REDES BAYESIANAS DE APRENDIZADO: ESTUDO APLICADO AOS DADOS DE *DIABETES MELLITUS TIPO I* NO ESTADO DO PARANÁ, BRASIL

Ian Fraga Bitar
ian-fraga@hotmail.com
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Elisangela Aparecida da Silva Lizzi
elisangelalizzi@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

RESUMO

O Diabetes Mellitus configura-se como uma epidemia mundial, traduzindo- se em um grande desafio para os sistemas de saúde de todo o mundo. O envelhecimento da população, a urbanização crescente e a adoção de estilos de vida pouco saudáveis como sedentarismo, dieta inadequada e obesidade são os grandes responsáveis pelo aumento da incidência e prevalência do diabetes. OBJETIVO: Estudar a aplicabilidade de redes bayesianas de aprendizado para dados de populacionais de diabetes mellitus tipo I no estado do Paraná no nível municipal. MÉTODOS: Para isto foram utilizados quatro algoritmos, para gerar as redes bayesianas, sendo eles: *Grow-Shrink*, *Incremental Association Markov Blanket*, *Max-Min Hill-Climbing* e *Hill-Climbing*. Sendo utilizados para análise e entendimento das interações entre as variáveis monstrando estes achados por meio de grafos acíclicos dirigidos. RESULTADOS E CONCLUSÃO: Estes algoritmos tiveram boa performance para entender a estrutura condicional das informações relativas a diabetes mellitus do tipo 1 e conduziram a resultados interessantes do ponto vista epidemiológico das relações estruturadas. Vale salientar que ao se ponderar os princípios de teoria dos grafos, teoria das probabilidades, otimização e ciência da computação o melhor algoritmo é o MMHC, pois obteve-se resultados análogos em termos da estrutura das redes bayesianas geradas e este método é mais robusto pois sua implementação é híbrida. Os métodos estudados e aplicados são úteis para entender a estrutura condicional probabilística entre as variáveis de interesse e podem ser utilizados como um indicador de políticas públicas locais para o controle e intervenções da diabetes mellitus no nível populacional.

PALAVRAS-CHAVE: Redes Bayesianas. Diabetes mellitus. Estatística aplicada. Inteligência artificial.

1. INTRODUÇÃO

O *Diabetes Mellitus* (DM) configura-se como uma epidemia mundial, representando um desafio para os sistemas de saúde. O envelhecimento da população, a urbanização crescente e a adoção de estilos de vida pouco saudáveis como sedentarismo, dieta inadequada e obesidade são os grandes responsáveis pelo aumento da incidência e prevalência do diabetes em todo o mundo (Brasil, 2006).

O diabetes é uma doença crônica que ocorre quando o pâncreas não produz insulina suficiente ou quando o corpo não pode efetivamente usar a insulina que produz, caracterizada como uma doença metabólica. A insulina é um hormônio que regula o açúcar no sangue. A hiperglicemia ou o aumento do nível de açúcar no sangue é um efeito comum da diabetes não controlada e, ao longo do tempo, causa sérios danos a diversos sistemas do corpo, especialmente os nervos e os vasos sanguíneos (WHO,2016). A *diabetes mellitus* tipo 1 é caracterizada por uma produção insuficiente de insulina e requer administração diária de insulina (WHO, 2016)

Mundialmente, estima-se que 422 milhões de adultos contraíram diabetes em 2014, comparado com 108 milhões em 1980. A taxa de incidência padronizada pela idade dobrou desde 1980, aumentando de 4,7% para 8,5% na população adulta. Responsável pela morte de 1,5 milhões em 2012, quarenta e três por cento destes 1,5 milhões de mortes ocorreram em pessoas inferior a 70 anos (WHO, 2016).

No âmbito da saúde pública a diabetes apresenta alta morbimortalidade, com perda importante na qualidade de vida. É uma das principais causas de mortalidade, insuficiência renal, amputação de membros inferiores, cegueira e doença cardiovascular. O diabetes representa carga adicional à sociedade, em decorrência da perda de produtividade no trabalho, aposentadoria precoce e mortalidade prematura (Brasil, 2006)

O intuito deste estudo é entender as interações entre várias variáveis de interesse sobre diabetes mellitus tipo I, trabalhando-se com dados do estado do Paraná. Foram utilizados vários algoritmos computacionais já implementados de redes bayesianas com o objetivo de entender as interações probabilísticas entre essas variáveis no nível municipal. Redes Bayesianas (BNs) são modelos gráficos, o que significa que eles contêm uma parte que pode ser representada como um gráfico, mais especificamente Grafos Acíclicos Dirigidos (DAG's). A motivação deste trabalho nesta classe de modelos está no interesse em entender o fenômeno e em derivar relacionamentos causa-efeito a partir de dados com a capacidade de representar relações causais baseando-se em funções de distribuições de probabilidade conjunta (Margaritis, 2003).

2. OBJETIVO

Analisar a estrutura causal utilizando redes bayesianas de aprendizado para entender as interações das variáveis disponíveis sobre *Diabete Mellitus Tipo 1* no nível populacional, mais especificamente relativos ao município de residência dos dados notificados no sistema TABNET-DATASUS para o estado do Paraná, no período de 2008 a 2012 (dados acumulados).

3. MÉTODOS

A estrutura de uma rede bayesiana representa um conjunto de relações probabilísticas condicionais que se mantêm no domínio do fenômeno em estudo. A resposta da rede bayesiana revela a estrutura causal subjacente sobre o diabete mellitus tipo 1 e como as variáveis estão relacionadas entre si, utilizando DAG's. O desenho do estudo é do tipo epidemiológico ecológico, onde a população em estudo é o estado do Paraná-Brasil.

Os dados obtidos foram organizados como banco de dados para *input* no programa que realiza as análises, sendo necessário aplicar alguns filtros aplicados como: informações duplicadas, valores perdidos/ausentes, validação do *link* dos municípios com as variáveis de interesse tabeladas, formando assim uma base final para análise. O banco de dados final foi formatado em estrutura matricial e consolidado para as múltiplas variáveis de interesse, sendo elas: Tabagismo, Sobrepeso, Sexo, Sedentarismo, Pé Diabético, Infarto Agudo do Miocárdio e Amputação por pé diabético para *Diabetes Mellitus* tipo 1. Essas informações representam um panorama instantâneo estático do fenômeno em um momento especifico. Por esta razão utilizaremos algoritmos de redes bayesianas estáticas, estes fornecem uma estrutura intuitiva e abrangente para modelar as dependências entre as variáveis, pois parte-se da ideia central de aprendizado da estrutura de relações causais a partir dos dados. As implementações computacionais foram realizadas com o auxílio do software R (Versão 3.2) e para a construção das redes bayesianas foi utilizado a biblioteca externa *bnlearn* (Scutari,2010; Scutari, 2017) com todos os algoritmos disponibilizados.

3.1. Algoritmos utilizados

Foram utilizados quatro algoritmos distintos para estudar a estrutura de dependência condicional para os dados de DM segundo as variáveis de interesse. Sendo eles: *Grow-Shrink, Incremental Association Markov Blanket, Hill Climbing* e *Max-Min Hill Climbing Restricted*.

Os algoritmos *Grow-Shrink* (GS) e *Incremental Association Markov Blanket(IAMB)*, pertencem a categoria de algoritmos com restrições na estrutura condicional, que se baseiam em técnicas de aprendizado da estrutura dos dados, gerando e analisando a relação probabilística implicada pela propriedade de probabilidade e independência condicional. Depois de estruturada as relações é possível construir um grafo que demonstre os caminhos possíveis e as inter-relações entre as variáveis. O algoritmo *Hill Climbing (HC)* pertence a categoria de algoritmos baseados em pontuações que atribuem um escore para cada rede bayesiana candidata, e depois maximiza-lo com algumas heurísticas baseadas em algoritmos gulosos. Por último, o algoritmo *Max-Min Hill Climbing Restricted (MMHC)* que pode ser classificado como algoritmo de aprendizado com estrutura híbrida, que mescla as funcionalidades de relação probabilística/independência condicional com algoritmos baseados em pontuação.

4. RESULTADOS E DISCUSSÃO

Abaixo segue Tabela 1 com a legenda das variáveis utilizadas nas redes bayesianas geradas.

Tabela 1- Legenda das variáveis de interesse utilizadas no processo de modelagem das redes bayesianas.

V1 – Hipertensão c/ Diabete Tipo 1 por Tabagismo	V12 – Hipertensão por Sexo Feminino
V2 – Hipertensão c/ Diabete Tipo 1 por Sobrepeso	V13 – Hipertensão por Sedentarismo
V3 – Hipertensão c/ Diabete Tipo 1 por Sexo Masculino	V14 – Hipertensão por Infarto Agudo do Miocárdio
V4 – Hipertensão c/ Diabete Tipo 1 por Sexo Feminino	V15 – Diabete Tipo 1 por Tabagismo
V5 – Hipertensão c/ Diabete Tipo 1 por Sedentarismo	V16 – Diabete Tipo 1 por Sobrepeso
V6 – Hipertensão c/ Diabete Tipo 1 por Pé Diabético	V17 – Diabete Tipo 1 por Sexo Masculino
V7 – Hipertensão c/ Diabete Tipo 1 por Infarto Agudo do Miocárdio	V18 – Diabete Tipo 1 por Sexo Feminino
V8 – Hipertensão c/ Diabete Tipo 1 por Amputação Pé Diabete	V19 – Diabete Tipo 1 por Sedentarismo
V9 – Hipertensão por Tabagismo	V20 – Diabete Tipo 1 por Pé Diabético
V10 – Hipertensão por Sobrepeso	V21 – Diabete Tipo 1 por Infarto Agudo do Miocárdio
V11 – Hipertensão por Sexo Masculino	

Fonte: Autoria Própria (2017)

Segundo os resultados expostos no quadro 1, com da estrutura das redes dos 4 algoritmos testados é possível fazer algumas comparações: o algoritmo HC

gerou a rede bayesiana com o maior número de arcos diretos contabilizando 98. Os demais algoritmos utilizados minimizaram esta estrutura, e por consequência resultaram em uma quantidade menor de arcos diretos. O HC e o MMHC foram otimizados, pois ambos trabalham com critérios de pontuação e devem ser otimizados para gerar o melhor caminho e não geraram arcos indiretos. É interessante ponderar os resultados das redes bayesianas obtidas pelos algoritmos GC, IAMB e MMHC pois conseguiram reduzir a dimensão da estrutura dos dados e tem comportamentos semelhantes neste caso.

Quadro1: Demonstrativo com os resultados obtidos das respectivas redes bayesianas geradas para os 4 algoritmos implementados

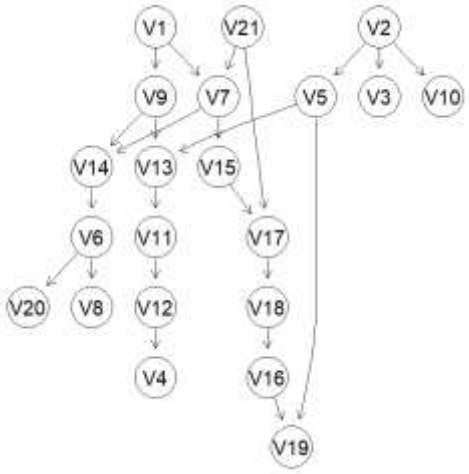
Algoritmos	Arcos Indiretos	Arcos Diretos
GS	7	16
IAMB	5	23
HC	0	98
MMHC	0	23

Fonte: Autoria própria (2017)

Os resultados dos algoritmos GS, IAMB e MMHC retornaram algumas relações de causa e efeito similares, por exemplo: infarto agudo do miocárdio está diretamente relacionado a hipertensão, diabetes e tabagismo, sendo a causa direta deste fenômeno no sexo masculino. Por sua vez, o sobrepeso está associado a diabetes e sedentarismo no sexo feminino. Com relação a diabetes, pé diabético, hipertensão e tabagismo exercem relação direta no acometimento de infarto agudo do miocárdio, independente do sexo.

Estes algoritmos tiveram boa performance para entender a estrutura condicional das informações relativas a *diabetes mellitus do tipo 1* e conduziram a resultados interessantes do ponto vista epidemiológico das relações estruturadas. Vale salientar que ao se ponderar os princípios de teoria dos grafos, teoria das probabilidades, otimização e ciência da computação o melhor algoritmo seria o MMHC, pois obteve-se resultados análogos em termos da estrutura das redes bayesianas geradas e este método é mais robusto pois sua implementação é híbrida. Logo abaixo segue a figura 1 com o grafo gerado pelo o algoritmo MMHC, neste trabalho escolhido como o melhor candidato para gerar a solução pretendida.

Figura 1 – Grafo da rede bayesiana obtida pelo algoritmo Max-Min Hill Climbing Restricted



Fonte: Autoria própria (2017)

5. CONCLUSÃO

Métodos de análise baseado em distribuições de probabilidade são úteis para entender fenômenos no nível populacional. As redes bayesianas de aprendizado aplicadas ao estudo da diabetes mellitus tipo 1, promoveu o entendimento da estrutura probabilística condicional entre as variáveis de interesse e podem ser utilizados como um indicador de políticas públicas locais para controle e intervenções da *diabetes mellitus* no nível municipal.

LEARNING BAYESIAN NETWORKS: APPLIED STUDY TO *DIABETES MELLITUS TYPE I* DATA IN THE STATE OF PARANA, BRAZIL.

ABSTRACT

Diabetes Mellitus is a worldwide epidemic, posing a major challenge for health systems around the world. Population aging, growing urbanization and the adoption of unhealthy lifestyles, such as sedentary lifestyle, inadequate diet and obesity are largely responsible for the increased incidence and prevalence of diabetes. OBJECTIVE: To study the applicability of learning Bayesian networks from population data of type I diabetes mellitus in the state of Parana at the municipal level. METHODS: Four algorithms were used to generate Bayesian networks: Grow-Shrink, Incremental Association Markov Blanket, Max-Min Hill-Climbing and Hill-Climbing. Being used to analyze and understand the interactions among the variables, showing these findings by means of directed acyclic graphs. RESULTS AND CONCLUSIONS: These algorithms performed well to understand the conditional structure of information regarding type 1 diabetes mellitus and led to interesting results from the epidemiological point of view of structured relationships. It is worth mentioning that when considering the principles of graph theory, probability theory, optimization and computer science, the best algorithm is the MMHC, since similar results were obtained in terms of the structure of the Bayesian networks generated and this method is more robust since it is a hybrid implementation. The methods studied and applied are useful to understand the probabilistic conditional structure among the variables of interest and can be used as an indicator of local public policies for the control and interventions of diabetes mellitus at the population level.

KEYWORDS: Bayesian Networks. Diabetes mellitus. Applied statistics. Artificial intelligence.

REFERÊNCIAS

BRASIL. Cadernos de Atenção Básica: Diabetes Mellitus .Secretaria de Atenção à Saúde. De-partamento de Atenção Básica. Ministério da Saúde, Secretaria de Atenção à Saúde, Departa-mento de Atenção Básica. Brasília : Ministério da Saúde, 2006.

EDERA, ALEJANDRO; STRAPPA, YANELA; BROMBERG, FACUNDO. The Grow-Shrink strategy for learning Markov network structures constrained by countext-specific inde-pendences. Departamento de Sistemas de Información, Universidade Tecnológica Nacional.

MARGARITIS, D. Learning Bayesian Network Model Structure from Data. Thesis of School of Computer Science Carnegie Mellon University Pittsburgh, 2003.

NAGARAJAN, D; SCUTARI, M; LÈBRE, S. Bayesian Networks in R: with Application in Systems Biology.

SCUTARI M. Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimized Implementations in the bnlearn R Package. Journal of Statistical Software, 77(2), 1-20:2017.

SCUTARI M. Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35(3), 1-22:2010. URL http://www.jstatsoft.org/v35/i03/.

WORLD HEALTH ORGANIZATION. Global report on diabetes. Diabetes Mellitus – epide-miology. 2. Diabetes Mellitus – prevention and control. 3. Diabetes, Gestational. 4. Chronic Disease. 5. Public Health. I. World Health Organization, 2016.

Recebido: 31 ago. 2017.
Aprovado: 02 out. 2017.
Como citar:
BITAR, I. F. e LIZZI, E.A.S. Redes bayesianas de aprendizado: estudo aplicado aos dados de diabetes mellitus tipo I no estado do Paraná, Brasil. In: SEMINÁRIO DE INICIAÇÃO CIENTÍFICA E TECNOLÓGICA DA UTFPR, 22., 2017, Londrina. **Anais eletrônicos...** Londrina: UTFPR, 2017. Disponível em: <https://eventos.utfpr.edu.br//sicite/sicite2017/index. Acesso em: XXX.
Correspondência:
Ian Fraga Bitar
Rua dos bandeirantes, número 180, Bairro centro, Cornélio Procópio, Paraná, Brasil.