Avaliando a Utilização do Aprendizado de Máquina em um Sistema de Apoio à Predição de Diagnósticos Médicos

Airan Poelking Camargo[1], **Julio Cesar Duarte**[1]

[1] Instituto Militar de Engenharia
Rio de Janeiro - RJ, Brazil

{airancamargo, duarte}@ime.eb.br

Abstract. *The diagnostic hypothesis is a preliminary decision-making process of the physician that is generated based on the clinical and laboratory data available at the end of the consultation. In order to assist the physician's decision-making process, we propose the use of the national registry of signs, symptoms and diseases (ICD-9) and patient records by applying machine learning techniques with the objective of generating hypotheses while providing a second opinion for the professional. We, also, present a qualitative evaluation with results generated by different algorithms applied in two scenarios of the same datasets. We propose a cascade model that was most efficient in predicting the CID chapter based on a classifier committee that obtained an accuracy rate of 54.19 % and after a classification algorithm based on Decision Trees, which obtained an average accuracy of 61.79% to identify each CID category.*

Resumo. *A hipótese diagnóstica é um processo preliminar de tomada de decisão do médico gerado em função dos dados clínicos e laboratoriais disponíveis ao final da consulta. De forma a auxiliar o processo de tomada de decisão de um médico, propomos a utilização do cadastro nacional de doenças (CID) e do prontuário de pacientes para aplicar técnicas de aprendizado de máquina com o objetivo de gerar hipóteses diagnósticas, fornecendo apoio ao diagnóstico para o profissional. Ao final, apresentamos uma avaliação qualitativa com resultados gerados por diferentes algoritmos aplicados em dois cenários do mesmo conjuntos de dados. Propomos um modelo composto que foi mais eficiente na predição do capítulo CID baseado em um comitê de classificadores que obteve uma taxa de acurácia de 54,19% e após isso um algoritmo baseado em Árvores de Decisão, que obteve uma acurácia média de 61,79% para identificar cada categoria CID.*

1. Introdução

O processo de tomada de decisões ocorre em diversos pontos da atividade do médico, desde o estudo de como a doença se espalha [Bezerra et al. 2017] até a hipótese diagnóstica. A hipótese diagnóstica é um diagnóstico preliminar, levantado pelo médico em função dos dados clínicos e laboratoriais disponíveis ao final da consulta, ou seja, uma hipótese de trabalho, que irá nortear de maneira geral e flexível, os próximos passos da investigação clínica. As Unidade de Pronto Atendimento (UPAs) funcionam 24 horas por dia, sete dias por semana, e podem resolver grande parte das urgências e emergências [Prefeitura RJ 2018]. Todo atendimento é registrado e armazenado em uma base de conhecimento que poderá ser utilizada para melhorar, cada vez mais, os processos médicos, de admissão e de classificação realizados nessas unidades [Ministério da Saúde 2018]. A Classificação Internacional de Doenças (CID) fornece códigos relativos à classificação de doenças e de uma grande variedade de sinais, sintomas, aspectos anormais, queixas, circunstâncias sociais e causas externas para ferimentos ou doenças. O código CID é composto por 4 dígitos e organizado em capítulos, divididos em agrupamentos, categorias e subcategorias, ex.: K35.0 representa apendicite aguda com peritonite generalizada. O Sallusys, sistema para operação e gestão de UPAs que realizam pronto atendimento,

utiliza o CID-9 no prontuário eletrônico do paciente para registrar antecedentes, queixa principal, exame físico e hipótese diagnóstico, mas também possui informações importantes em texto livre e sinais vitais.

O objetivo deste trabalho é utilizar uma base histórica de diagnósticos médicos formatada com o cadastro nacional de sinais, sintomas e doenças (CID-9) em conjunto com técnicas de AM para gerar um sistema de informação de apoio à decisão de hipóteses diagnósticas a partir dos sintomas identificados e informados pelo especialista, gerando um apoio ao diagnóstico para o profissional da medicina. Dentre as várias abordagens existentes de aprendizado de máquina (AM), daremos ênfase a preditiva também conhecida como supervisionada. Alguns dos algoritmos de AM mais comumente utilizados, nesse caso, são o Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN) e Support Vector Machine (SVM) [dos Santos 2016]. O resultado deste estudo fornece um modelo computacional de apoio à decisão do especialista médico gerando maior velocidade no atendimento e uma menor possibilidade de erro, assim como fornecer informações e novas visões para realizar o diagnóstico, mas não tem a intenção de substituir os médicos. O restante deste artigo está organizado da seguinte forma: o estado da arte revisado na Seção 2. O conjunto de dados e o método proposto é detalhado na Seção 3 e 4, seguido por relatórios de avaliação dos resultados na Seção 5. Concluímos o artigo na Seção 6, onde são apresentadas algumas oportunidades de trabalhos futuros.

2. Trabalhos Relacionados

Várias tentativas foram feitas para atribuir automaticamente códigos CID-9 a documentos clínicos desde a década de 1990. [Atutxa et al. 2019] avaliou diferentes arquiteturas de redes neurais para classificação de documentos de várias classes. Utilizaram três conjuntos de dados, francês, italiano e húngaro. Os resultados da codificação multilíngue do CID-10, superaram várias abordagens alternativas, o sistema RNN-CNN obteve uma medida F de 0,838 para francês, CNN-RNN obteve 0,963 para húngaro e RNN-RNN obteve 0,952 para italiano. [Baumel et al. 2018] investigou quatro modelos para atribuição de múltiplos códigos CID-9 (SVM, CBOW, CNN e HA-GRU) e contaram com o conjunto de dados clínicos MIMIC II e MIMIC III. O modelo HA-GRU alcançou o melhor desempenho, 55.86% F1 em MIMIC III e no MIMIC II o HA-GRU alcançou 7% de melhoria absoluta no F1. [Nguyen et al. 2018] apresentou uma abordagem baseada em PLN utilizando métodos de normalização padrão de texto livre para SNOMED CT e tabelas de mapeamento SNOMED CT para CID-10 na codificação de diagnóstico. Os dados foram obtidos de três hospitais australianos. A abordagem proposta alcançou sensibilidade de 54,1% e valor preditivo positivo de 70,2%. [Zhang et al. 2017] estudou o problema do desequilíbrio de dados na tarefa de atribuição automática do código CID-9-CM e coletaram dados estrategicamente do PubMed para enriquecer os dados de treinamento. Validaram os métodos no conjunto CMC-dataset e os resultados da avaliação indicaram que o método pode melhorar o desempenho dos classificadores no nível de macro-média. [dos Santos 2016] analisou a aplicação de algoritmos de AM no diagnóstico de dengue. Concluíram que o algoritmo Naive Bayes obteve o pior desempenho e o SVM foi o melhor obtendo 0,89 de F-measure. [Kavuluru et al. 2015] avaliou abordagens de AM supervisionada para atribuir automaticamente a CID-9-CM codificadas para prontuários eletrônicos. Utilizaram um conjunto de dados com visitas de pacientes do Centro Médico da Universidade de Kentucky. [Kamkar et al. 2015] utilizou um modelo Tree-Lasso e o

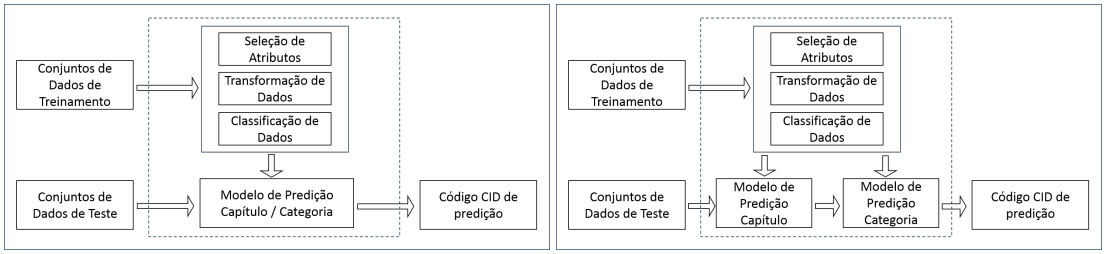
comparou com outros métodos de seleção de atributos. Utilizaram dados sintéticos e dois conjuntos de dados do mundo real (Câncer e Infarto Agudo do Miocárdio). Mostraram que o desempenho de classificação do Tree-Lasso é comparável ao Lasso e melhor que outros métodos. [Chiaravalloti et al. 2014] apresentou um sistema de suporte à codificação para auxiliar médicos na atribuição de códigos CID-9-CM. o sistema proposto é baseado em PLN e pesquisa de texto em bases de conhecimento. Os resultados das experiências mostram uma precisão de 92%.

3. Metodologia

O trabalho utiliza um conjuntos de dados de 67.451 registros de atendimentos da especialidade clínica médica no período do mês de abril de 2012 ao mesmo mês em 2013 do sistema Sallusys da UPA de Senador Camará - RJ. A base de dados fornecida é completamente anonimizada, sem nenhum registro que permita identificar o paciente. O principal objetivo dessa pesquisa é utilizar o grande conjuntos de dados da aplicação Sallusys para gerar um modelo de predição de doenças (códigos CID) a partir da anamnese registrada pelo especialista no prontuário médico do paciente. Os principais atributos selecionados são: antecedentes, queixa principal, traumas e causas externas, exame físico e protocolo (CID-9) da especialidade clínica médica. Ainda não foi utilizado em pesquisa um modelo de dados de uma aplicação onde são registrados códigos CID-9 para sinais e sintomas no prontuário do paciente e, por esse motivo, espera-se que o modelo de predição seja mais eficiente.

O estudo avaliou 2 modelos diferentes onde, em ambos, foram aplicados uma seleção de atributos, transformação de dados no pré-processamento e aplicação de algoritmos de AM para gerar um modelo de predição. Na seleção de atributos foram geradas duas visões diferentes do mesmo conjunto de dados. Uma visão agrupada dos códigos CID de sinais / sintomas para cada grupo de atributos da anamnese e uma visão individualizada de cada código CID de sinais e sintomas da anamnese, onde os valores são estritamente binários. Na visão agrupada, por exemplo, 0 ou mais códigos CID, separados por um delimitador são valores possíveis para o atributo queixa principal e na visão individualizada 0 ou 1 são valores possíveis para cada código CID. No caso da visão agrupada foram aplicadas conversões de valores nominais em valores numéricos para aplicação das técnicas de AM. Um valor marcado como "N/D" significa que tal atributo não foi preenchido durante o diagnóstico. Ainda no pré-processamento, uma análise no conjunto de dados foi aplicada. O conjunto de dados inicial possui 570 categorias (classes) e 51.618 registros (diagnósticos). Os capítulos XVIII, XIX, XX, XXI são códigos de sinais e sintomas utilizados na anamnese, por esse motivo não representarão uma hipótese de diagnóstico e os capítulos XV e XVI possuem uma quantidade muito pequena de diagnósticos e, por esse motivo, foram removidos. Os códigos CID (categoria) com menos de 3 diagnósticos e amostras com menos de 3 instâncias também foram desconsideradas. A linguagem utilizada para executar os experimentos de aprendizado de máquina foi o Python em conjunto com as bibliotecas Scikit-Learn, Pandas, Numpy e a ferramenta Eclipse + Pydev. O **modelo único** utiliza os atributos para obter um capítulo CID de doença ou uma categoria CID de doença como hipótese diagnóstico. Os algoritmos de AM: NB, DT, RF, KNN e SVM foram utilizados no treinamento para gerar o modelo de predição. Os parâmetros dos algoritmos foram configurados utilizando testes iniciais em subamostras menores da base de dados, como a profundidade máxima da árvore = 15 para

DT e RF e o parâmetro k = 1 para KNN. Um estudo mais profundo ainda deverá se aplicado na avaliação desses parâmetros de forma a obter os melhores resultados. O **modelo composto** utiliza os mesmos atributos para identificar um capítulo CID de doença e a partir disso obter uma categoria CID de doença como hipótese diagnóstico. Os algoritmos de AM: Comitê de classificação e DT, o qual apresentou o melhor resultado no modelo único, foram utilizados no treinamento para gerar o modelo de predição. O comitê de classificação foi aplicado para identificar o capítulo CID de doença e o algoritmo DT foi aplicado para identificar categoria CID de doença na sequencia do modelo. Os algoritmos com os melhores resultados no modelo único foram aplicados no comitê de classificação, DT, RF e KNN com pesos 2, 2 e 1 respectivamente e voto flexível. Os mesmos parâmetros do modelo único foram aplicados no algoritmo DT.



(a) Diagrama do modelo único **(b) Diagrama do modelo composto**

Figura 1. Diagramas de proposta de modelo de predição

4. Avaliação dos resultados

A validação cruzada foi utilizada para avaliar os resultados. O método de validação cruzada denominado k-fold consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para estimação dos parâmetros e calcular a acurácia do modelo. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste. A métrica mais comum para classificação é a acurácia, que é a fração de amostras prevista corretamente, acurácia = ((VP + VN) / (VP + VN + FP + FN)). O nosso estudo utiliza essa métrica para comparar e analisar os resultados.

4.1. Análise do Modelo Único

Como pode ser observado, a Tabela 1 apresenta os resultados da validação cruzada (3-Fold) do modelo de predição da categoria CID para ambas as visões do conjuntos de dados (agrupado e individual) de cada algoritmo aplicado. A pequena quantidade de registros em alguns capítulos inviabilizou a abordagem mais tradicional (10-fold). Os ajustes no conjuntos de dados realizados no pré-processamento também estão representados a fim de comparar a evolução dos resultados, mas o estudo considera como resultado final o conjunto de dados com números de amostras maior que 3. O conjunto de dados com amostras maior que 120 indica a necessidade de uma quantidade maior de dados para elevar as taxas de acurácia. O conjunto final na visão agrupada considerou 5 atributos e 14.733 registros, onde os algoritmos NB e KNN apresentaram as menores taxas de acurácia, 1,43% e 29,25% respectivamente e os algoritmos DT, SVM e RF apresentam resultados similares, 41,67%, 42,22% e 41,85% respectivamente. Por outro lado, o conjunto final na visão individualizada considerou 185 atributos e 25.911 registros, onde

os algoritmos NB e KNN também apresentaram as menores taxas de acurácia, 2,76% e 14,18% respectivamente e os algoritmos DT, SVM e RF também apresentam resultados similares, 22,98%, 20,70% e 24,52% respectivamente. Com isso, podemos concluir que o algoritmo SVM e a visão agrupada dos dados apresentaram o melhor resultado, com uma acurácia de 56,45%.

Tabela 1. Validação cruzada de categorias CID-9

Conjunto de Dados	Atributos	# Registros	Naive Bayes	Desison Tree	SVM	Random Forest	KNeighbors
Caso 1: Visão agrupada							
nenhum ajuste aplicado	5	51.618	0,38%	22,52%	21,03%	21,91%	12,83%
capítulos de 15 à 21 removidos	5	32.216	0,60%	24,76%	23,66%	24,80%	15,97%
categorias $<$ 3 diagnósticos removidos	5	31.941	0,63%	24,87%	23,87%	25,29%	16,22%
número de amostras $>$ 3	5	14.733	1,43%	41,67%	42,22%	41,85%	29,15%
número de amostras $>$ 120	5	3.222	76,38%	83,09%	83,09%	83,09%	77,62%
Caso: 2 Visão individualizada							
nenhum ajuste aplicado	185	51.618	0,26%	15,97%	15,88%	18,03%	6,53%
capítulos de 15 à 21 removidos	185	32.216	0,34%	19,38%	17,66%	20,18%	8,27%
categorias $<$ 3 diagnósticos removidos	185	31.941	0,37%	19,55%	17,81%	20,82%	9,40%
número de amostras $>$ 3	185	25.911	2,76%	22,98%	20,70%	24,52%	14,18%
número de amostras $>$ 120	185	9.410	30,54%	47,45%	44,84%	47,45%	32,82%

4.2. Análise do Modelo Composto

Como pode ser observado, A Tabela 2 apresenta os resultados da validação cruzada 3-Fold para o formato do conjunto de dados agrupado e algoritmo comitê de classificadores para identificar o capitulo CID. O modelo composto utilizou apenas o conjunto de dados ajustado no pré-processamento com números de amostras maior que 3, onde foram encontrados 5 atributos e 14.733 registros e uma taxa de acurácia de 54,19%. A partir do capitulo CID identificado o modelo utiliza o algoritmo DT e o mesmo conjunto de dados ajustado, porém isolado por capítulo para identificar a categoria CID. A menor taxa de acurácia encontrada foi do capitulo XI, 46,19%, a maior taxa de acurácia foi 85,88% e taxa média de acurácia foi 61,79%, desconsiderando os capítulos II e III que apresentaram taxa de 100% devido a pequena quantidade de registros.

Tabela 2. Resultados do modelo composto

Conjunto de Dados	Atributos	# Registros	Qualidade (Comitê)
Todos os Capítulos	5	14.733	54,19%
Conjunto de Dados	**Atributos**	**# Registros**	**Qualidade (Árvore)**
Categoria por Capítulo (média)	5	14.733	61,79%

5. Conclusão

A hipótese diagnóstica é um processo de tomada de decisão do médico, onde os dados clínicos e laboratoriais disponíveis são utilizados para gerar um diagnóstico preliminar. O uso do cadastro nacional de sinais e sintomas e doenças (CID-9) para aplicar técnicas de AM e gerar um modelo de predição de hipóteses de diagnóstico eficiente, com o propósito de diminuir o erro médico e melhorar a velocidade no diagnóstico preliminar apresentou excelentes resultados. Com isso, concluímos que o modelo composto obteve o melhor resultado na predição da categoria CID em conjunto com o algoritmo DT, após identificar o capítulo CID com uma taxa média de acurácia de 61,79% e um resultado similar ao

modelo único para identificar o capitulo CID em conjunto com o algoritmo Comitê de Classificadores, com uma taxa de acurácia de 54,19%. Como trabalhos futuros, é interessante adicionar atributos de característica dos pacientes e aplicar algoritmos de AM em dados não estruturados da anamnese do paciente. Assim como utilizar dados mais recentes e não remover instâncias para algumas configurações.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.
Este material é baseado no trabalho suportado pela Pesquisa Científica do Escritório da Força Aérea sob o número de concessão FA9550-19-1-0020.

Referências

Atutxa, A., de Ilarraza, A. D., Gojenola, K., Oronoz, M., and Perez-de Viñaspre, O. (2019). Interpretable deep learning to map diagnostic texts to icd-10 codes. *International Journal of Medical Informatics*, 129:49–59.

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., and Elhadad, N. (2018). Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Bezerra, A., Filho, J. J. B., Braga, R., Oliveira, C., and Oliveira, M. (2017). Dengosa: Um sistema de informação geográfica para apoio à decisão no controle de epidemias. In *Anais Estendidos do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 179–183, Porto Alegre, RS, Brasil. SBC.

Chiaravalloti, M. T., Guarasci, R., Lagani, V., Pasceri, E., and Trunfio, R. (2014). A coding support system for the icd-9-cm standard. In *Healthcare Informatics (ICHI), 2014 IEEE International Conference on*, pages 71–78. IEEE.

dos Santos, A. C. (2016). Aprendizado de máquina aplicado ao diagnóstico de dengue. In *XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*.

Kamkar, I., Gupta, S. K., Phung, D., and Venkatesh, S. (2015). Stable feature selection for clinical prediction: Exploiting icd tree structure using tree-lasso. *Journal of biomedical informatics*, 53:277–290.

Kavuluru, R., Rios, A., and Lu, Y. (2015). An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine*, 65(2):155–166.

Ministério da Saúde (2018). Ministério da saúde. http://shorturl.at/sKUX6.

Nguyen, A. N., Truran, D., et al. (2018). Computer-assisted diagnostic coding: Effectiveness of an nlp-based approach using snomed ct to icd-10 mappings. In *AMIA Annual Symposium Proceedings*, volume 2018, page 807. American Medical Informatics Association.

Prefeitura RJ (2018). Prefeitura rj - upa senador camara. `http://shorturl.at/dhkBW`. [Online; acessado em 22 de Outubro de 2018].

Zhang, D., He, D., Zhao, S., and Li, L. (2017). Enhancing automatic icd-9-cm code assignment for medical texts with pubmed. In *BioNLP 2017*, pages 263–271.