Big Data, situación y desafíos para la ciencia actuarial

ANTONIO BERLANGA
aberlan@ia.uc3m.es
Prof. Titular CC. de Computación e Inteligencia Artificial. Dpto. Informática, Universidad Carlos III de Madrid

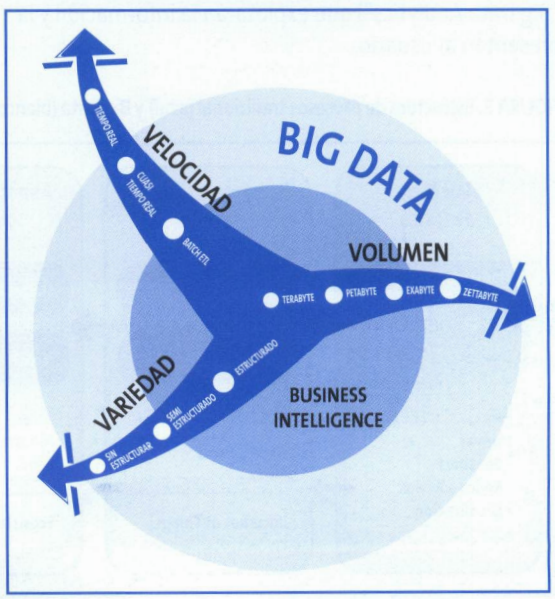
La ciencia actuarial, desde sus inicios, requirió para su desempeño de la acumulación y procesamiento de datos. Con esos datos se podían, entre otros cálculos, realizar tablas de vida o de probabilidad de eventos que permitían evaluar riesgos. En estadística se estableció, con el teorema del límite central, la forma de la distribución de probabilidad de variables aleatorias cuando se tienen muestras suficientemente grandes. Es decir, la calidad de un modelo de inferencia estadístico estará directamente relacionado, entre otras cosas, con la cantidad de datos involucrados. Por tanto, es objetivo prioritario para la ciencia actuarial, disponer de datos, en cantidad y calidad, que permitan la generación de sus modelos.

La posibilidad de tener datos asociados a un evento está relacionada con tres capacidades: la disponibilidad del dato, es decir, la capacidad para poder medirlo ya sea directa o indirectamente, la capacidad para almacenarlo y finalmente, para procesarlo. La evolución de estas capacidades ha estado estrechamente ligada con la evolución en el análisis de riesgos y como consecuencia, con la oferta de servicios en sector de seguros. Estas tres capacidades han experimentado en el pasado reciente un incremento exponencial y se prevé la misma tendencia a futuro. Los dispositivos de almacenamiento masivo a bajo coste, el procesamiento paralelo y de altas prestaciones en la nube y la evolución hacia el registro digital de todos los datos junto a la posibilidad del Internet of Things para ampliar este registro a cualquier dispositivo, alumbra un escenario completamente diferente. Comparando la situación actual con el pasado cercano se puede apreciar la gran evolución experimentada; el coste para almacenar todos los libros escritos en la historia de la humanidad, en torno a 60TB hoy es de 1.000 euros, la capacidad de cómputo de un Apple iphone5 es 2,7 veces la del supercomputador Cray-2 del año 1985 y la previsión para el año 2020 de dispositivos conectados a internet será de 50.000 millones, el doble de los que hay hoy. Sin embargo, las arquitecturas de proceso tradicionales, se ven superadas por el volumen y variedad de información a tratar. Por ejemplo, en media hora de vuelo, los sensores de un avión generan 10TB de datos, cada minuto en Facebook se hacen 500.000 comentarios (360.000 en Twitter), se hacen 300.000 actualizaciones de estado y se cargan 140.000 imágenes, Google cada minuto procesa 2.400.000 búsquedas, necesitando 1.000 ordenadores para dar respuesta a una simple búsqueda en 0,2 segundos. Es para tratar con estas magnitudes de datos y procesos para las que se acuña el término de Big Data.

El concepto de Big Data surge para caracterizar la situación en la que se tienen datos que exceden la capacidad de procesamiento de los sistemas tradicionales de cómputo. El término lo introdujeron, M. Cox y D. Ellsworth, en una conferencia en 1997, aunque fue posteriormente en 2001, D. Laney en un informe técnico, desarrolla el concepto y establece sus características. Es el bien conocido modelo de variedad, volumen y velocidad, o de las "tres uves". Este modelo ha sido extendido a cinco y siete "uves" pero más con un criterio comercial que por una justificación académica.

La situación actual en las empresas es tener sus datos en soportes estructurados, principalmente en ba-

FIGURA 1. Modelo de 3Vs de Big Data



ses de datos. Se registra la información directamente de sus procesos de producción, sin alcanzar el volumen del exabyte y procesando los datos periódicamente, como soporte para la toma de decisión de la dirección estratégica. El concepto Big Data hace referencia a la integración de fuentes heterogéneas (variedad); procesos, ficheros de log, cámaras, textos, sensores, redes sociales, voz, etc., con la necesidad de tomar decisiones en tiempo real o muy cercano al tiempo real (velocidad). Por ejemplo, The Weather Company en alianza con IBM han desarrollado un modelo de predicción meteorológica con datos procedentes de más de 100.000 sensores meteorológicos junto con otros millones de datos que proceden de aviones, edificios, vehículos y teléfonos móviles. En total, se procesan 2.000 millones de datos para realizar 10.000 millones de predicciones diarias. Estas predicciones son explotadas, entre otras, por empresas de generación de energía para el ajuste de su producción y consumo y por compañías aseguradoras que pueden advertir a sus clientes de condiciones meteorológicas adversas que podrían dañar sus bienes.

TECNOLOGÍAS Y RECURSOS

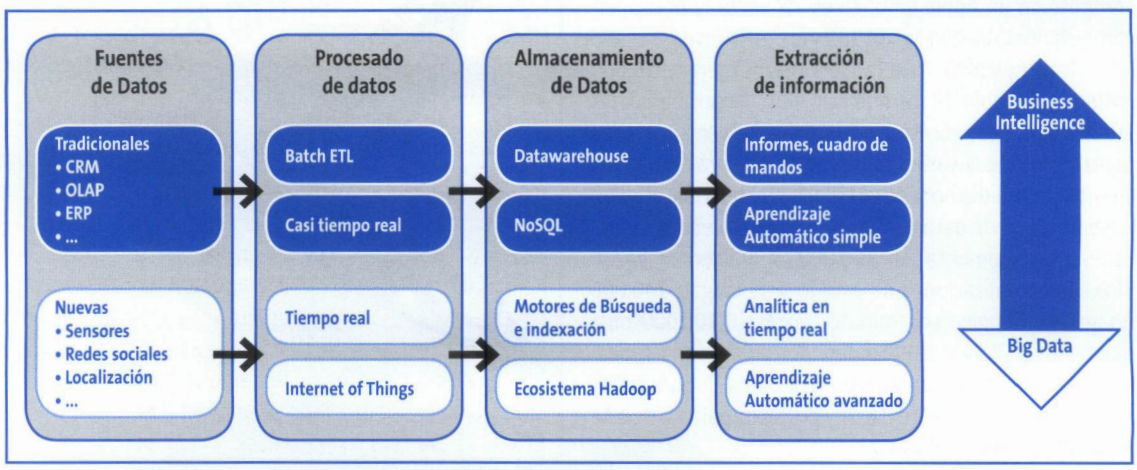
En su origen, el Big Data exponía la necesidad de buscar otras formas para organizar la información y los algoritmos que la tratan y poder superar las limitaciones de espacio local y en red, un problema de arquitectura de sistemas. Pero muy pronto, junto al problema de la arquitectura de sistemas, quedó asociado el problema de determinar qué algoritmos son los más adecuados para extraer la información. De forma, que hoy el concepto de Big Data engloba tanto las soluciones de arquitectura hardware, la arquitectura software de procesos y soporte de información y los algoritmos y herramientas (englobados bajo la denominación de "Big Data Analytics") que explotarán la información y la presentan al usuario.

Estos algoritmos son los propios del análisis estadístico y del aprendizaje automático y buscan estructuras, patrones, relaciones y modelos. El aprendizaje automático engloba a los algoritmos que utilizan un modelo de inducción de conocimiento a partir de ejemplos, en este contexto, a partir de los datos. Algunas de las técnicas más habituales del aprendizaje automático son: redes de neuronas y bayesianas, modelos de Markov, SVM, aprendizaje por refuerzo, algoritmos evolutivos, aprendizaje de árboles de decisión, algoritmos de agrupamiento (K-means, expectation maximization, clusterización jerárquica, etc.).

En función del volumen de datos y la necesidad de procesamiento, el tipo de infraestructura podrá cambiar, pero un requisito indispensable será su escalabilidad. La arquitectura de sistemas habitual será la de un sistema de ficheros y procesos distribuidos sobre una red o clúster de ordenadores, aunque hay una tendencia creciente a virtualizar el hardware, realizar la computación en la nube, utilizando a demanda los recursos de cómputo, proporcionales al volumen de las tareas. Como soporte para el sistema de información y procesos está muy extendido el uso de HDFS (Hadoop Distributed File System), sobre el que se instalan data warehouses y sistemas distribuidos de procesos (Apache Hadoop y Apache Spark).

Las necesidades específicas de infraestructura y software del Big Data y la evolución hacia un negocio orientado por los datos, ha tenido impacto en las organizaciones creando nuevas posiciones directivas. Ha aparecido la figura del director de datos (Chief Data Officer, CDO) encargado del procesado y mantenimiento de los datos. Compañías como HSBC, Generali Group o QBE ya tienen este puesto y Gartner estima que para el 2019 el 90 por ciento de las grandes compañías habrán creado esta posición. Surge también el director de analítica (Chief Analytics Officer, CAO) que se centra en establecer las estrategias que permiten extraer información que puede dar valor al negocio.

FIGURA 2. Estructura de procesos tradicional (azul) y Big Data (blanco)



NUEVAS OPORTUNIDADES

Una aplicación inmediata va a ser en la mejora de los modelos predictivos que experimentaran un sensible incremento en la precisión, aunque el gran cambio estará en la posibilidad de ofertar productos hiper-personalizados. Ya hay compañías que están aplicando bonificaciones a asegurados por tener buenos hábitos de conducción. Mediante dispositivos que se instalan en los automóviles, se pueden obtener datos de velocidades máximas, medias, periodos de descanso, etc. con estos datos se pueden establecer patrones de conducción de riesgo.

Otra área importante es en la interacción personalizada con el cliente en el ámbito de la salud. En este sector, por un lado, crece la demanda de servicios personalizados ya que se ha establecido una relación entre la calidad del servicio y el nivel de personalización.

Por otro, la ciencia médica ha puesto muchas expectativas en las terapias individualizadas, ajustadas al perfil genético y epigenético de cada persona. Simplemente portando un teléfono móvil y una banda de actividad física se podrá trazar, en tiempo real, el nivel de actividad y el estado físico general. El desafío será procesar el gran volumen de fuentes de datos e integrarlos con los datos del perfil genético e historial clínico. Se podrán establecer bonificaciones a los asegurados que desarrollen una actividad física saludable y tener un sistema reactivo de alerta médica.

DESAFÍOS

Hay un conjunto de cuestiones abiertas que tendrán que ser resueltas para que el Big Data pueda consolidarse aportando valor en las empresas. En primer lugar, se debe realizar un esfuerzo en la integración de los sistemas de información que dispone la empresa, dotándolos de la suficiente flexibilidad y escalabilidad para adaptarse a futuros requisitos. En segundo lugar, se deberá facilitar la identificación de los datos que son correctos y útiles. Acumular y procesar datos ruidosos y erróneos suponen un gran desperdicio de recursos de almacenamiento y cómputo y pueden introducir sesgos en los modelos. Un aspecto de especial relevancia para la aplicación del Big Data en el sector de los seguros concierne a los aspectos éticos y de regulación legal. Las directrices europeas y los marcos reguladores estatales han introducido límites tanto a la forma y uso final de los datos personales, perfilados y decisiones automatizadas. Se imponen requisitos estrictos de seguridad, dada la naturaleza de los datos, tanto para su almacenamiento como su trasmisión y se establecen límites para el uso de los perfilados. No se pueden utilizar perfiles que tengan un efecto discriminatorio por razón de raza u origen étnico, religión o creencias, ideas políticas y estado de salud o características genéticas. Se debe informar al propietario de los datos de la finalidad del perfilado y del derecho que tiene a oponerse para que se use o incluso que se le realice un perfil. Además, tendrá derecho a migrar sus datos a otra compañía, lo que permite poner en evidencia ante la competencia su estrategia de análisis de datos.

Finalmente, será necesario incorporar profesionales que puedan manejar las tecnologías involucradas y con capacidad para poder interpretar los datos. Para la gestión de la infraestructura de datos, los perfiles técnicos relacionados con las TIC pueden dar el soporte necesario, si bien, debe estar versado en arquitectura de sistemas, gestión de sistemas de información, computación de altas prestaciones y optimización de algoritmos. Para gestionar el análisis de los datos, están surgiendo nuevos perfiles académicos que complementan a los existentes en TIC con especialidades en inteligencia artificial. Se están creando grados y postgrados que persiguen dar formación para una figura de reciente creación, el científico de datos. Si bien, existe cierta controversia al respecto ya que desde algunas posiciones académicas no se aprecia una sustancial diferencia entre el científico de datos y el estadístico. Podría interpretarse al científico de datos como la intersección entre el estadístico, el informático y el experto en la materia de estudio. El papel de experto debe desempeñarlo el actuario, cuando el sector de aplicación es el financiero y la industria aseguradora. Sería labor del actuario orientar los análisis, dicho de otra forma, es quien debe formular las preguntas y validar las respuestas.

CONCLUSIÓN

En el pasado, en las compañías, se aplicaba análisis descriptivo a los datos para realizar informes que servían de ayuda a la toma de decisión. La información extraída permitía explicar la situación. En la actualidad, el incremento en el volumen y variedad de datos disponibles, permite realizar análisis en tiempo real, teniendo la capacidad, de tomar decisiones también en tiempo real. El conocimiento que se extrae de los datos se aplica de forma reactiva, es el Big Data. Para el futuro, los sistemas buscarán la eficacia, anticipándose, aplicando análisis predictivo con el fin de orientar las acciones que estarán dirigidas por los datos. En este escenario, la industria aseguradora se encuentra ante desafíos importantes. Parte de una posición de ventaja, desde sus inicios, la aplicación de la ciencia actuarial siempre ha estado ligada a la acumulación y procesado de datos para realizar sus modelos. Sin embargo, para el presente y futuro, requiere de la adaptación a las nuevas tecnologías, procesos y marcos regulatorios implicados en el Big Data. Los profesionales actuarios siempre tendrán, como expertos del negocio, un papel que desarrollar en los procesos de análisis, pero la cuestión es si actualizaran sus competencias para poder liderarlos.