Análise não supervisionada de K-means de dados de tuberculose no Brasil

Autores: Angelo Rossini[1], Newton Shydeo Brandão Miyoshi[2]

Colaboradores: Domingos Alves[3], Vitor Cassão[4]

[1,2]Centro Universitário Barão de Mauá

[3,4]Escola Médica de Ribeirão Preto, Universidade de São Paulo

[1]angelorossini96@gmail.com - Ciência da Computação, [2]newton.miyoshi@baraodemaua.br

Resumo

Este artigo tem como objetivo demonstrar os resultados obtidos por meio da análise e aplicação de um algoritmo não supervisionado *K-means* na base de dados do SINAN de 2001 a 2022 no Brasil, com o propósito de compreender quais estados têm o maior número de casos de tuberculose e identificar semelhanças entre eles que podem contribuir para uma taxa de casos mais elevada em relação à população local. Iniciaremos com uma breve introdução histórica, seguida de um panorama das características relacionadas à transmissão da tuberculose. Posteriormente, discutiremos os resultados obtidos a partir da análise ano a ano dos dados coletados. Conforme destacou Barbosa et al. (2013), a análise de dados epidemiológicos pode oferecer *insights* valiosos sobre a distribuição e evolução da tuberculose em diferentes regiões. Ao aplicar técnicas de agrupamento, como o algoritmo *K-means*, é possível identificar padrões e semelhanças entre os estados, contribuindo para uma compreensão mais aprofundada dos fatores que influenciam a incidência da doença.

Introdução

A tuberculose continua sendo uma das doenças infecciosas mais mortais do mundo. Todos os dias, mais de 4.000 pessoas morrem de tuberculose em todo o mundo e aproximadamente 30.000 ficam doentes com esta doença evitável e curável. Nas Américas, mais de 70 pessoas morrem e cerca de 800 adoecem com tuberculose todos os dias. Estima-se que em 2020 havia 18,3 mil crianças com TB nas Américas, metade delas com menos de cinco anos. Globalmente, no mesmo ano, 7,0 milhões de casos de tuberculose pulmonar confirmados bacteriologicamente tiveram 3,9 milhões (55%) de contactos avaliados tanto para infecção como para doença, conforme afirmado em [12]. Os anos de 2020 e 2021 também foram impactados pela COVID-19, que afetou a prestação e o acesso aos serviços essenciais de tuberculose (TB).

Objetivos

O estudo tem como objetivo identificar padrões de evolução temporal dos casos de tuberculose por meio da análise e desenvolvimento de um algoritmo de agrupamento utilizando séries temporais com base em dados fornecidos pelo SINAN para casos de tuberculose no Brasil. As etapas foram divididas da seguinte forma: definição do problema e busca de fontes de dados confiáveis relacionadas à tuberculose no Brasil; tratamento dos dados coletados, como limpeza de dados, normalização, tratamento de valores faltantes e integração de dados; transformação dos dados em um formato adequado para entrada do modelo de aprendizado de máquina; interpretação dos resultados obtidos do aprendizado de máquina por meio de análise de *cluster* baseada em séries temporais, permitindo avanços futuros para potencialmente projetar o modelo como uma ferramenta de tomada de decisão sobre os possíveis caminhos de evolução da tuberculose regionalmente.

Materiais e métodos

Descoberta de conhecimento no fluxo de trabalho de pesquisa e análise de banco de dados

Este estudo foi desenvolvido com base no fluxo de trabalho *Knowledge Discovery in Databases (KDD)* aplicado à análise de dados de tuberculose no Brasil. Seguindo essa metodologia, as etapas foram divididas na seguinte ordem: (i) definição do problema e busca de fontes de dados confiáveis relacionadas à tuberculose no Brasil; (ii) tratamento dos dados coletados, como limpeza de dados, normalização, tratamento de valores faltantes e integração de dados; (iii) transformar os dados em um formato adequado para entrada do modelo de aprendizado de máquina; (iv) interpretar os resultados obtidos do aprendizado de máquina por meio de análise de *cluster* baseada em séries temporais, permitindo avanços futuros para potencialmente projetar o modelo como uma ferramenta de tomada de decisão sobre os possíveis caminhos de evolução da tuberculose regionalmente.

Conjunto de dados

Todos os dados utilizados neste trabalho foram obtidos do *DATASUS* por meio da biblioteca *PySUS*, que permite uma coleta de dados mais eficiente no formato *CSV (Comma-Separated Values)*. Os dados coletados em questão são originários do SINAN (Sistema de Informação de Agravos de Notificação), que tem como objetivo coletar, transmitir e divulgar dados gerados rotineiramente pelo Sistema de Vigilância Epidemiológica. É importante ressaltar que todos os dados são abertos e possuem total anonimidade das pessoas cujas informações foram coletadas, sendo formatados conforme as diretrizes estabelecidas pela Lei Geral de Proteção de Dados (LGPD), garantindo a ética do uso dos mesmos. As notificações recebidas pelo SINAN entre os anos de 2001 e 2022 são preenchidas pelas unidades de saúde para cada paciente quando há suspeita de problema de saúde notificável de interesse nacional, estadual ou municipal. Embora raro, existe a possibilidade de um mesmo paciente gerar mais de uma notificação. Há várias informações disponíveis nos arquivos *CSV* coletados, representando diferentes conjuntos de dados. Para a análise da evolução da tuberculose entre os estados ao longo de meses e anos, foram selecionadas apenas as informações contendo dados sobre os estados brasileiros, inicialmente desconsiderando os municípios. Este repositório recebe atualizações diárias desde 2001. Os dados filtrados para análise incluem os seguintes indicadores: data, estado, total de casos por ano e mês, e total de casos por 1.000 habitantes. Neste estudo, foi utilizada a série temporal do número de óbitos por 1.000 habitantes notificados em 11 anos para cada estado brasileiro.

Análise Não Supervisionada

A escolha dos algoritmos não supervisionados foi feita devido à sua capacidade de descobrir padrões e formar grupos a partir de um conjunto de dados sem informações prévias. Existem vários tipos de algoritmos, cada um com sua aplicação específica, que podem ter melhor ou pior desempenho dependendo do caso de uso. No entanto, em geral, são frequentemente aplicados para reconhecimento de padrões [8], tornando-os uma ferramenta útil neste contexto. Além disso, esses algoritmos permitem a disseminação de informações por meio de visualização *web*, como mapas dinâmicos ou *APIs*.

Algoritmos de *clustering* referem-se a uma ampla gama de técnicas usadas para encontrar subgrupos ou *clusters* dentro de um conjunto de dados. Cada *cluster* identificado é composto por observações (ou instâncias) que são altamente semelhantes com base em alguma métrica pré-definida (como a distância euclidiana). A técnica escolhida neste estudo é *K-means*, que envolve particionar N elementos em K grupos ou *clusters*, onde N é o número total de elementos sendo analisados, e K é o número total de *clusters*. O valor de "K" pode ser determinado pela experiência do usuário ou utilizando técnicas disponíveis para calcular um valor possível, como o método do cotovelo ou o método da silhueta. A validação de estruturas de *cluster* é muitas vezes a parte mais desafiadora e frustrante da análise de *cluster*, o que, como enfatizado por [8], pode fazer com que a análise de *cluster* pareça uma caixa preta.

K-means inicializa um valor inicial para o ponto central do *cluster*, também conhecido como medoide, geralmente selecionando aleatoriamente um valor dentro do intervalo de dados. O algoritmo atribui cada ponto de dados ao medoide mais próximo com base em uma métrica de distância, normalmente a distância euclidiana, e então recalcula o medoide considerando os novos dados associados, até sua convergência (onde não há mudanças significativas entre os elementos dentro dos *clusters*) ou as iterações do algoritmo terminam.

Tecnologias

Todo o código desenvolvido foi criado na linguagem de programação *Python*. Para visualizações gráficas, foi utilizada a biblioteca *Matplotlib*. *Scikit-learn* foi usado para algoritmos básicos de aprendizado de máquina, como *K-means*. *Pandas* e *NumPy* foram empregados para limpeza e transformações de dados.

Pré-processamento de dados

O conjunto de séries temporais foi criado a partir dos registros de casos por 1.000 habitantes em um período de 11 anos, com os dados segmentados pelo número total de casos por mês para cada ano, a partir do primeiro registro do SINAN em janeiro de 2001.

Clustering de Séries Temporais

A etapa de agrupamento teve o valor de K escolhido como forma de mapear a propagação da doença com base no conjunto de dados agrupados em 3 grupos de controle (k = 3), visando agrupar a resposta à progressão da pandemia no Brasil em 3 níveis relativos: ruim, moderado e bom.

Resultados

Ao analisar a apresentação dos resultados do agrupamento da série temporal com k=3 foi notado que as notificações por ano a cada mil habitantes em relação aos respectivos anos, demonstrou que os estados do Acre, Amazonas, Roraima, Pernambuco e Rio de Janeiro como os estados com pior desenvolvimento de casos quando comparados com a média do seu respectivo *cluster*.

Anais, v. 9, 2024 – ISSN 2594-3723 – DOI: 10.56344/enic2024-7

Durante a análise, foi notado que as médias para os grupos 1 e 0 são semelhantes, e que a maioria dos estados teve sua evolução abaixo da média no grupo 1, aproximando-se ou superando-a apenas no início e no final. Por outro lado, os estados em do grupo 0 exibiram uma distribuição um pouco mais ampla, com três estados ultrapassando a média, e a maioria permanecendo próxima, mas abaixo dela, caracterizando-se como os casos melhor e moderado, respectivamente. Além disso, no grupo 2, onde se encontram os piores casos, Amazonas e Rio de Janeiro ficaram consideravelmente acima da média, sendo classificados como os estados com os piores casos de tuberculose do Brasil.

Discussão

No presente estudo, buscamos identificar como os estados brasileiros têm lidado com a tuberculose ao longo dos anos por meio da análise de agrupamento utilizando o número de casos por 1.000 habitantes. Optamos por essa métrica por sua confiabilidade em relação a outras medidas disponíveis, que podem sofrer com um maior grau de subnotificação. Utilizar uma métrica que considera a população permite comparações entre diferentes estados, sendo uma escolha criteriosa. Caso tivéssemos utilizado relatórios de notificação absoluta, os estados de São Paulo e Rio de Janeiro teriam formado um *cluster* isolado devido aos maiores números absolutos de casos. Na análise realizada com k=3, observamos que as principais diferenças entre os *clusters* residem no número total de notificações por 1.000 habitantes. Os estados que apresentaram maior número de casos foram Acre, Amazonas, Roraima, Pernambuco e Rio de Janeiro. Também notamos uma maior densidade entre 20 e 60, com uma ligeira diminuição à medida que ambos os lados se aproximavam de 40, sugerindo que 30 a 50 casos por 1.000 habitantes é uma média comum de ser encontrada.

Com base na comparação entre as regiões sudeste e norte, conforme uma nota técnica do IEPS (Instituto de Estudos de Políticas de Saúde) [10], entre os anos de 2010 e 2020, os estados das regiões Norte e Nordeste apresentaram os piores indicadores de saúde em 14 indicadores, enquanto os das regiões Sul e Sudeste apresentaram, em sua maioria, os melhores indicadores de recursos e mortalidade. Isso pode explicar por que a região Norte apresenta um maior número de casos notificados, e entre as Unidades da Federação, o Amazonas demonstrou ter um número superior ao de estados com maior densidade populacional, como Rio de Janeiro e São Paulo.

Ao analisar a distribuição do número total de casos de tuberculose ao longo dos anos, observamos uma clara melhoria de 2005 a 2016, atribuída às campanhas de sensibilização, à implementação de uma rede de diagnóstico rápido de doenças em 2014 e à descentralização do tratamento para a Atenção Primária durante este período. Houve também um ligeiro aumento nas notificações de 2018 para 2019, atribuído à redução da cobertura vacinal contra tuberculose. Até 2018, a taxa de vacinação manteve-se acima dos 95%, enquanto em 2019 a cobertura não ultrapassou os 88%. O aumento a partir de 2020 pode dever-se potencialmente a erros de diagnóstico ou às consequências da pandemia.

Outro fator importante a considerar é a qualidade dos dados das notificações do SINAN. Sendo um sistema de âmbito nacional, o SINAN-TB pode ser utilizado como um sistema orientado burocraticamente e algumas limitações são destacadas por Rocha et al. [11], como relatos incompletos e inconsistentes de variáveis-chave, como detalhes demográficos (escolaridade, raça/cor da pele), condições associadas (AIDS, alcoolismo, diabetes) e informações de acompanhamento do tratamento. Além disso, dados essenciais relacionados com populações especiais e beneficiários de programas governamentais são muitas vezes captados de forma inadequada. A presença de registros não fechados ou fechados de forma imprecisa, juntamente com a falta de atualizações atempadas dos testes laboratoriais em curso, dificulta ainda mais a representação e avaliação precisas das intervenções realizadas. O SINAN-TB não possui um identificador único de pessoa e é difícil promover a integração perfeita com outros sistemas. É necessário um esforço para mudar sua arquitetura e adotar tecnologias modernas para agilizar a transferência e análise de dados no contexto do controle da TB no Brasil.

Conclusão

O agrupamento de séries temporais tem se mostrado uma ferramenta importante para analisar dados e encontrar semelhanças comportamentais na progressão da tuberculose entre os estados brasileiros. Inicialmente, escolheu-se como parâmetro de análise o número de notificações por 1.000 habitantes. O número de *clusters* foi definido como 3, em uma tentativa de categorizar a progressão da doença em 3 grupos de controle.

Apesar dessas constatações, considerando um país de dimensões continentais como o Brasil, ainda é necessária uma análise mais aprofundada. É preciso identificar as razões das diferenças na progressão da doença entre os estados brasileiros. Como trabalho futuro, serão incluídos parâmetros adicionais no processo de

agrupamento para considerar outras informações sobre os estados, como feito em referências anteriores. Alguns estados possuem características específicas que podem não ser totalmente captadas por um parâmetro que considere apenas a população residente total. Estados como Minas Gerais e Amazonas possuem vastas áreas verdes, necessitando de uma análise mais criteriosa.

Outra melhoria que pode ser feita em trabalhos futuros é automatizar o processo de encontrar o número ideal de *clusters* através de uma abordagem de *clustering* hierárquico. Seria interessante comparar a hierarquia entre os estados e extrair dela características significativas. Além disso, a utilização de métricas que consideram toda a série temporal durante o agrupamento, destacando variações na velocidade de propagação da doença e outros fatores, como o algoritmo de agrupamento *Dynamic Time Warping (DTW)*, é outro ponto de melhoria.

A integração com outros conjuntos de dados dos estados tem um imenso potencial para aumentar a profundidade e a eficácia da análise da TB. Ao fundir dados clínicos, demográficos, socioeconômicos e geográficos, podemos alcançar uma visão holística da incidência, prevalência e fatores de risco associados da TB. Esta abordagem abrangente ajuda a identificar padrões complexos e potenciais determinantes da TB, permitindo intervenções e políticas personalizadas que abordam os desafios únicos enfrentados pelas diferentes regiões.

Os resultados obtidos podem auxiliar na identificação de padrões e na avaliação do sucesso das ações e estratégias adotadas pelos estados brasileiros no combate à tuberculose. Esta análise pode fornecer insights para orientar ações futuras e determinar as melhores respostas para casos semelhantes.

Referências

[1] **BARBOSA, Isabelle Ribeiro et al**. Análise da distribuição espacial da tuberculose na região Nordeste do Brasil, 2005-2010. **Epidemiologia e Serviços de Saúde**, Brasília, v. 22, n. 4, p. 687-695, dez. 2013. Disponível em: http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742013000400015&lng=pt&nrm=iso. Acesso em: 27 mar. 2023. http://dx.doi.org/10.5123/S1679-49742013000400015.

[2] **BATISTA, Gustavo Enrique de Almeida Prado Alves**. Pré-processamento de dados em aprendizado de máquina supervisionado. 2003. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2003. doi:10.11606/T.55.2003.tde-06102003-160219. Acesso em: 27 mar. 2023.

[3] **BIERRENBACH, A. L. et al.** Incidência de tuberculose e taxa de cura, Brasil, 2000 a 2004. **Revista de Saúde Pública**, v. 41, p. 24–33, 1 set. 2007.

[4] **CAMILO, Cássio Oliveira; SILVA, João Carlos da.** Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFG)**, v. 1, n. 1, p. 1-29, 2009.

[5] **CASSÃO, Victor et al.** Unsupervised analysis of COVID-19 pandemic evolution in brazilian states. **Procedia Computer Science**, v. 196, p. 655-662, 2022.

[6] **GARETH, James; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert.** An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

[7] **Incidência da tuberculose cai 20,2% no Brasil em uma década.** Disponível em: https://www.unasus.gov.br/noticia/incidencia-da-tuberculose-cai-202-no-brasil-em-uma-decada. Acesso em: 14 set. 2023.

[8] **JAIN, Anil K.; DUBES, Richard C.** Algorithms for clustering data. Prentice-Hall, Inc., USA, 1988.

[9] **Ministério da Saúde lança campanha de combate à tuberculose e reforça ações para eliminação da doença no Brasil.** Disponível em: https://www.gov.br/saude/pt-br/assuntos/noticias/2023/marco/ministerio-da-saude-lanca-campanha-de-combate-a-tuberculose-e-reforca-acoes-para-eliminacao-da-doenca-no-brasil#:~:text=Uma%20das%20principais%20formas%20de. Acesso em: 14 set. 2023.

[10] **RACHE, Beatriz et al.** A Saúde dos Estados em Perspectiva Comparada: Uma Análise dos Indicadores Estaduais do Portal IEPS Data. **Instituto de Estudos para Políticas de Saúde**, 2022.

[11] **ROCHA, M. S. et al.** Sistema de Informação de Agravos de Notificação (Sinan): principais características da notificação e da análise de dados relacionada à tuberculose. **Epidemiologia e Serviços de Saúde**, v. 29, n. 1, mar. 2020.

[12] **SILVA, S.** Implicações dos Fulanos na Global Tuberculosis Report 2021. Disponível em:

https://www.who.int/publications/digital/global-tuberculosis-report-2021.

[13] **SPOLAÔR, Newton et al.** Um estudo da aplicação de clustering de séries temporais em dados médicos. In: **III Congresso da Academia Trinacional de Ciências**, 2008. p. 1-10.