**Probabilistic methods for multi-source and temporal biomedical data quality assessment**

Actualmente, la investigación y toma de decisiones en entornos biomédicos dependen en gran medida de los datos almacenados en los sistemas de información. En consecuencia, una falta de calidad en los datos (CD) puede afectar significativamente a la interpretación de los mismos, lo cual puede dar lugar a decisiones sub-óptimas o dificultar los procesos y resultados de las investigaciones derivadas.

Generalmente, la CD es evaluada mediante diversas métricas de las denominadas dimensiones de calidad sobre problemas fundamentales como datos incompletos, inconsistentes o incorrectos. Sin embargo, los actuales desarrollos sobre repositorios de datos biomédicos masivos (Big Data), basados en infraestructuras de compartición de datos multi-institucionales o transfronterizas, requieren nuevas aproximaciones para una evaluación eficiente y desde perspectivas generales de la CD.

Esta tesis tiene como propósito la investigación y desarrollo de métodos para evaluar dos problemas especialmente importantes en repositorios multi-sitio masivos adquiridos durante largos periodos de tiempo: (1) la variabilidad de las distribuciones de probabilidad de los datos entre diferentes fuentes o sitios---variabilidad multi-fuente---y (2) la variabilidad de las distribuciones de probabilidad de los datos a lo largo del tiempo---variabilidad temporal. Esta variabilidad puede estar causada por diferencias en los métodos de adquisición de datos, protocolos o políticas de atención sanitaria; a errores sistemáticos o aleatorios durante la entrada o gestión de datos; diferencias geográficas o demográficas en las poblaciones; o incluso por falsificaciones en los datos. Si esta variabilidad no es gestionada, puede complicar el análisis de los datos, sesgar los resultados, o minimizar la generalización de modelos o hipótesis basadas en los datos.

Hasta la fecha, la variabilidad multi-fuente y temporal han recibido poca atención como problemas de CD, y hasta donde sabemos no cuentan con métodos adecuados para su evaluación. Esta tesis aporta métodos para detectar, medir y caracterizar dicha variabilidad, los cuales han sido especialmente diseñados para superar los problemas que las aproximaciones estadísticas clásicas pueden tener con datos biomédicos multi-tipo, multivariantes y multi-modales, y sin ser afectados por tamaños muestrales grandes en entornos Big Data. Para ello, hemos definido un marco probabilístico común basado en Teoría y Geometría de la Información que da soporte a los métodos desarrollados. Este marco está basado en la inferencia de variedades de Riemann no-paramétricas a partir de distancias probabilísticas normalizadas entre las distribuciones de varias fuentes de datos o a lo largo del tiempo.

Basadas en dicho marco probabilístico se han aportado las siguientes contribuciones:

Para la evaluación de la variabilidad multi-fuente se han definido dos métricas y un gráfico para visualización: (1) la *Global Probabilistic Deviation* (GPD), la cual mide el grado de variabilidad global entre las distribuciones de las diferentes fuentes---como un estimador equivalente a la desviación estándar entre distribuciones; y (2) la *Source Probabilistic Outlyingness* (SPO), la cual mide la disimilaridad entre la distribución de una fuente de datos dada y la distribución de una fuente promedio o global latente definida. Estas métricas están basadas en la construcción de un simplex geométrico (la variedad de máxima dimensionalidad) mediante las distancias entre fuentes. Adicionalmente, se ha definido el *Multi-Source Variability* (MSV) *plot*, para una visualización exploratoria basada en tal simplex, que permite detectar patrones de agrupamiento o desagrupamiento entre fuentes.

Para la variabilidad temporal el método desarrollado proporciona dos herramientas principales: (1) el *Information Geometric Temporal* (IGT) *plot*, para una visualización exploratoria de la evolución temporal de las distribuciones de datos, basada en la proyección de la variedad estadística de las relaciones entre lotes temporales; y (2) el *Probability Distribution Function Statistical Process Control* (PDF-SPC), un algoritmo para la monitorización y detección automática de cambios en las distribuciones de datos. Adicionalmente, este método permite monitorizar la variabilidad multi-fuente a lo largo del tiempo.

Los métodos han sido aplicados en casos de estudio reales en repositorios biomédicos, incluyendo: El Registro de Salud Pública de Mortalidad y el de Cáncer de la Comunidad Valenciana, España; el conjunto de datos de enfermedades del corazón del repositorio UCI; el conjunto de datos NHDS de los Estados Unidos; un conjunto de datos español de Cáncer de Mama; y un conjunto de datos de Fecundación In-Vitro. En particular esta tesis incluye una descripción detallada de los hallazgos de variabilidad multi-fuente y temporal del Registro de Mortalidad, incluyendo: una partición del repositorio en dos subgrupos temporales probabilísticamente separados siguiendo un cambio en el Certificado Médico de Defunción en 2009, anomalías temporales puntuales debidas a incrementos puntuales en el número de datos perdidos, así como departamentos de salud anómalos y agrupados debido a diferencias en poblaciones y en las prácticas.

La aplicación sistemática de los métodos a los casos de estudio ha contribuido al desarrollo de un conjunto de herramientas software, el cual incluye los métodos GPD, SPO, MSV plot, IGT plot, PDF-SPC, otras herramientas básicas de CD, y la generación automática de informes de CD. Finalmente, se ha definido la base teórica de un marco general de CD biomédicos, el cual ha sido utilizado en tres aplicaciones: en el proceso de construcción de repositorios de calidad asegurada para la alimentación del lactante, en la contextualización de datos para el reuso en Sistemas de Ayuda a la Decisión Médica usando un *wrapper* HL7-CDA, y en un servicio on-line para la evaluación y clasificación de la CD de repositorios biomédicos.

Los resultados de esta tesis han sido publicados en ocho contribuciones científicas (en publicaciones en revistas indexadas y en artículos en congresos), en las áreas de Estadística y Probabilidad, Sistemas de Información, Minería de Datos, Informática Médica e Ingeniería Biomédica. Una de las publicaciones fue seleccionada por la IMIA como una de las mejores publicaciones en 2013 en la sub-área de Sistemas de Información de Salud. Adicionalmente, los resultados de esta tesis han contribuido en varios proyectos de investigación, y han facilitado los primeros pasos hacia la industrialización de los métodos y aproximaciones desarrolladas para la auditoría y control de la CD biomédica.