tanto, con la imputación clásica de datos faltantes que requiere un modelo estadístico para imputar valores faltantes y en el cual se debe de considerar el mecanismo de generación de los valores faltantes.

El concepto de fusión de datos se inició y extendió en los años 60 y 70 en Alemania y Francia para asociar información de medios impresos y hábitos de audiencia televisiva con información de compra debida a las necesidades de planeación de medios.
En Estados Unidos y Canadá, las encuestas son emparejadas desde 1970 por las oficinas federales para conseguir una mejor información de los ingresos familiares.
En España, el instituto que mide las audiencias de televisión es SOFRES AUDIENCIA DE MEDIOS. El Instituto dedicado a la medición del consumo es DYMPANEL. Pertenecen al grupo TAYLOR NELSON SOFRES.
Las bases que se fusionan pueden ser virtualmente de cualquier tipo.

Algunos ejemplos de proyectos de fusión de datos:

- En Argentina, Brasil y México, IBOPE Información de Medios opera paneles de medición de audiencia televisiva TAM (Television Audience Measurement). En estos mismos países, Kantar Media Research, junto con socios locales, conduce estudios de uso de productos multimedia[2] conocido como TGI (Target Group Index). Estas bases de datos se fusionan en una sola fuente de datos que proporciona información detallada de audiencia televisiva para los usuarios de productos y servicios TGR (Target Group Ratings).[3]

- "Data Enhancements to Improve the Scope and Reliability of Micro simulation Models".
  El objetivo del proyecto: producir un conjunto de datos con información detallada de gastos e ingresos en familias de Gran Bretaña. Los datos generados serán una fuente de información para un modelo de micro simulación de beneficios de impuestos para investigar el impacto de cambios en la política gubernamental. Se desarrolló un método para mejorar la encuesta de recursos familiares 1995/6 (FRS Family Resources Survey) con datos de gastos de la encuesta de gastos familiares (FES Family Expenditure Survey).[4]

Se tratará en este trabajo el problema de completar dos archivos que contienen un subconjunto de variables comunes observadas. La propuesta presentada se basa en la técnica de imputación hot-deck por vecinos más cercanos empleando el algoritmo de Fukunaga/Narendra. Los objetivos de esta tesis en términos generales son:

- Desarrollo de un sistema completo de fusión de datos.

[2] El concepto multimedia designa todas las posibles combinaciones de las computadoras, las telecomunicaciones y la informática; las aplicaciones multimedia comprenden productos y servicios que van desde la computadora hasta las comunicaciones virtuales que posibilita Internet, pasando por los servicios de vídeo interactivo en un televisor y las videoconferencias.
[3] Soong, R. (2001). *47th Annual Conference of the Advertising Research Foundation.* New York
[4] Cohen, ML (1991)

- Propuesta de criterios de validación de la calidad en una fusión de datos
  Se contempla en esta propuesta, el manejo, dentro del sistema a desarrollar, de variables categóricas y continuas, así como de las opciones determinísticas y estocásticas para el proceso de imputación. En lo que se refiere al algoritmo para la búsqueda de los vecinos más cercanos, se han hecho modificaciones y mejoras. Se proponen diferentes alternativas de imputación, así como el determinar los valores adecuados de los parámetros para cada alternativa que así lo requiera.

Los objetivos buscados en la imputación son:

- Coherencia
  Evitar resultados imposibles (Por ejemplo, imputaciones en las cuales los resultados sean: hombres embarazados)

- Precisión en la imputación.

- Reproducción de la distribución condicional f (Y|X).
  Se desea medir el grado de correlación entre las distintas variables, tratando de establecer si esta se presenta en los receptores de manera similar como se presenta en los donantes. Para esto se definirá el concepto de homogeneidad interna y externa

  - interna (correlaciones entre las variables específicas)

  - externa (correlaciones entre las variables comunes y específicas)

Sin importar la sofisticación matemática que se incluya, la esencia del problema debe ser si la información disponible permite que el proceso de fusión tenga éxito o no. El proceso de fusión involucra procesos sobre variables que son comunes a los archivos que se van a fusionar y por lo tanto la calidad del resultado dependerá de la efectividad de estas variables comunes. Si estas variables comunes resultan irrelevantes para el proceso en

cuestión, ninguna sofisticación matemática ayudará. Es por eso que resulta importante el establecer mecanismos de validación de los resultados de las imputaciones.

Esta evaluación toma en consideración lo siguiente:

- capacidad predictiva de las variables comunes sobre las variables específicas.
- características del método seleccionado de fusión.
- métodos usados para evaluar estadísticos (coeficientes de correlación / regresión, tablas de contingencia, etc.)

Así mismo, se utilizarán los métodos EM y PLS con los mismos archivos de datos para tener un marco de referencia y poder establecer comparaciones.
La investigación desarrollada y aquí presentada está contenida parcialmente en el proyecto europeo ESIS[5] .

Este trabajo se presenta en 5 capítulos
El primer capítulo está dedicado a la definición y aproximación del concepto "fusión de datos". Se presentan algunos enfoques para el tratamiento de los datos, se revisan, no de manera exhaustiva, algunos algoritmos relacionados con el tema y se muestran aplicaciones de la fusión de datos relacionados con otras herramientas.
En el capítulo 2 se presenta una revisión de los métodos de búsqueda de los k vecinos más cercanos. Se describe con detalle el algoritmo elegido para implantar en el sistema.
El tercer capítulo describe el desarrollo de la propuesta de investigación. Muestra los componentes del sistema. Se presentan los fundamentos y las bases que se han seguido para el desarrollo y la implantación del sistema. Muestra también las ideas propuestas para medir la calidad de la fusión.

[5] (European Satisfaction Index System) Contract IST-2000-31071. Tiene el objetivo de investigar, desarrollar e implementar una herramienta de software para la medición de la satisfacción del cliente en Europa, así como desarrollar un sistema de almacén de datos (datawarehouse) capaz de recolectar y administrar las sucesivas encuestas. Es dentro de esta última parte que la fusión se realiza con el objetivo de mantener completos los archivos y así poder utilizar el modulo de estimación del índice de satisfacción. Esta propuesta está contenida en el WP3 en la tarea T3.4 (módulo de fusión) <www.esisproject.com>.

El capítulo cuarto está dedicado a la experimentación. Está dividido en dos partes. Para esto se emplea un archivo sintético creado para la verificación del sistema y análisis de los resultados obtenidos.
El capítulo cinco reporta las conclusiones y propuestas resultado de este trabajo.

La metodología de fusión para el desarrollo de esta investigación es la siguiente:

- Establecimiento del poder predictivo de los atributos comunes sobre los atributos que se quieren imputar.
- Distribución de las variables comunes en las 2 muestras.
- Definición de un espacio común.
- Búsqueda de los K-vecinos más cercanos.
- Imputación de variables.
- Validación

Los tres primeros puntos de esta metodología se desarrollan aprovechando las prestaciones existentes en el paquete estadístico SPAD[6].
Los últimos tres puntos forman parte de la propuesta en esta tesis.

[6] Sistema estadístico desarrollado por DECISIA, coordinador del proyecto ESIS.