Predição de Diabetes com Redes Neurais Artificiais Bayesianas

Ítalo Della Garza Silva[0000−0002−7848−9147]

Departamento de Ciência da Computação, Universidade Federal de Lavras,
Lavras MG 37200-000, BR
dcc@dcc.ufla.br
https://ufla.br/

Resumo O presente trabalho avalia o desempenho das Redes Neurais Artificiais Bayesianas no contexto da predição da doença Diabetes, utilizando uma base de dados bastante explorada na literatura, e avaliando o sistema quanto a acurácia, precisão e revocação.

Keywords: Redes Neurais Bayesianas · Mineração de Dados · Diabetes

1 Introdução

A Mineração de Dados (MD) é uma área que tem tido avanços consideráveis, graças ao crescente aumento da facilidade para se adquirir os equipamentos necessários à armazenagem combinado à importância que os dados têm para empresas e governos [6]. A MD combina conhecimentos advindos de Computação (Aprendizado de Máquina, Inteligência Artificial e Banco de Dados), Probabilidade e Estatística objetivando a extração de informações relevantes em bases de dados para uso posterior. Segundo [8], a MD pode ser entendida como a aplicação de um ou mais algoritmos para se extrair padrões de bases de dados, sendo um dos passos no processo de Descoberta de Conhecimento em Bases de Dados (KDD). [6] afirma que as tarefas abordadas em MD são Descrição, Estimação, Predição, Agrupamento, Associação e Classificação, foco do presente trabalho.

Cada vez mais, os campos de Mineração de Dados e Inteligência Artificial têm contribuído para o avanço da medicina [12]. Graças à essa união, avanços consideráveis na qualidade dos tratamentos e diagnósticos médicos fazem com que cresça a qualidade de vida e a longevidade da população. Por exemplo, [1] mostra como tarefas da mineração de textos são aplicadas em várias contextos da Saúde, citando sumarização, sistemas de respostas, extração de informação e ontologias biomédicas. Nesse contexto, predição eficaz da presença de doenças baseada em características previamente adquiridas de um dado paciente é fundamental para prevenir o desenvolvimento daquela doença em estágios mais avançados ou até mesmo irreversíveis. Baseando-se nisso, o presente trabalho propõe um método para classificar pacientes como diabéticos e não-diabéticos baseando-se em dados biológicos coletados dos mesmos.

Este trabalho busca, a partir de uma seleção e conversão adequada dos dados presentes, além de um método de classificação adequado com Redes Neurais Artificiais Bayesianas, propor um sistema de predição de alta eficácia a ser utilizado no presente contexto.

2 Referencial Teórico

Segundo [12], a Classificação, foco desse trabalho, é amplamente usada nos trabalhos que aplicam MD em Medicina. A Classificação compreende, ainda segundo [12], "uma função de aprendizagem preditiva que classifica um item de dado em uma dentre várias classes predefinidas". [9] também cita, em sua conclusão, a importância da MD para a Saúde.

Em 2.1 é apresentado uma descrição a respeito das Redes Neurais Bayesianas. Na subseção 4.1 é fornecida uma descrição da base de dados utilizada no presente trabalho.

2.1 Redes Neurais Bayesianas

Redes Neurais Bayesianas (RNB) são modelos de redes neurais cuja abordagem é diferente da abordagem convencional dos *Multilayer Perceptrons*. A escolha por esse modelo se deve a sua robustez e seu bom comportamento sobre bases de dados pequenas[17]. De fato, segundo [20], algoritmos que fazem uso do aprendizado Bayesiano figuram-se entre os mais usados na literatura. A combinação da teoria de Bayes com o modelo de aprendizagem baseado no cérebro humano resultam em um método resistente ao *overfitting*. Conforme descrito em [4] e [10], ela se dá por meio de aprendizagem bayesiana no qual a partir dos pesos desse neurônio pode ser obtido cada resposta através de uma função probabilística P ligada a esse neurônio, ou seja,

$$P(\mathbf{y}|\mathbf{x}, \mathbf{w}, \beta),$$

onde $\mathbf{y}$ é a resposta desejada, $\mathbf{x}$ é o vetor de entrada, $\mathbf{w}$ é o conjunto de pesos associados ao neurônio e β é a precisão (variância inversa), usada para compor a função de distribuição. É possível usar técnicas de inferência bayesiana para a realização do aprendizado, isto é, busca-se inferir os pesos a partir da distribuição dos dados, a resposta desejada e a precisão. No entanto, a complexidade dos cálculos demandaria muito tempo para a aquisição. A maneira possível para acelerar esse processo é o uso de métodos de inferência mais baratos, mesmo que haja possibilidade de se diminuir a eficácia, como Markov Chain Monte Carlo (MCMC).

Na classificação, em vez de se fixar valores de pesos e *bias* para a rede, são feitas várias iterações, cada uma amostrando um conjunto de pesos e bias diferente. Além disso, a RNB retorna um conjunto de respostas, que compreende uma distribuição probabilística. A RNB é capaz de calcular a incerteza de uma classificação feita, consequentemente tendo a opção de não considerar um conjunto de classificações com alta taxa de incerteza. A Figura 1 ilustra a estrutura de uma RNB.

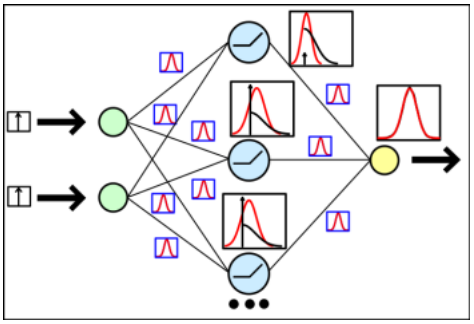


Figura 1: Ilustração de uma estrutura de Rede Neural Bayesiana

3 Trabalhos Relacionados

Muitas abordagens na área da Saúde exploram a RNB. [19] fez uso de RNBs convolucionais para predizer a incerteza gerada por ruídos de prontuários eletrônicos. A base de dados usada continha muitos ruídos por ser proveniente de uma grupo muito diversificado de pacientes. Logo, foi adicionado uma variável aleatória representando o ruído presente na base de dados. Uma rede foi então treinada através de *MCMC* usando o otimizador Adam, o mesmo utilizado no presente trabalho. Foi utilizado em [16] um modelo de *Deep Learning* Bayesiano para a detecção de retinopatia diabética em imagens fundoscópicas, superando os métodos de estado-da-arte até então. O trabalho define uma incerteza associada à presença da retinopatia diabética e define um conjunto de modelos treinado através de MCMC. Foram testadas duas arquiteturas de rede neural, uma definida manualmente e outra baseada em um dos trabalhos mais bem pontuados no Kaggle, de onde o trabalho buscou a base de dados.

Para a base de dados sobre a qual o presente trabalho foi feito (ver subseção 4.1), houveram sucessivas tentativas de solução. Em [2] foram testadas técnicas de *Deep Learning* na base de dados PIMA, porém fazendo uma análise preditiva nos dados de forma a melhorar os resultados. O trabalho primeiramente fez uma seleção de atributos baseada em sua singularidade. Em seguida, realiza o treinamento em uma Máquina de Boltzmann Restrita para fazer a classificação. Foi atingida alta precisão para ambos os doentes e não doentes, porém baixa revocação para os doentes. A acurácia resultante foi 81%. Em [11] foi testado o algoritmo *ADAP Learning*. Uma curva de especificidade $\times$ sensibilidade para o algoritmo foi gerada contendo um *crossover* no ponto 0.76 para ambos. Houve também uma tentativa de solução baseada em SVM[13] com um *kernel* RBF, com um diferencial na etapa de separação dos dados em treino e teste. Os dados utilizados para treinamento foram somente os dados mais coerentes, enquanto os dados difíceis de se agrupar foram os dados de teste, podendo dessa forma lidar com as anomalias da base separadamente. Obteve acurácia média de 82.74%. Em [7], foi proposta uma solução que combina *Naive Bayes* para classificação e

algoritmo genético para a seleção de *features*. Os resultados foram avaliados em diversas métricas e o algoritmo obteve alta acurácia (76.95%).

4 Metodologia

A base de dados foi adquirida no *Kaggle Datasets*[15] sendo que nenhum campo dentre os dados estava vazio ou incorretamente preenchido, não tendo sido necessário, portanto qualquer tratamento de correção. A classificação foi realizada através de Redes Neurais Bayesianas. Dado que os dados encontravam-se em escalas diferentes e como a maioria das funções de ativação não lidam bem com dados com alta variação, foi fundamental uma normalização adequada dos dados para que estes se encontrassem entre 0 e 1. Logo, a normalização *Minimax* foi aplicada na solução, implementada na biblioteca *scikit-learn*[18]. Para lidar com a existência de atributos menos significativos para a realização da classificação, foi conveniente selecionar somente os melhores atributos, os quais deveriam maximizar as métricas finais. Sendo assim, a estratégia para a seleção de *features* consistiu na enumeração de 6 redes neurais, com suas respectivas entradas contendo de 3 a 8 neurônios. Ou seja, a primeira rede continha 3 neurônios de entrada, a segunda continha 4, e assim sucessivamente, até chegar a 8 entradas para a última rede. Para a seleção dos melhores atributos, foi utilizada a função *SelectKBest* do *scikit learn*, utilizando o teste χ^2 para a comparação entre os atributos. Foi calculado, para critério de comparação com os trabalhos anteriores, a acurácia. A construção da Rede Neural Artificial Bayesiana foi realizada através das bibliotecas *Pyro*[3] e *Pytorch*. O procedimento metodológico é ilustrado na Figura 2.

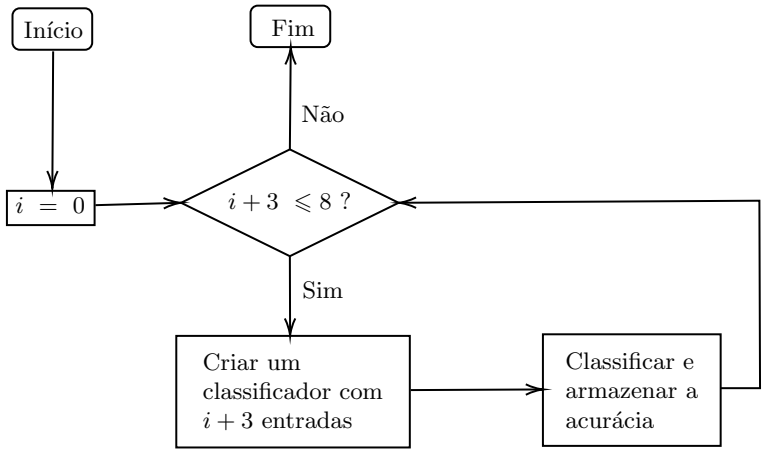


Figura 2: Fluxograma ilustrativo que descreve o processo metodológico do presente trabalho.

A arquitetura de rede usada no presente trabalho é simples, contendo uma camada intermediária com 4 neurônios (verificou-se que produzia-se, com essa quantidade, resultados mais satisfatórios) e 2 nós na camada de saída, com as entradas variando de 3 a 8. Os dados foram divididos em 48 *batches* iguais de 16 dados cada, para a leitura correta pelo *Pytorch*. 30 *batches* foram definidos como os dados de treinamento e o restante como dados de teste, sendo que a divisão dos dados é aleatória. Para o treinamento da rede, foi definido um modelo de inferência variacional [5] usando o otimizador Adam[14], submetido a 100 iterações. Para a obtenção dos resultados, foram feitas 200 amostragens e o resultado foi dado pela média dos resultados.

Durante a etapa dos testes, na tentativa de melhorar os resultados e reduzir a variação da acurácia (ver Seção 5), a classificação foi incrementada a uma característica particular das Redes Bayesianas, que é a possibilidade de, através de uma estimativa da certeza dos resultados, a rede pode optar ou não pela classificação. Nesse caso, durante a etapa de testes, são obtidos 100 resultados para cada entrada da rede, sendo que esses resultados são, ao invés da classe em si, os valores de cada nó de saída aplicados normalizados por *log softmax*. Esses resultados são convertidos em probabilidades através do cálculo de seu exponencial. Há então a extração da mediana dessas probabilidades (50º percentil) e, caso esse valor seja maior que um determinado limiar, a classificação é feita. Foram testados os limiares de 0.6, 0.7 e 0.8. A solução foi disponibilizada *online* para pesquisas futuras[1].

4.1 Base de Dados

O trabalho foi feito sobre a base de dados PIMA[15], bastante explorada na literatura, que se trata do resultado de uma coleta de dados realizada sobre a população indígena Pima residente próxima *Phoenix, Arizona*, nos Estados Unidos. A população esteve sendo monitorada regularmente desde 1965 pelo *National Institute of Diabetes and Digestive And Kidney Diseases*. Cada os dados que compõem essa base de dados foram extraídos de mulheres indígenas acima de 21 anos. São registrados, para cada pessoa nesta base, A quantidade de gravidez, a concentração de glucose no plasma sanguíneo, a pressão sanguínea (em mm Hg), a espessura da pele (em mm), a taxa de insulina (em mu U/ml), o índice de massa corporal (em Kg/m^2), a idade e a presença ou não de diabetes.

5 Resultados

Os resultados mostram alta acurácia, embora a mesma tenha sofrido boa variação entre diferentes execuções. A diferença de resultados entre as redes de diferentes números de entrada varia conforme a execução, alternando a cada momento a rede com melhor acurácia. Quando se adiciona o filtro das incertezas, percebe-se que as acurácias aumentam e tendem a variar menos, embora

[1] Disponível em: https://github.com/italodellagarza/TPMineracaoDados

não de maneira considerável a ponto de se distinguir a melhor configuração de rede. A acurácia varia entre 70% e 80% quando a rede é forçada a classificar, ignorando as incertezas.

Na Figura 3, ilustram-se as diferentes curvas de perda para as diferentes redes testadas em uma execução, onde uma dada rede de número i possui $i + 3$ entradas. A taxa de erro é obtida a cada iteração das otimizações para um *batch*. Para esse modelo, a máxima acurácia obtida com limiar fixado em 0.6 foi 80%, obtido pela Rede 1, que classificou 76% das entradas. Ignorando as incertezas, essa mesma rede obteve uma acurácia de 76%. Para um limiar de 0.7, a rede com o melhor resultado foi a Rede 3, que obteve uma acurácia de 86%, classificando 57% dos dados e, quando forçada a classificar todos os dados, obtendo 71% de acurácia. A Rede 4, com o limiar em 0.8, foi a que obteve a melhor acurácia, de 91%, porém classificou somente 33% dos dados. Essa mesma rede obteve acurácia de 75% quando ignorou as incertezas.

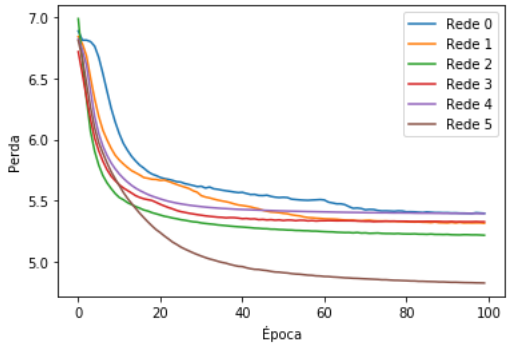
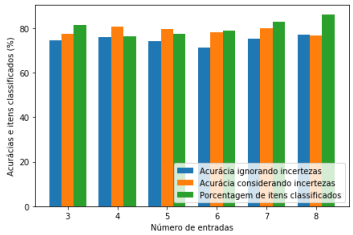


Figura 3: Curvas de perda para as diferentes redes testadas

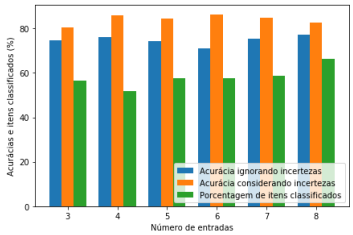
Para essa mesma execução, a Figura 4 mostra as acurácias e porcentagem de itens classificados para cada variação de limiar adotada, diferenciando o número de entradas da rede.

6 Discussão

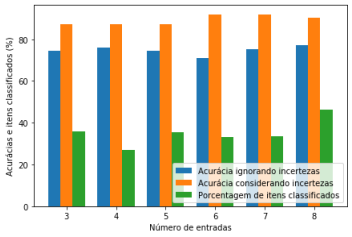
Os resultados são similares aos que exploraram a mesma base de dados (ver subseção 3), mostrando que o modelo aqui descrito executa de maneira satisfatória a classificação nesse caso. É importante ressaltar que, para limiares muito altos, o modelo obteve alta acurácia, porém muitos dados foram ignorados, e isso inutiliza a aplicação do modelo em muitos dados, já que estes têm boas chances de não serem classificados pelo modelo. Verifica-se também que a presença de mais atributos tendem a gerar resultados melhores, porém isso não ocorre em todos os casos. A acurácia ainda é bastante inconsistente, portanto trabalhos



(a) limiar em 60%



(b) limiar em 70%



(c) limiar em 80%

Figura 4: Variação da porcentagem e acurácias para diferentes limiares de probabilidade mínima × número de entradas na rede.

futuros poderiam tentar diminuir essa variação, aprimorando, por exemplo, a forma como os dados de treino e teste são separados. Para melhores resultados, acredita-se que mais atributos fornecidos, ou seja, mais dados biológicos coletados, poderiam fornecer melhores resultados. Também é possível, em trabalhos futuros, adicionar mais camadas ao modelo, aumentando a profundidade. Outro ponto que poderia ser explorado seriam alterações nos algoritmos adotados pelo presente trabalho, mantendo-se a estratégia aqui definida, trocando por exemplo o método de inferência, o otimizador ou até mesmo o pré-processamento. Assim sendo, a estratégia aqui definido ainda tem muito a ser explorada.

Referências

1. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: A brief survey of text mining: Classification, clustering and extraction techniques (2017)
2. Balaji, H., Iyenger, N.C.S.N., Caytiles, R.: Optimal predictive analytics of pima diabetics using deep learning. International Journal of Database Theory and Application **10**, 47–62 (09 2017). https://doi.org/10.14257/ijdta.2017.10.9.05
3. Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., Goodman, N.D.: Pyro: Deep Universal Probabilistic Programming. Journal of Machine Learning Research (2018)
4. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)

5. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for statisticians. Journal of the American Statistical Association **112**(518), 859–877 (Feb 2017). https://doi.org/10.1080/01621459.2017.1285773, http://dx.doi.org/10.1080/01621459.2017.1285773
6. Camilo, C.O., da Silva, J.C.: Mineração de dados: Conceitos, tarefas, métodos e ferramentas. Tech. rep., Universidade Federal de Goiás, Departamento de Informática (08 2009)
7. Choubey, D., Paul, S., Kumar, S., Kumar, S.: Classification of pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. pp. 451–455 (11 2016). https://doi.org/10.1201/9781315364094-82
8. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine **17**(3), 37 (Mar 1996). https://doi.org/10.1609/aimag.v17i3.1230, https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230
9. Galvão, N.D., Marin, H.d.F.: Técnica de mineração de dados: uma revisão da literatura. Acta Paulista de Enfermagem **22**, 686 – 690 (10 2009)
10. Hinton, G.E., Neal, R.M.: Bayesian learning for neural networks (1995)
11. Johannes, R.S.: Using the adap learning algorithm to forecast the onset of diabetes mellitus. Johns Hopkins APL Technical Digest **10**, 262–266 (1988)
12. Jothi, N., Rashid, N.A., Husain, W.: Data mining in healthcare – a review. Procedia Computer Science **72**, 306 – 313 (2015). https://doi.org/https://doi.org/10.1016/j.procs.2015.12.145, http://www.sciencedirect.com/science/article/pii/S1877050915036066, the Third Information Systems International Conference 2015
13. Karatsiolis, S., Schizas, C.N.: Region based support vector machine algorithm for medical diagnosis on pima indian diabetes dataset. In: 2012 IEEE 12th International Conference on Bioinformatics Bioengineering (BIBE) (2012). https://doi.org/10.1109/BIBE.2012.6399663
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014)
15. Learning, U.M.: Pima indians diabetes database (2016), https://www.kaggle.com/uciml/pima-indians-diabetes-database
16. Leibig, C., Allken, V., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. bioRxiv (10 2016). https://doi.org/10.1101/084210
17. Mullachery, V., Khera, A., Husain, A.: Bayesian neural networks. ArXiv **abs/1801.07710** (2018)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python . Journal of Machine Learning Research **12**, 2825–2830 (2011)
19. Qiu, R., Jia, Y., Hadzikadic, M., Dulin, M., Niu, X., Wang, X.: Modeling the uncertainty in electronic health records: a bayesian deep learning approach. CoRR **abs/1907.06162** (2019), http://arxiv.org/abs/1907.06162
20. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 algorithms in data mining. Knowledge and Information Systems **14**(1), 1–37 (Jan 2008). https://doi.org/10.1007/s10115-007-0114-2, https://doi.org/10.1007/s10115-007-0114-2