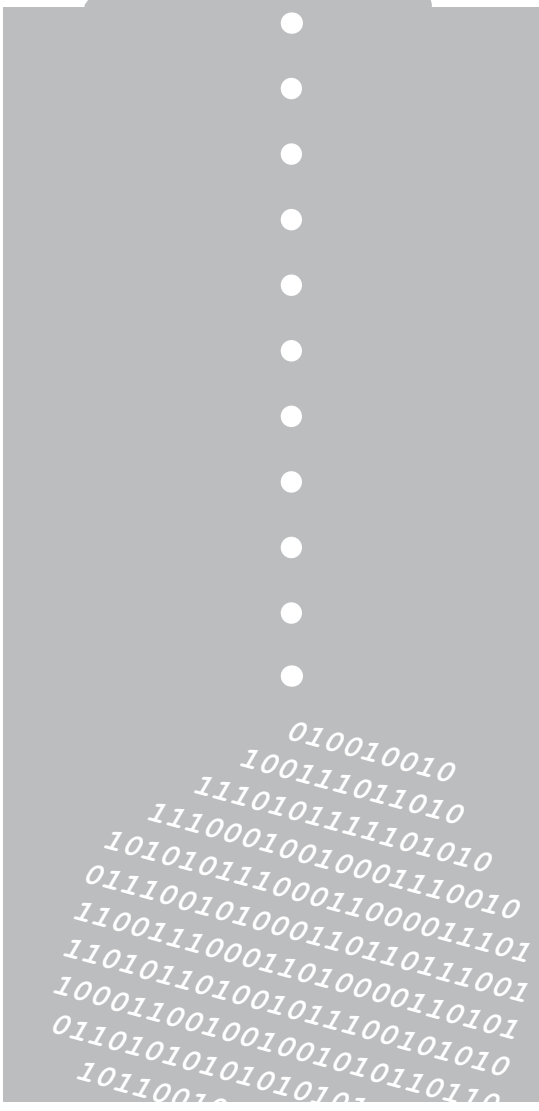
Data mining en Exactas

Mucho mejor que deshojar una margarita

Por Guillermo Mattei | gmattei@df.uba.ar



Empate con sabor amargo: el pase a la semifinal de la Copa Mundial 2006 de Fútbol lo deciden los penales. Leonardo Franco y Jens Lehmann son los principales protagonistas del duelo. Con frialdad germánica, un asistente le acerca un papel a Lehmann. En ese acto el arquero europeo adquiere información de alto valor agregado: nada menos que las regularidades y tendencias estadísticas, extraídas de numerosos datos, correspondientes a la crónica de los penales ejecutados por todos los jugadores de la selección argentina en lo que va de sus carreras deportivas. Para elegir la punta a la cual tirarse ante cada ejecutor argentino de penales, Lehmann no usa azar sino conocimiento. La fuente de ese conocimiento se llama data mining, o "minería de datos". Y Lehmann se ataja todo.

La Dirección de Sistemas de la Facultad de Arquitectura, Diseño y Urbanismo (UBA) funciona en el cuarto piso del pabellón III de Ciudad Universitaria y se encarga de coordinar todos los recursos informáticos de esa casa de estudios. El licenciado en sistemas de la Facultad de Ingeniería (UBA), Fernando Zimenspitz, es su director desde hace varios años. Como ocurre con otros profesionales de perfil similar a Zimenspitz -es decir, en puestos claves del área de informática, tanto en el ámbito público como en el privado- la Maestría en Explotación de Datos y Descubrimiento del Conocimiento (Facultades de Ciencias Exactas y Naturales e Ingeniería de la UBA) logra nuclearlos en un ámbito donde confluyen el conocimiento de vanguardia y sus aplicaciones en medios productivos o de servicios.

"En 2005 me incorporé a la Maestría que, entre sus estudiantes, se conoce como *Data Mining* o Minería de Datos, por la analogía con encontrar el oro, aunque sea en pequeñas cantidades, que está oculto en medio de muchos otros minerales de escaso valor económico", grafica Zimenspitz, y agrega: "nuestro oro es el conocimiento enmascarado en cantidades colosales de datos acumulados a partir de procesos de adquisición muy diversos".

Datos y más datos

No es muy dificil advertir que las bases de datos de organizaciones y estructuras tales como bancos, universidades, compañías de telecomunicaciones y organismos estatales, entre muchos otros, tuvieron crecimientos brutales en los últimos años. Luego, casi naturalmente, surgió la pregunta "¿qué hacemos con estos datos? A los millones de transacciones mundiales de los últimos cinco años, ¿cómo les sacamos algún valor que trascienda la mera acumulación preventiva?" Zimenspitz explica: "Lo que busca el data mining es extraer, de grandes volúmenes de datos, las regularidades relevantes, la información valiosa escondida, y hasta ese momento desconocida, de una manera rápida, eficiente y escalable (es decir, que las técnicas sigan siendo viables aunque el volumen de datos crezca considerablemente)". En estas escalas monstruosas, no vale explorar todas las combinaciones posibles entre los datos y ver qué da, porque el tiempo de procesamiento sería ridículamente inmenso y, en la práctica, el problema resultaría incomputable. Luego, no quedó otra que aguzar el ingenio y el resultado fue el *data mining*.

La fiebre del oro

Zimenspitz opina que, como sucede en muchas ramas del conocimiento, las buenas ideas no necesariamente dan frutos inmediatamente sino que, a veces, hay que esperar a que la tecnología, tanto por evolución como por caída de costos, se ponga a su altura para poder fructificar. En otras palabras, el punto de inflexión tecnológico catapultó al *data mining* como novedad.

El *data mining* tiene tres pilares de sustentación: las bases de datos, por un lado; el aprendizaje automatizado, los algoritmos y la inteligencia artificial, por otro y, finalmente, la estadística.

"En principio, lo que hacemos son análisis exploratorios, distintos a los de tipo confirmatorio que da la estadística, que nos permiten ´ver´ –algunas veces en forma literal– cómo están dispuestos los datos, cómo es su distribución, si se forman grupos o configuraciones", explica Zimenspitz. Siempre y cuando se tomen los recaudos pertinentes a la hora de interpretar, las técnicas de visualización disponibles son muy útiles en la exploración de datos.

En la jerga de las bases de datos, una *query* es una consulta a una cierta base. Por ejemplo: consultar a la base de datos de una Facultad acerca de los alumnos ingresantes en tal año, ordenados por su promedio actual y que hayan aprobado entre 20 y 30 materias. En ese caso, la persona que consulta sabe exactamente lo que quiere pero, en el análisis exploratorio del data mining eso no ocurre: "no sabemos qué hay, y vamos a ver qué encontramos", acota Zimpenspitz. Por ejemplo, consultar acerca de "todos los clientes que tienen una probabilidad de irse de un banco mayor al 80%" es una query que no se resuelve directamente sino que, antes, hay que elaborar un modelo matemático capaz de predecir comportamientos y resultados del fenómeno bajo análisis.

Cuando los expertos quieren construir un modelo, los datos sirven para estimar su rango de validez y aplicación. "Lo que hacemos es destinar una parte de los datos como conjunto de generación, entrenamiento y ajuste del modelo y otro conjunto diferente para validar el modelo", indica el experto.

Zimenspitz aclara: "Nuestro punto de partida es contra qué nos comparamos. En principio, nos comparamos con el azar. En primera instancia, tenemos que ser mejores que el azar pero, en segunda, *mucho* mejores. Cuanto mejor predigas, mejor habrá sido el análisis de los datos". Por ejemplo, *predecir* la lista de los clientes con mayor probabilidad de darse de baja del banco, le permite a la gerencia

Los fierros de la minería

Una PC permite hacer algo de *data mining* pero la clave es el volumen de datos y los objetivos a alcanzar. Si bien hay algoritmos que son eficientes y responden en tiempo razonable, la utilización de memoria es muy importante, dado que es la que fija cotas a lo posible. Con grandes volúmenes de datos, el hardware y software especializado es indispensable.

El hardware es solo cuestión de dinero: su elección se basa en la capacidad de proyectar certeramente sus aplicaciones para no gastar de más. Por supuesto que existen megaproyectos, por ejemplo de mediciones astronómicas por medio de satélites, que implican volúmenes de datos lo suficientemente descomunales como para requerir grillas de computadoras y procesamiento en paralelo sincronizado.

Por su parte, existe ya software comercial y de código abierto apto para manejar grandes y variados volúmenes de datos. Como, en general, ese software carga datos y los analiza en memoria, el sistema operativo, que es necesariamente acotado, marca los límites.

Cómo retener clientes sin mucho costo

La información mensual relacionada con la cartera de clientes de un banco incluye, por ejemplo, el dato de los clientes que deciden darse de baja o no de una tarjeta de crédito. Con la información correspondiente, el equipo de trabajo de Zimenspitz en la Maestría se planteó el objetivo de encontrar características de clientes con más chances de darse de baja al mes siguiente, tratando de optimizar los costos de las acciones de retención que el banco debería hacer para evitar la fuga y del beneficio que se lograría si el cliente decidía continuar con ese producto.

Para resolver el problema, los mineros utilizaron los llamados algoritmos de árboles de decisión, que permiten extraer reglas para inferir, a partir de los datos, si un cliente se daría de baja o no de la tarjeta. En esta lógica, los especialistas construyeron una cantidad siginificativa de árboles para luego seleccionar el algoritmo y el modelo más conveniente para optimizar la función de costo. Esta selección se realiza visualizando las propiedades comparativas de los árboles construidos.

Una vez seleccionados el algoritmo y el modelo más adecuados, los mineros determinaron las reglas que permitieron ordenar a los clientes de mayor a menor chance de darse de baja de la tarjeta.

Finalmente, el equipo entregó a la gerencia del banco la lista de los clientes (los primeros de la lista anterior hasta un cierto punto de corte sobre los cuales la entidad debería aplicar una política de retención de costo mínimo.

De esta manera, el banco solo invirtió en promociones y publicidad focalizadas sobre la fracción más inestable de sus clientes, tal cual habían descubierto los maestrandos. Maximizar la retención y minimizar los gastos es la fórmula de la felicidad del banquero y hallar los datos de oro entre los de roca, la del minero informático.

respectiva decidir la aplicación de políticas selectivas de retención y tomar medidas para minimizar razonablemente las bajas de ciertos clientes. En este caso, no hace falta ofrecer promociones y beneficios a la cartera completa de clientes, porque su costo sería mayor. Por lo tanto, el *data mining* contribuye a hacer ajustes finos en la toma de decisiones y, por cierto, en escalas de grandes empresas, estos ajustes pueden significar sumas millonarias.

Sin embargo, en muchos casos, el miedo del minero es no encontrar *el oro*. Puede suceder que, de entre los datos, surja un patrón evidente pero, al presentarlo a una gerencia, el miembro más veterano del directorio exclame: "¡esto ya los sabemos por olfato comercial desde hace años!, ¿para esto te pagué, pibe?"

Si bien los algoritmos son exactos, probados y calculan a la perfección, eventualmente, pueden no ser los mejores para emplear en una determinada situación: ahí juega el criterio analítico del minero. El buen analista no solo debe manejar conocimiento sino, además, ser capaz de dialogar constructivamente con el experto mejor relacionado con el costado más fáctico de los datos.

"Debemos probar nuestros modelos con datos que no participan en sus generaciones para no caer en lo que se conoce como *overfitting* (o incapacidad del modelo para aportar reglas suficientemente generales no influenciadas por la *suciedad* de los datos) y también para no errarle en una medida tal que corra riesgo nuestro puesto laboral en la empresa...", ironiza Zimenspitz. Muchas veces los errores pueden significar *sólo* la pérdida del cincuenta por ciento del capital de una empresa pero, en ciertos casos, puede ir la vida de un paciente en un diagnóstico deficiente, o el despegue indeseado de un misil intercontinental.

Las transacciones en las bases de datos, ¿siempre están bien hechas?; los datos ¿siempre están perfectamente cargados?; los sistemas, ¿siempre validan? No, eso no sucede en la realidad. Por ejemplo, un operador telefónico carga datos de un nuevo cliente en su computadora. Cuando pregunta por el código postal, el cliente le dice que no lo recuerda, y entonces el operador pone "0000" o "1111". En la jerga, estos pseudodatos introducen un metafórico *ruido* indeseado. "Si no hay limpieza y preparación de datos, los resultados pueden poner la eficiencia de la predicción por debajo del azar", afirma Zimenspitz. Independientemente de lo sofisticado que sea el dispositivo, el setenta por ciento del trabajo es limpieza y preparación de datos.

La escuela de minería

En este momento Zimenspitz termina su segundo año de la Maestría y defenderá una tesis para obtener el grado de magíster durante el año 2007. Las orientaciones posibles en esas tesis de magíster son dos: temas comerciales y financieros o de perfil académico y científico. "Yo me inclinaría por una especie de combinación de perfiles tal que, usufructuando el desarrollo de nuevas tecnologías y algoritmos, impacte en cuestiones prácticas pero novedosas", confiesa Zimenspitz.

En la maestría hay alumnos graduados en sistemas, en ciencias económicas y hasta hay físicos que no manejan programación a nivel profesional. "Yo mismo me dedico a la gestión más que a la programación", admite Zimenspitz. Pero siempre es útil que, en un equipo de mineros, alguno sea más afín a los desarrollos de software, tanto para comprar programas, hacerles modificaciones, o encargar software a medida.

"En lo personal, la maestría me cambió la visión conceptual de mi trabajo al poder *ver* los datos de otra manera o permitiéndome aprovecharlos para el día a día", indica Zimenspitz y concluye: "En este momento no sé decir si el *data mining* es una revolución del conocimiento pero seguro puedo decir que produce cambios trascendentales acerca de cómo describir y encarar problemas, aunque las bases de datos transaccionales sigan operando de la misma manera."

La dinámica actual de esta área del conocimiento muestra que, en breve, la cresta de la ola de demanda de profesionales de *data mining* arribará a estas costas. Luego, todo indica que la oportunidad de las Facultades de Ciencias Exactas y Naturales y de Ingeniería de la UBA de hacer historia con las primeras camadas de expertos en *data mining* es inmejorable, única y altamente promisoria.