



Utilização de técnicas computacionais de mineração de dados para auxiliar no diagnóstico de diabetes *mellitus* tipo 2

Matheus Maciel¹ (PIBITI/CNPq/Unioeste), Wu Feng Chung^{1 2} (Orientador), Huei Diana Lee^{1 2}, Newton Spolaôr¹, e-mail: matheusm1999@live.com.

¹Universidade Estadual do Oeste do Paraná/Centro de Engenharia e Ciências Exatas/Foz do Iguaçu, PR.

²Universidade Estadual de Campinas/Faculdade de Ciências Médicas/Campinas, SP.

Área/subárea: Ciências Exatas e da Terra/Ciência da Computação

Palavras-chave: Diabetes melito, Informática em saúde, Tomada de decisões

Resumo

A diabetes *mellitus* é uma das principais doenças crônicas não transmissíveis. Dentre os tipos dessa doença, destaca-se a do tipo 2 como o mais comum em indivíduos adultos acima de 40 anos e está intimamente relacionada com sedentarismo e obesidade. Ainda, em diversos casos, o diagnóstico é realizado somente depois do surgimento de complicações graves e, decorrente à complexidade da enfermidade, processos computacionais inteligentes têm sido aplicados para apoiar profissionais da saúde no diagnóstico dessa doença. Assim, este trabalho avaliou o uso de Mineração de Dados para auxiliar no diagnóstico da diabetes *mellitus* tipo 2. Para tanto, foi escolhida a base de dados *Pima Indians Diabetes*, constituída por oito atributos e 768 indivíduos. As etapas de pré-processamento, extração de padrões e pós-processamento foram instanciadas com os algoritmos ReliefF e J48 e a estratégia de validação cruzada. O J48 foi aplicado para gerar uma árvore de decisão com 50% dos melhores atributos selecionados pelo ReliefF e outra com todos os atributos. Classificados os casos com e sem diabetes *mellitus* e validada em 10 *folds*, foi realizada análise estatística para comparar as médias de erro dos modelos gerados ($p\text{-valor} > 0,05$). Apesar da ausência de significância estatística, o modelo gerado a partir dos atributos selecionados se destacou pela redução da dimensionalidade dos dados e pela menor complexidade da árvore correspondente.

Introdução

A diabetes *mellitus* (DM) é uma das principais doenças crônicas não transmissíveis, e afeta cerca de 463 milhões de indivíduos no mundo. Esta doença origina elevado número de mortes e promove, em média, a cada oito segundos, o

falecimento de um indivíduo (International Diabetes Federation, 2019). Dentre os tipos de DM, o do tipo 2 (DM2) é o mais comum entre adultos, e é basicamente definido por duas principais características fisiopatológicas. Uma delas é o comprometimento da função secretora das células beta pancreática, responsáveis pela produção de insulina. A outra é a resistência ao hormônio insulina pelas células corporais, aumentando assim, o nível de glicemia (Sociedade Brasileira de Diabetes, 2014). Adicionalmente, a DM2 é ocasionada por fatores como o estilo de vida sedentário, obesidade e alimentação com elevada quantidade de carboidratos e de lipídios (Goldman & Ausiello, 2011). Ainda, esta doença apresenta dificuldade de diagnóstico, principalmente na fase inicial, pois apresenta evolução insidiosa e, de modo frequente, somente é detectada depois de complicações decorrentes à enfermidade (Brasil, 2013).

Diante do exposto, neste trabalho, aplicou-se o processo computacional de Mineração de Dados (MD) no auxílio e na extração de conhecimentos em Base de Dados (BD) de DM2 (Kavakiotis *et al.*, 2017; Zhang *et al.*, 2018). Com isso, espera-se contribuir para o apoio de especialistas da saúde no diagnóstico dessa grave enfermidade.

Material e Métodos

Para este trabalho foram utilizados os seguintes materiais: a) Base de dados (BD) *Pima Indians Diabetes* (Dua & Graff, 2019), constituída por 768 indivíduos do sexo feminino, sendo que 268 indivíduos apresentam DM2 enquanto que em 500 não são acometidos pela doença. Essa BD é composta pelos atributos/características Glicemia (G), Pressão Arterial (PA), Número de Gestações (NG), Espessura da Prega Cutânea do Tríceps Braquial (EP), Nível de Insulinemia (NI), Índice de Massa Corporal (IMC), Características Hereditárias (CH) e Idade (I); b) *Notebook* Lenovo g40 com 8GB de memória RAM, processador Intel i7 5500U 2.40GHZ e sistema operacional Windows 10; c) *Software* Weka versão 3.8.2; d) Máquina Virtual Java versão 1.8; e) Linguagem de programação e interpretador Python versão 3.7.

O processo de MD aplicado à base *Pima* consistiu nas etapas de pré-processamento, extração de padrões e pós-processamento (Kavakiotis *et al.*, 2017). O pré-processamento foi instanciado pela Seleção de Atributos (SA) e a etapa de extração de padrões foi representada pela tarefa de classificação. Por fim, o pós-processamento foi concretizado por meio do método de validação cruzada, o qual foi adotado para avaliar o desempenho da classificação.

Com o auxílio do *software* Weka, o algoritmo ReliefF foi empregado para a SA, ranqueando as características que melhor diferenciam os casos positivos dos negativos (pré-processamento). Após, o algoritmo de classificação J48 foi aplicado para extrair padrões da BD a partir dos indivíduos descritos por 50% dos melhores atributos ranqueados pelo ReliefF e por todos os atributos (extração de padrões). Na

sequência, as árvores J48 geradas foram utilizadas para classificar os indivíduos como portadores ou não da DM2 e então avaliadas por meio do método de validação cruzada *k folds*, sendo o valor de *k* igual a 10 (pós-processamento). Depois desses procedimentos, utilizou-se a linguagem de programação Python para implementar e realizar testes estatísticos descritivos e analíticos, fixando-se o intervalo de confiança em 95%.

Resultados e Discussão

A Tabela 1 representa a ordem dos atributos ranqueados pelo algoritmo ReliefF, sendo a primeira posição destinada ao atributo com maior importância e a última destinada ao menos importante.

Tabela 1 – Ordem de importância dos atributos selecionados pelo algoritmo ReliefF

Ranque	1	2	3	4	5	6	7	8
Atributo	G	IMC	EP	NG	I	CH	PA	NI

O atributo classificado como mais importante foi a Glicemia, fator esse fundamental para detecção de DM2, uma vez que o alto nível de glicemia é situação determinante da DM2. Esse atributo é seguido por IMC e é digno de nota que a obesidade também é um importante indicador dessa enfermidade (Alves *et al.*, 2017). O atributo EP, em terceiro lugar, apresenta correlação com DM2 pela associação com a obesidade, fator esse que está diretamente relacionado com DM2. Além disso, o atributo NG, em quarto lugar, pode apresentar-se como indicador de desenvolvimento de futura condição de DM2. Sabe-se que é possível o surgimento de DM gestacional, ou seja, durante a gravidez ocorre o aumento de glicemia no corpo, e, estes sujeitos apresentam maior probabilidade de desenvolver DM2 ao longo da vida (Sociedade Brasileira de Diabetes, 2014). Já o atributo NI pode ter sido selecionado como o menos importante no ranqueamento desse trabalho pelo fato de DM2 ter como uma das bases fisiopatológicas a resistência à ação da insulina nas células corporais e não a diminuição e ou ausência desse hormônio decorrentes da falência na produção pelas células pancreáticas (Barbosa & Camboim, 2016).

A taxa média e, inseridos em parênteses, os respectivos desvios-padrões, do erro da classificação realizada pelas duas árvores geradas são apresentadas na Tabela 2.

O modelo que foi gerado utilizando 50% dos melhores atributos conseguiu classificar corretamente 571 instâncias das 768, totalizando 74,35% de acurácia (complemento do erro médio). Já o modelo gerado a partir de todos os atributos da BD foi capaz de classificar de maneira correta 567 instâncias, o que representa

73,82% de acurácia. Esse resultado se aproxima daquele publicado em Iyer *et al.* (2015), ainda que os autores tenham efetuado operações adicionais de pré-processamento na mesma BD para apoiar o algoritmo J48.

Tabela 2 – Média e desvio-padrão dos erros das árvores de decisão geradas

Modelo com 50% dos melhores atributos	Modelo com todos atributos
25,65 (5,30)	26,18 (5,90)

A análise estatística foi feita inicialmente por meio do teste de normalidade de Shapiro-Wilk na taxa de erro atingida em 10 *folds* pelas árvores de decisões. Como resultado, observou-se que esses dados seguem uma distribuição normal (p -valor $> 0,05$). Posteriormente, o teste estatístico T emparelhado foi utilizado e não foi encontrada diferença estatisticamente significativa entre as médias de erro dos modelos gerados (p -valor $> 0,05$) para o intervalo de confiança de 95%. Desse modo, a árvore gerada por meio dos atributos selecionados pelo ReliefF pode ser considerada superior ao modelo gerado utilizando todos os atributos, uma vez que alcançou um desempenho de classificação semelhante enquanto reduzia pela metade o número de características. Ainda, a quantidade de nós produzidos foi de 17 contra 39 do classificador construído com todas as características. Dessa maneira, é possível compreender e interpretar a árvore de decisão de modo mais simples, uma vez que o número de nós do modelo gerado com SA foi menor.

Adicionalmente, vale ressaltar que o processo de SA é fundamental, pois possibilita encontrar um subconjunto de características que define a BD tão bem quanto o conjunto original. Além disso, esse processo evita o problema da “maldição da dimensionalidade”, condição esta associada à alta dimensionalidade de dados (Liu & Motoda, 2007). Dessa maneira, é possível remover atributos redundantes, ou seja, significativamente correlacionados e/ou irrelevantes, inúteis na discriminação das classes. Esses procedimentos aceleram o processo de aprendizado dos algoritmos e podem contribuir na melhora dos resultados.

Conclusões

O algoritmo ReliefF foi eficaz na seleção de atributos e as árvores de decisão foram capazes de prever casos negativos e positivos de DM2. Portanto, foi possível utilizar o processo de MD na classificação dessa doença.

Agradecimentos

Ao PIBITI/Unioeste/CNPq pelo financiamento da bolsa de Iniciação Tecnológica.

Referências

Alves, C.R., Nogueira, S.P. & Barbosa, M.L.C. (2017). Fármacos para o tratamento do diabetes tipo 2: uma visita ao passado e um olhar para o futuro. *Revista Virtual de Química* **9**, 514-534.

Barbosa, S.A. & Camboim, F. (2016). Diabetes mellitus cuidados de enfermagem para controle e prevenção de complicações. *Temas em Saúde* **16**, 404-417.

Brasil (2013). *Diabetes mellitus: Estratégias para o cuidado da pessoa com doença crônica*. Brasília: Ministério da Saúde.

Dua, D. & Graff, C. (2019). UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>. Acesso em 10 de Setembro de 2020.

Goldman, L. & Ausiello, D. (2011). *Cecil medicina interna*. Rio de Janeiro: Elsevier.

International Diabetes Federation (2019). IDF diabetes atlas. <https://www.diabetesatlas.org/en/>. Acesso em 1 de Setembro de 2020.

Iyer, A., Jeyalatha, S. & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *International Journal of Data Mining & Knowledge Management Process* **5**, 1-15.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal* **15**, 104-116.

Liu, H. & Motoda, H. (2007). *Computational methods of feature selection*. Boca Raton: Chapman & Hall.

Sociedade Brasileira de Diabetes (2014). *Diretrizes da Sociedade Brasileira em Diabetes 2014-2015*. Rio de Janeiro: AC Farmacêutica.

Zhang, J., Liu, J. & Qin, X. (2018). Advance in early biomarkers of diabetic nephropathy. *Revista da Associação Médica Brasileira* **64**, 85-92.