Exploração Semântica de Comorbidades da COVID-19: Ferramentas para predição de mortes.

Felipe Calmon, Fernanda Baião e Lucas Maddalena

Departamento de engenharia industrial, Pontifícia Universidade Católica do Rio de Janeiro, Rua Marquês de São Vicente 225, Rio de Janeiro, 22451-900, Brasil.

Abstract

A ontologia na ciência de dados é uma ferramenta que auxilia formalmente a estrutura de um sistema, na identificação das suas entidades e as suas relações entre si. Dessa forma, em um mundo globalizado, onde os dados são produzidos massivamente, entretanto, em muitas esferas sem padronizações ou em organizações que não necessariamente exprimem a real semântica da realidade, a ontologia, se apresenta como excelente alternativa para o encadeamento e gerenciamento logico dos dados. Posto isto, o seguinte trabalho teve por objetivo avaliar um conjunto de dados provenientes do SIVEP-Gripe, sob o enriquecimento semântico da Ontologia de Doenças Humanas (DO), e examinar as benesses desse artifício as mais variadas técnicas de Machine learning empregadas no processo de pesquisa. Em resumo, todo esse processo objetiva na predição de morte dos pacientes provenientes do SIVEP-Gripe e busca compreender como esse conjunto de doenças ou não, se correlatam com esses óbitos.

Keywords

Ontologia de doenças, Clustering, COVID-19

1. Introdução

O historial da COVID-19 se principia na sua descoberta em dezembro de 2019 até a sua seriação como pandemia pela Organização Mundial da Saúde (OMS) em março de 2020. No primeiro momento, identificado apenas na cidade de chinesa de Wuhan, o novo coronavírus, chamado SARS-CoV-2, rapidamente tomou proporções globais, revertendo-se em uma emergência de saúde pública internacional declarada pela OMS em janeiro de 2020, e em seguida declarando-se como uma pandemia em março do mesmo ano. Posto isto, após a afirmativa da doença, foi imperativo a coordenação nacional para combater os avanços da doença, e mitigar os riscos do vírus sobre os mais diversos tecidos sociais do globo [Fonte].

As Ontologias são eximiras aliadas da ciência de dados, visto que, proporcionam uma estrutura semântica que consente uma compreensão mais profunda e precisa dos dados ao definir relações e classificações para um domínio específico. Dessa forma, elas ofertam uma base comum de conhecimento compartilhado, e por consequência, facilitam a integração de

* Corresponding author.
† These authors contributed equally.
aleksandr.ometov@tuni.fi (A. Ometov); t.princesales@utwente.nl (T. P. Sales); manfred.jeusfeld@acm.org (M. Jeusfeld)
0000-0003-3412-1639 (A. Ometov); 0000-0002-5385-5761 (T. P. Sales); 0000-0002-9421-8566 (M. Jeusfeld)

dados de diversas fontes e a interoperabilidade entre sistemas. Ao que tange, a COVID-19, diversos trabalhos já propuseram a aplicação de ontologias, como Fonte [Fonte].

O trabalho a seguir, manipula os dados provenientes do SIVEP-Gripe sobre as comorbidades de pacientes diagnosticados com a COVID-19 no Estado do Rio de Janeiro. O propósito substancial é observar a implicação que a abordagem semântica permiti ao conseguir predizer ou não o óbito dos pacientes e por conseguinte averiguar o subconjunto de doenças que mais se assemelham a esse obituário. Para isso, após a aplicação de técnicas de clustering foi preciso separar o conjunto de dados em dois casos, o primeiro caso, sem o enriquecimento semântico, onde cada hospitalização representava um vetor binário, e após a submissão da métrica de similaridade entre os cossenos (A definir).

Ademais, no que tange ao segundo caso, o domínio de ontologia escolhido foi a DO (Ontologia de Doenças), que é um recurso elementar na integração entre vocabulários médicos, facilitando a associação das comorbidades em análise com os conceitos da DO. Para o cálculo da similaridade, foi empregado a métrica que usa os termos da DO, onde cada paciente pode apresentar mais de uma comorbidade, logo, é preciso que a avaliação da métrica seja conjunta, para que a comparação seja entre subconjuntos e não sobre termologias individuais. Deste modo, por permitir uma análise mais robusta, se consideramos hipoteticamente um conjunto de doenças como “Hipertensão essencial, Diabetes mellitus, Doença cerebrovascular, Doença renal crônica”, a métrica empregada embasa-se no cálculo de similaridades em grupo entre termos na DO, promovendo uma compreensão mais profunda e contextualizada das relações entre as doenças.

Em resumo, a abordagem semântica de forma consciente quando de encontro a subconjuntos semelhantes de hostilizações, é avaliada na etapa de Pós-Processamento de Dados usando métricas de qualidade de cluster. Prontamente, será realizada a previsão de riscos dos usuários e a observação das doenças mais prevalentes nesse processo. A investigação na parte final se inclina na busca de demonstrar a eficácia dos clusters resultantes de cada cenário na segmentação dos subconjuntos de doenças e na sua capacidade encontrar mecanismos de predição óbitos de seus usuários, pois a importância da abordagem semântica na identificação de padrões e tendências é relevante para a saúde pública.

2. Ontologia e DO (Diseiase Ontology)

A ontologia é abordada de duas maneiras distintas: do ponto de vista filosófico, envolve o estudo da natureza e estrutura da realidade; já na ciência da computação, trata-se de um instrumento que modela a estrutura de um sistema [fonte]. Nesse sentido, a Disease Ontology (DO) é uma base de conhecimento que centraliza informações sobre doenças humanas, fornecendo um guia para análises biomédicas. Seu propósito é simplificar a integração de dados genômicos e clínicos, oferecendo uma estrutura precisa e compreensível para a investigação e entendimento das doenças. Sob a liderança de Lynn Schriml, da Escola de Medicina da Universidade de Maryland, a DO visa unificar dados diversos e promover uma visão completa das enfermidades humanas [Fonte].

3. Ciclo de vida na ciência de dados e os benefícios da semântica ontológica

O ciclo de vida na ciência de dados pode ser entendido como o processo que guia projetos desde a compreensão do problema até o feedback das ações implementadas. Posto isto, é possível destacar algumas fases pelos quais se compreendem nesse ciclo [Fonte].

- **Coleta de dados:** Consiste em obter os dados corretos, considerando a complexidade do mundo real e as diferentes perspectivas dos indivíduos.
- **Pré-processamento de Dados:** Nesta etapa, os dados são analisados e manipulados para serem tratados e transformados adequadamente, preparando-os para a etapa seguinte de Mineração de Dados. Isso envolve lidar com valores ausentes, atípicos e selecionar e organizar características relevantes.
- **Mineração de Dados:** Aqui, técnicas computacionais e estatísticas são aplicadas para extrair conhecimento dos dados. Isso inclui Aprendizado de Máquina supervisionado e não supervisionado, usando algoritmos de classificação, regressão e agrupamento para descobrir padrões nos dados e prever características.
- **Pós-processamento de dados:** Após a Mineração de Dados, os resultados são interpretados, avaliando sua qualidade e considerando possíveis vieses. Os resultados são utilizados para reavaliar o modelo selecionado e a avaliação pode ser feita usando métricas como o Coeficiente de Silhueta.

O uso de ontologias é crucial na ciência de dados, oferecendo uma estrutura semântica que aprofunda a compreensão dos dados ao definir relações e classificações para um domínio específico. Na análise da COVID-19, ontologias têm sido aplicadas em diversos estudos. Por exemplo, um trabalho que utiliza dados do SIVEP-Gripe sobre comorbidades de pacientes com COVID-19 no Rio de Janeiro busca predizer óbitos e identificar doenças mais associadas. Utilizando técnicas de clustering, os dados são comparados em dois cenários: um sem enriquecimento semântico e outro com a aplicação da Ontologia de Doenças (DO). Esta última abordagem permite uma análise mais robusta, considerando grupos de doenças em conjunto, o que aprimora a compreensão das relações entre elas. Na etapa de Pós-Processamento de Dados, métricas de qualidade de cluster são utilizadas para avaliar a eficácia da abordagem semântica na segmentação de doenças e na previsão de óbitos. Essa análise é fundamental para a saúde pública, destacando a importância da abordagem semântica na identificação de padrões e tendências [Fonte].

4. Coleta e análise de dados

O SIVEP-Gripe foi criado em 2000 para monitorar o vírus da gripe no Brasil, expandindo-se em 2009 para monitorar a SRAG. Com a pandemia da COVID-19 em 2020, houve uma adaptação do sistema para incluir o novo coronavírus, visando fortalecer a vigilância de vírus respiratórios. Esta adaptação foi feita pela Secretaria de Vigilância em

Saúde do Ministério da Saúde, orientando o Sistema Nacional de Vigilância em Saúde para lidar com a circulação simultânea do SARS-CoV-2, Influenza e outros vírus respiratórios durante a emergência de saúde pública [Fonte].

5. Pré-Processamento dos dados

5.1 Tratamento da base de dados

O processamento dos dados teve início com a importação das bibliotecas essenciais, tais como pandas, numpy. O arquivo CSV contendo os dados foi então lido e atribuído ao DataFrame `srag_bruto`, com tipos de dados específicos definidos para algumas colunas. Em seguida, a coluna de datas foi convertida para o tipo datetime e o DataFrame foi filtrado para incluir apenas datas dentro de um intervalo específico. Os pacientes com evolução "Ongoing" foram removidos e uma nova coluna 'IS_DEATH' foi criada para indicar a ocorrência de óbito.

Diversas colunas foram selecionadas para compor o DataFrame final, incluindo informações como sexo, idade, comorbidades e localização geográfica. Colunas contendo valores "Yes" foram convertidas para 1 e demais valores para 0, para representar dados binários. Após esta etapa, o DataFrame estava preparado para a próxima fase, que consistiu na conversão de valores categóricos em binários, como na coluna "CS_SEXO".

Adicionalmente, foram criadas colunas adicionais para representar diferentes grupos étnicos com base na coluna "CS_RACA". Outras etapas envolveram o escalonamento dos valores da coluna "NU_IDADE_N" usando MinMaxScaler e a importação de um novo DataFrame contendo o Índice de Desenvolvimento Humano (IDH) dos municípios brasileiros.

Os valores de IDH foram ajustados e transformados em um dicionário para mapeamento, sendo então adicionados ao DataFrame principal. Ao final deste processo, o DataFrame `srag_bruto` estava limpo e pronto para análises posteriores, sendo então exportado para um novo arquivo CSV denominado "srag_tratada.csv".

5.2 Enriquecimento semântico

O processo de enriquecimento semântico dos dados teve início com a importação da base de dados tratada, que foi lida a partir de um arquivo CSV e atribuída ao DataFrame `srag_tratado`. Em seguida, foram realizadas operações de limpeza e preparação dos dados, incluindo a remoção da coluna de identificação `ID` e a seleção das colunas relevantes para o enriquecimento semântico.

Um dicionário foi empregado para renomear as colunas selecionadas, atribuindo a cada uma um identificador único correspondente ao vocabulário da Ontologia de Doenças (DOID). Essa etapa foi fundamental para estabelecer uma estrutura semântica que possibilitasse uma compreensão mais precisa dos dados.

Após a renomeação das colunas, apenas aquelas contendo informações sobre a presença ou ausência de determinadas doenças foram mantidas. Uma função foi

desenvolvida para renomear as linhas do DataFrame, substituindo os índices numéricos por uma lista ordenada das doenças presentes em cada registro.

O DataFrame resultante foi ajustado e organizado, com as linhas renomeadas como índices e as colunas selecionadas para manter apenas informações relevantes para a análise. Identificaram-se também as linhas não numéricas que precisavam ser enriquecidas com informações semânticas.

A partir de um arquivo CSV contendo medidas de similaridade entre as doenças, foram realizadas operações para substituir os valores das linhas não numéricas no DataFrame principal pelos valores correspondentes de similaridade. Esse processo foi crucial para enriquecer os dados com informações semânticas, permitindo uma análise mais aprofundada e contextualizada das relações entre as doenças.

Por fim, o DataFrame resultante do enriquecimento semântico foi combinado com um conjunto selecionado de colunas do DataFrame original, incluindo informações como sexo, idade e localização geográfica, entre outros fatores relevantes. Essa etapa concluiu o processo de tratamento e enriquecimento dos dados, preparando-os para análises posteriores.

6. Mineração dos dados

A exploração dos dados teve início com a normalização, uma prática comum para garantir uma distribuição mais homogênea e simplificar a análise. Em seguida, aplicou-se a Análise de Componentes Principais (PCA), uma técnica para reduzir a dimensionalidade dos dados mantendo sua informação essencial.

Para determinar o número ideal de componentes principais, utilizou-se o método do cotovelo, avaliando a variância explicada cumulativa em relação ao número de componentes. Isso proporcionou uma compreensão mais clara de como cada componente contribui para a variabilidade dos dados.

Posteriormente, geraram-se gráficos de dispersão dos dados reduzidos em duas dimensões, visualizando a distribuição dos registros em relação aos componentes principais. Essa visualização facilitou a identificação de padrões e agrupamentos nos dados.

Para avaliar a qualidade dos agrupamentos formados, foram testados diferentes métodos, como o KMeans e o GMM, utilizando métricas como BIC, silhueta e índice Davies-Bouldin. Os resultados foram registrados em um DataFrame, permitindo uma comparação objetiva entre os métodos e números de agrupamentos testados.

Além disso, foram gerados gráficos para visualizar os agrupamentos formados pelo KMeans e pelo GMM, oferecendo insights visuais sobre a estrutura dos dados.

References

[Em construção]