¡Señor... si yo me olvido de ti,
nunca te olvides de mí!

Capítulo 5
- Conclusiones -

El aspecto práctico de este trabajo, ha sido el desarrollo de un sistema de fusión de datos completo. Se ha intentado desarrollarlo de una manera flexible y amigable a la vez que útil y con posibilidades de crecimiento.

La revisión bibliográfica ha permitido analizar el problema del análisis de datos faltantes como una aproximación al tema de la fusión de datos, considerando que la fusión de datos es un caso particular del análisis de datos faltantes. En este caso, se está hablando de bloques de datos faltantes y en muchas ocasiones, datos faltantes por diseño.

Como se ha visto, son muchas las técnicas que existen para tratar la reconstrucción de los datos faltantes. En este trabajo, el enfoque se ha hecho sobre la técnica de imputación Hot deck. Esta técnica se basa en el reemplazo de la información faltante, utilizando la información de los individuos más parecidos. Se requiere por lo tanto de la definición de semejanza, que en este caso se define a partir de distancia. Para este objetivo, se estudiaron los distintos procedimientos no paramétricos de discriminación para efectuar la búsqueda de los individuos más cercanos (parecidos).

Existen algunos algoritmos diseñados para la búsqueda de los vecinos más cercanos. Este trabajo se basó en el algoritmo de Fukunaga/Narendra. El enfoque básico consiste en descomponer jerárquicamente las muestras (entrenamiento) en subconjuntos disjuntos, y después aplicar a los grupos resultantes el método de Branch and Bound. Este algoritmo se modificó para incluir las opciones de búsqueda con restricciones y búsqueda de los vecinos más cercanos para los donantes, con el fin de efectuar validaciones posteriores. El algoritmo requiere de la construcción de un árbol para la búsqueda, por lo tanto, es importante considerar algunos aspectos para su recorrido como son los nodos terminales o el número de variables empleadas.

Se construyeron algunas gráficas para este objetivo y para intentar evaluar el costo del algoritmo.

Se han propuesto distintas formas de imputación:

- determinística
  - El vecino más cercano (con probabilidad de repetir donantes)
  - Los K vecinos más cercanos

- estocástica.
  - El vecino más cercano
  - Los K vecinos más cercanos

Se propusieron diferentes medidas de calidad para evaluar el proceso de fusión, resumidas como:

- Comparación de estadísticos marginales sobre variables comunes: E[X0] y E[X1].

- Comparación de estadísticos marginales sobre variables específicas: E[Y0] y E[Y1].

- Comparación de coherencia interna: Cor (Y_0) y $\mathrm{Cor}\left(\hat{Y}_1\right)$.

- Reproducción de las correlaciones(X0, Y0) sobre los valores imputados $(X_1, \hat{Y}_1)$.

- Cálculo de errores (medida de exactitud).

Se construyeron algunas gráficas como ayuda a la interpretación de los resultados.

Para probar el sistema desarrollado, se construyeron dos escenarios para simular los distintos métodos de imputación. Estos escenarios incluyeron la selección aleatoria y no

aleatoria de los individuos para la formación de los conjuntos de donantes y receptores. Así mismo, las bases empleadas en cada escenario fueron construidas con distintos niveles de variabilidad (ruido).

Los resultados obtenidos y por lo tanto las conclusiones, se refieren a un conjunto de datos concretos; concesión de productos y servicios bancarios. Por lo tanto, no se puede decir que estos sean generales.

La importancia del parámetro P (probabilidad de repetir un donante) se pone de manifiesto en los resultados del ensayo T1DM. Con este método se asegura la coherencia de los datos imputados. El valor de P = 0 en los ensayos desarrollados parece dar los mejores resultados.

El método TKDM muestra buenos resultados para el índice de precisión y no parece depender del valor de K, solo en el caso de ruido aleatorio amplificado. En el caso de ASL y ACD, los mejores resultados se obtienen para valores pequeños de K (entre 5 y 10). Vale la pena mencionar el efecto agujero negro (un desplazamiento de los valores asignados hacia el centro de la distribución). Esto se presenta en una distribución de individuos con una zona de muy fuerte densidad rodeada de capas con menor densidad.

El método TKSM mejora la reproducción de la desviación estándar. Los valores del índice ASL mejoran para el caso de ruido aleatorio amplificado. El valor del índice de precisión disminuye en el caso de ruido aleatorio amplificado especialmente.

Los valores del índice de precisión mejoran en función de los valores crecientes de π. Parece no influir su efecto en el índice de homogeneidad interna y en el caso de homogeneidad externa, el valor del índice disminuye en función del valor de π. Los mejores valores se pueden observar para valores de π cercanos a 1.

En la parte III de la simulación se trabajó sobre un conjunto de datos proporcionados por el Instituto de Estadística de Cataluña (Idescat) relacionados con el estudio sobre la utilización de las Tecnologías de la Información y las comunicaciones (TIC) en los hogares

de Cataluña correspondientes al mes de noviembre de 2002. Los datos representan una parte de la muestra y están relacionados con los sujetos que se conectan a Internet. En esta muestra se hace un seguimiento especial para conocer el uso de Internet en sus distintas categorías. Estos datos permitieron hacer uso de las variables mixtas y el manejo de transformaciones a las variables.

Los ensayos se realizaron también con el método EM y el método PLS para tener un marco de referencia y punto de comparación. Tal vez por la forma como fueron construidos las bases de datos para los distintos ensayos (modelos lineales) se explique la ligera ventaja de estos métodos en algunos índices.

No se puede concluir a partir de estos resultados el método que debe ser usado en un proceso de fusión de datos. Algunos métodos tienen mejor rendimiento para algún índice de validación que otros. Los resultados de un proceso de fusión de datos, en términos generales, van a depender de la naturaleza de los datos y de las relaciones que existan entre las variables.

Es por esto que el método o los métodos que se escojan para la imputación, deberán tener en cuenta el objetivo que se busca. Sin embargo, es posible marcar algunas líneas generales de actuación como referencia a partir de las características de cada uno de los métodos.
Así por ejemplo, decidir entre escoger un vecino o K vecinos, de manera determinística, se debe considerar lo siguiente: se obtendrá mejor precisión empleando K vecinos, aunque la homogeneidad interna como la externa será mejor evaluada si la elección es de un vecino.
En cuanto al número de vecinos, parece razonable el uso de pocos vecinos (entre 5 y 10).

Los resultados obtenidos, algunos esperados, resultaron satisfactorios y permiten ver en este sistema posibilidades de mejora. Sin embargo, es importante insistir en la "contradicción" en los objetivos: Precisión vs Homogeneidad. Como se ha visto, se pueden conseguir buenos índices de precisión con los métodos deterministicos, a través de la media condicional y con el valor de K = 1 asegurando coherencia de los datos imputados. Por su parte, buenos índices de homogeneidad se pueden conseguir con los métodos estocásticos.

Resulta difícil establecer una regla para el usuario que le permita establecer el valor adecuado de los parámetros a usar en un proceso de fusión, debido a la naturaleza de los datos. De ahí la importancia de mencionar el sistema "completo" de fusión de datos. Esta idea permite trabajar con variables continuas y categóricas y algo que resulta importante, es la posibilidad de determinar el valor del parámetro adecuado por medio de las distintas corridas que se puedan desarrollar y graficar para la mejor toma de decisión. Queda entonces a la decisión del usuario a partir del objetivo que busque y de la información adicional de la que disponga, la determinación del parámetro o parámetros de deberá emplear.

Aunque mucho está escrito sobre la imputación Hot deck, siempre habrá espacio para proponer nuevos métodos. Evaluar la calidad de la fusión seguirá siendo tema de interés. Queda claro que la fusión de datos satisface la necesidad de muchos de proporcionar una sola fuente de datos completa a los usuarios finales. Sin embargo se debe tener cuidado al hacer uso de esta información (son estimaciones, no valores reales observados).

Este trabajo está contenido parcialmente en el proyecto europeo ESIS[1]. El desarrollo y la implantación del sistema se hicieron empleando visual C++ en su mayor parte, Visual Basic para las interfaces y MatLab para las gráficas. A partir de este proyecto se derivó este trabajo de tesis extendiendo su alcance.

Perspectivas

Tal vez, entre los temas importantes que se pueden continuar investigando, está la parte teórica relacionada con estos temas. Determinar las distribuciones de probabilidad teóricas de los índices empleados, para establecer niveles de significación. Esto podría se abordado a través de simulaciones para construir la distribución de los distintos índices y de esta manera, tener otra herramienta para evaluarlos (intervalos de confianza).

[1] (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>.

Uno de los pendientes de los métodos multivariantes con K vecinos, es el índice de homogeneidad, tanto interna como externa para la relación continua-categórica.
Resulta atrayente entonces, el desarrollo de algún método de imputación multivariado conjunto que permita preservar esta relación.

Como ya se ha mencionado con anterioridad, sería deseable disponer de datos sobre diferentes aspectos para una muestra suficientemente grande y representativa. Tal cantidad de datos sería ideal aunque en la realidad no es fácil. Es común que la información sobre distintos aspectos de un mismo colectivo esté recogida en diferentes bases de datos. Por ejemplo, encuestas sobre temas diferentes se llevan a cabo sobre muestras diferentes de una misma población.

El problema que se presenta es el siguiente:

Archivo A

$X_1, \ldots, X_p$	$Y_1, \ldots, Y_q$	?

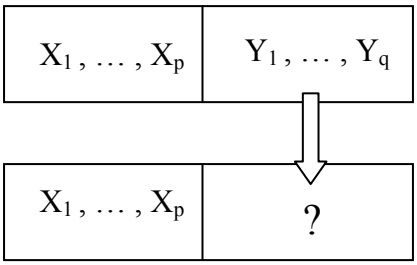
Archivo B

$X_1, \ldots, X_p$	?	$Z_1, \ldots, Z_r$

Se dispone de $N = N_A + N_B$ observaciones. Existe un conjunto común de variables y un grupo específico de variables. El objetivo es enlazar la información de ambos archivos para aproximar los resultados que se hubieran obtenido con los datos completos.

Se puede analizar parte de la relación que existe entre las variables específicas que se puede transmitir a través de las variables comunes. No se puede analizar la relación que existe entre Y y Z cuando X es constante ya que no se dispone de alguna observación conjunta de los tres grupos de variables. El problema antes mencionado, puede ser atendido con el sistema desarrollado en este trabajo a través de un proceso secuencial.

En una primera etapa se imputan los valores faltantes $\hat{Y}$ con alguno de los métodos incluidos en el sistema.



A partir de aquí, el problema será la imputación de los valores Z. Ahora el problema es el siguiente: