ABSTRACT

The growing digitalization of healthcare systems and the resulting availability of biomedical Big Data have profoundly transformed the way information is acquired, managed, and exploited in clinical and research settings. Yet the heterogeneity of data sources, the incompleteness of biomedical records, the semantic variability of domain terminology, and the demand for interpretable results pose significant challenges to the design of reliable data-driven biomedical processes. These challenges directly affect the ability of domain experts to explore, understand, and trust the information produced by such processes. In this thesis, the entire data-driven biomedical pipeline is examined end-to-end, with a particular focus on how different types of metadata (Documentary and Profiling) can enhance each stage and support domain experts in making informed and interpretable decisions. The first contribution concerns the Information Access phase. A semantic search engine for biomedical literature is developed, leveraging Documentary metadata, graph-based representations and semantic analysis techniques to overcome the limitations of keyword-based retrieval. The system uncovers conceptual relationships between terms, mitigating the impact of terminological variability typical of the domain and enhancing the quality of retrieval, with direct benefits for subsequent pipeline stages. To improve data quality within biomedical datasets, the thesis investigates the interplay between Profiling Metadata Discovery and Data Imputation in incremental scenarios. Through two complementary pipelines, it shows how the progressive discovery of Relaxed Functional Dependencies (RFDs) can guide more effective imputation processes and, conversely, how incremental imputation contributes to the stability and completeness of the discovered dependencies. In addition to this, the thesis presents an RFD-based data augmentation algorithm aimed at addressing class imbalance while preserving the structural and semantic constraints of the original data, thereby further enhancing data quality before analytical tasks. Within the broader area of Meta-

data Discovery, the thesis presents an algorithm for the discovery of Property Graph Keys (*PG Keys*), a new type of structural metadata that identifies uniqueness constraints and relational patterns in graph-based representations. This contribution enhances the consistency, interpretability, and queryability of biomedical graphs, reinforcing the organization and quality of information throughout the entire pipeline. To support the final stage of Interpretation and Explainability, the thesis introduces a tool for analyzing the temporal evolution of RFDs. By integrating metadata analysis with language models, the tool provides insights into how and why dependencies change over time, facilitating model validation and the interpretation of underlying phenomena. Finally, the thesis proposes a comprehensive framework for evaluating synthetic datasets along different quality dimensions, supported by an interactive web platform that ensures interpretability and reproducibility.

ABSTRACT

La crescente digitalizzazione dei sistemi sanitari e la conseguente disponibilità di Big Data biomedici hanno trasformato il modo in cui le informazioni vengono acquisite, gestite e sfruttate in ambito clinico e di ricerca. Tuttavia, l'eterogeneità delle fonti di dati, l'incompletezza delle cartelle cliniche, la variabilità semantica della terminologia di dominio e la richiesta di risultati interpretabili pongono sfide significative alla progettazione di processi biomedici affidabili basati sui dati. Queste sfide influiscono direttamente sulla capacità degli esperti di dominio di esplorare, comprendere e fidarsi delle informazioni prodotte da tali processi. In questa tesi, l'intera pipeline biomedica basata sui dati viene esaminata end-to-end, con particolare attenzione a come diverse tipologie di metadati (documentali e di profilazione) possano migliorare ogni fase e supportare gli esperti di dominio nel prendere decisioni informate e interpretabili. Il primo contributo riguarda la fase di Accesso alle Informazioni. Viene proposto un motore di ricerca semantico per la letteratura biomedica, sfruttando metadati documentali, rappresentazioni basate su grafi e tecniche di analisi semantica per superare i limiti del recupero basato su parole chiave. Il sistema scopre relazioni concettuali tra i termini, mitigando l'impatto della variabilità terminologica tipica del dominio e migliorando la qualità del recupero, con benefici per le fasi successive della pipeline. Per migliorare la qualità dei dati all'interno di dataset biomedici, la tesi indaga l'interazione tra la scoperta di Profiling Metadata e l'imputazione dei dati in scenari incrementali. Attraverso due pipeline complementari, mostra come la scoperta progressiva di dipendenze funzionali rilassate (RFD) possa guidare processi di imputazione più efficaci e, viceversa, come l'imputazione contribuisca alla stabilità e alla completezza delle dipendenze scoperte. Inoltre, la tesi presenta un algoritmo di Data Augmentation basato su RFD, volto a correggere lo squilibrio di classe preservando al contempo i vincoli strutturali e semantici dei dati originali, migliorando ulteriormente la qualità dei dati prima delle attività analitiche.

Nell'ambito più ampio della scoperta di metadati, la tesi presenta un algoritmo per la scoperta di Property Graph Key (*PG Keys*), un nuovo tipo di metadati strutturali che identifica vincoli di unicità e pattern relazionali nelle rappresentazioni basate su grafi. Questo contributo migliora la coerenza, l'interpretabilità e la interrogabilità dei grafi biomedici, rafforzando l'organizzazione e la qualità delle informazioni lungo l'intera pipeline. A supporto della fase finale di Interpretazione e Spiegabilità, la tesi introduce uno strumento per l'analisi dell'evoluzione temporale degli RFD. Integrando l'analisi dei metadati con i modelli linguistici, lo strumento fornisce informazioni su come e perché le dipendenze cambiano nel tempo, facilitando la convalida del modello e l'interpretazione dei fenomeni sottostanti. Infine, la tesi propone un framework completo per la valutazione di set di dati sintetici lungo diverse dimensioni di qualità, supportato da una piattaforma web interattiva che garantisce interpretabilità e riproducibilità.